White Paper

Technology
GenAI

Improving Intel Technical Sellers' Effectiveness and Customer Engagement with Help of a Generative AI Chatbot

Authors

Kiran K. Kondru

Sales Strategy and GenAI Applications, Datacenter and AI Segment, Sales and Marketing Group, Intel Corp

Kevin D. Johnson

CTO/GM, Datacenter and AI Segment, Technology Acceleration Office, Sales and Marketing Group, Intel Corp

Einat Ronen

Director, AI Solutions Group, Software and Advanced Technology Group, Intel Corp

Amit Perry

Senior Director, Sales AI, Sales and Marketing Group, Intel Corp

Abstract

The swift advancement of generative artificial intelligence (GenAI) is revolutionizing the business landscape. Companies are tapping into AI's potential to secure a competitive edge, improve decision-making, and streamline operations. However, they must address challenges related to accuracy and the use of current information to ensure reliable outcomes. Intel Corporation, a global leader in semiconductor technology, sought to equip its sales force with easier access to accurate technical data on its products. This whitepaper explores the creation, implementation, and impact of Project Virtual Intel Technical Assistant (VTA), a GenAI-powered chatbot designed for Intel's Data Center and AI (DCAI) business unit and soon to be expanded to other products and business units. By harnessing the power of GenAI, Intel aims to increase sellers' speed and precision in answering customer questions, boost customer engagement, and ultimately drive business expansion. VTA enables sales representatives to quickly obtain answers, ranging from simple to complex inquiries, backed by reliable sources.

Introduction: Navigating the Knowledge Terrain

In the dynamic realm of technology integration, Intel's sellers are at the forefront, crafting tailored solutions for client issues. They engage in deep consultations to understand client challenges and collaborate to design practical, actionable solutions that also inform Intel's product development. These sales experts often tackle novel and intricate problems. Their role is vital in ensuring that Intel's innovative products are seamlessly integrated with the needs of builders, cloud providers, original design manufacturers (ODM), original equipment manufacturers (OEM), and enterprise end customers—including internet providers.

These sellers face the formidable challenge of navigating through vast amounts of complex technical documentation scattered across multiple repositories to deliver timely and accurate answers to customers. With thousands of documents, some of which are hundreds of pages long, the search for relevant and specific content is not just daunting but also time-intensive.

Intel's portfolio consists of thousands of innovative products, and teams are continuously producing and updating material on those products, technologies, services, methods, and solutions. This content is dispersed across many domains of knowledge—a vast landscape of information stored in diverse formats and locations. The answer to a client's question might be hidden within dense technical manuals, unstructured web portals, or scattered emails. The inability of a salesperson to quickly locate accurate and relevant information can lead to customer dissatisfaction. The sprawling domains of knowledge have become a critical bottleneck, impeding the speed and precision of problem-solving.

To give our sellers, partners, and customers the benefit of that vast landscape of knowledge, we envisioned a digital assistant that could provide complete, accurate, and attributable answers that the sellers can trust. It would require a platform that can democratize knowledge by sifting through complex data in diverse locations and providing precise information, complete with source attribution.

Utilizing an LLM through a custom retrieval-augmented generation (RAG) architecture, we developed VTA, a GenAI chatbot that returns instant, accurate, summarized solutions for sellers to bring to their customers. VTA uses emerging AI capabilities to elegantly solve a real-world problem.

Retrieval-Augmented Generation

By leveraging a combination of industry-standard large language models (LLMs) and a customer RAG framework, we created VTA, a GenAI chatbot that provides instant, precise, and concise solutions for our sales representatives to present to their clients. VTA addresses this real-world challenge by harnessing the power of emerging AI technologies.

RAG Workflow

1. The user provides an input prompt (i.e., a question).
2. The RAG system queries an internal knowledge base (e.g., a database or vector index) to retrieve all pieces of information relevant to the prompt.
3. The retrieved information is included as additional context with the user's question as an LLM prompt.
4. The LLM generates output that is grounded in the retrieved information.

Key Benefits of RAG Architecture

- **Factual accuracy:** The retrieved information helps ground the LLM output in up-to-date facts and domain knowledge.
- **Reduced hallucinations:** Explicit context reduces the LLM's tendency to generate plausible but incorrect statements.
- **Efficient customization:** Retrieving domain-specific knowledge allows easy adaptation to new domains and use cases without expensive fine-tuning.
- **Recency of Information:** By using the latest data repository, the answers are always the most up to date and accurate in comparison to the 6- to 12-month-old data existing in the LLM.

By utilizing RAG to pull information from a vast array of dispersed data sources across the company, the VTA chatbot has become a powerful new tool for Intel sales information and solution discovery.

An Overview of the Virtual Intel Technical Assistant (VTA)

The VTA chatbot combines RAG with a generative pre-trained transformer (GPT) LLM to generate direct and full answers to any question a user might have on Intel products. When asking a question, users do not need to specify the location of relevant documents from which answers will be derived.

In addition to a generated answer, VTA provides both exact content sections and direct links to the original full documents used to generate the answers, giving the user the ability to easily access and read the source material to evaluate the quality of generated answers and find relevant documentation to support customer solutions. The direct content links are also categorized as Public, non-disclosure agreement (NDA) required, Restricted, and Restricted-Secret classifications, enabling sellers to share only appropriate information based on specific customer relationships and agreements. VTA is designed to support various GenAI large language models, though its initial implementation relies on GPT-4 accessed by the OpenAI API.

Designing the VTA

The VTA platform design prioritizes accuracy and completeness. In addition, the platform is designed to support a variety of Intel business organizations, some of which were concerned about data security and hesitant to use public cloud solutions.

Portable and Scalable Architecture

In addition to being portable, VTA needed to be scalable. Though the platform began with just a few hundred users, it soon expanded to 1,000 users and needed to be able to eventually scale to up to 20,000 users. Combining this requirement with a need for portability made building on Kubernetes an obvious choice. Kubernetes is a container orchestration platform designed around automation, management, and scaling, meaning that VTA could also be deployed in any private data center or public cloud that supports Kubernetes-based solutions.

Intel Hardware

Intel offers advanced hardware solutions for large-scale applications of this sort. The platform is built with Intel hardware, running on Intel® Xeon® processors with Intel® Accelerator Engines. Apart from calls to OpenAI's cloud-based GPT-4 API for summarization, VTA is built entirely on Intel hardware.

Calls to external public LLM models are optional, and can be replaced with internal, private, fine-tuned LLMs as needed. A future version of VTA that will run exclusively on Intel hardware, including Intel® Gaudi® 3-based hardware for fine-tuning open-source models like Llama, StarCoder, or Mixtral is currently in development.

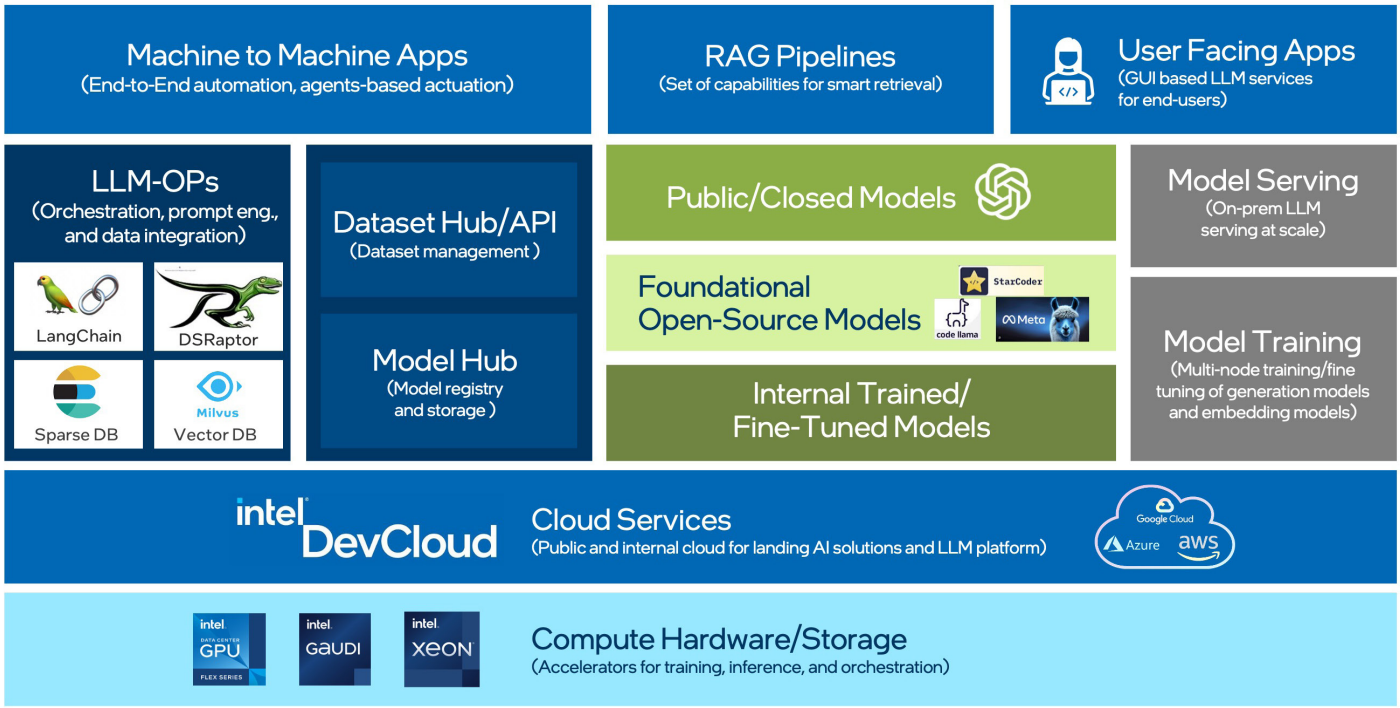

Figure 1. The VTA platform architecture consists of four logical layers.

VTA Architecture

As the architecture diagram shows, the VTA platform is composed of four logical layers.

1. At the bottom of the diagram, **Layer 1** contains compute and storage, including accelerators such as Intel Gaudi and other GPUs which tend to be heavily used in GenAI workloads.
2. **Layer 2** is cluster management, consisting of Kubernetes.
3. **Layer 3** includes several LLM-Ops capabilities such as batch and online workflow orchestration, data integration, dataset management, and a registry for storing and managing internally trained or fine-tuned models.
4. **Layer 4** contains the Gen AI applications and use-cases that run atop the first three layers.

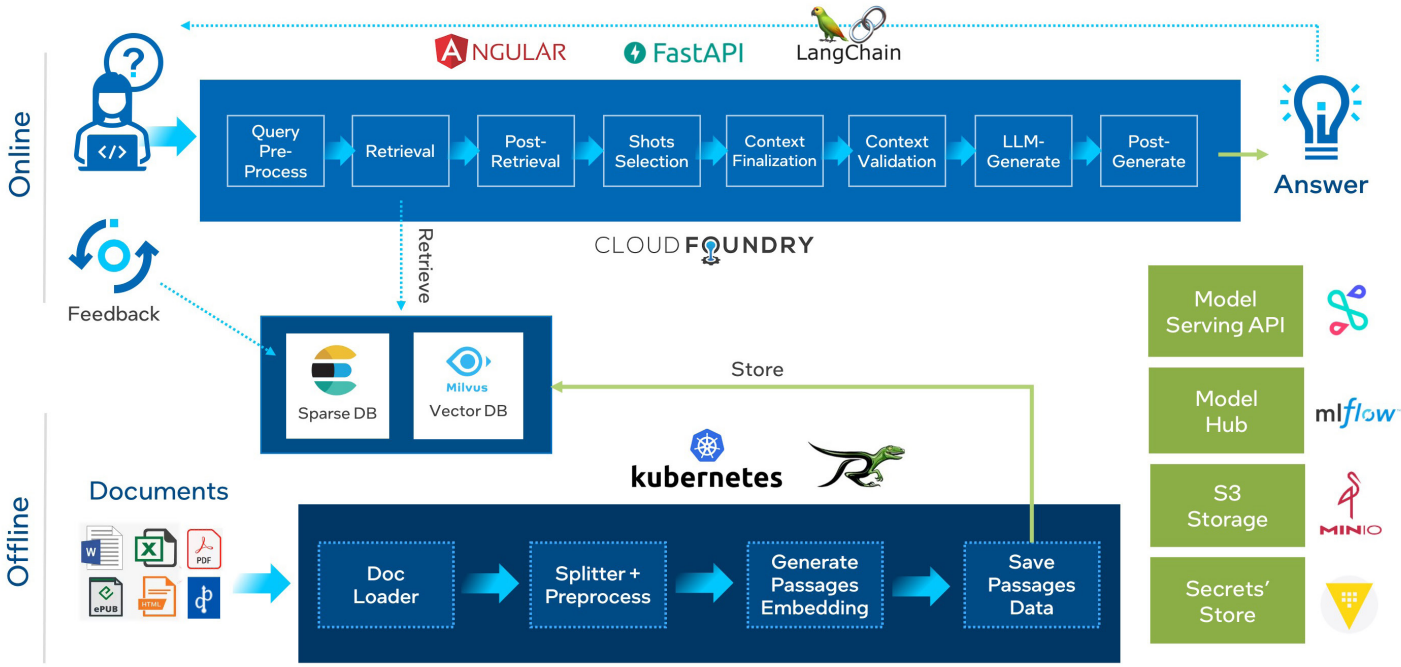

Figure 2. The RAG system is populated from diverse data sources.

Populating and Managing RAG

The platform can work with a variety of data sources, including documents, code, and database data. This diagram illustrates the entire data processing flow, including an offline process to load and pre-process documents, a set of databases to store the processed data, and an online process which enables user interaction. The entire flow is supported by a set of common services.

Offline pre-processing is optimized for RAG, which is responsible for loading documents into databases and updating them as needed. Pre-processing includes intelligent parsing and chunking, allowing for both sparse (text-based) and dense (semantic) searches.

Three different databases are used for different features. Elasticsearch is used to store sparse data and to conduct fast text searches, while Milvus is used for large-scale embedding (vector) storage. Finally, MongoDB stores structured or unstructured documents in a JSON-like format. All three databases are built with a scale-out architecture to handle large volumes and the ability to scale vertically or horizontally to accommodate growing data loads.

The online process is implemented as a set of APIs that are connected to the user interface. The main API is a LangChain sequential flow that implements all required data processing steps such as pre-retrieval, retrieval, ranking, context creation, generation, and post-generation logic. The system can run and manage many flows with different configurations to support various use cases and domains.

Common reusable services include the model serving APIs, the model hub, S3 compatible storage, and a secrets store. The serving API can make any model which is registered in the model hub available for inference as an endpoint. This API is used for serving the embedding model and can also be used to serve other generation models on-premises, if required.

Privacy and Security

Data privacy and security are paramount, and GenAI addresses these concerns by operating entirely within the client's security perimeter. By maintaining data isolation, Intel ensures that sensitive information remains protected. This approach aligns with responsible AI practices and builds trust with clients.

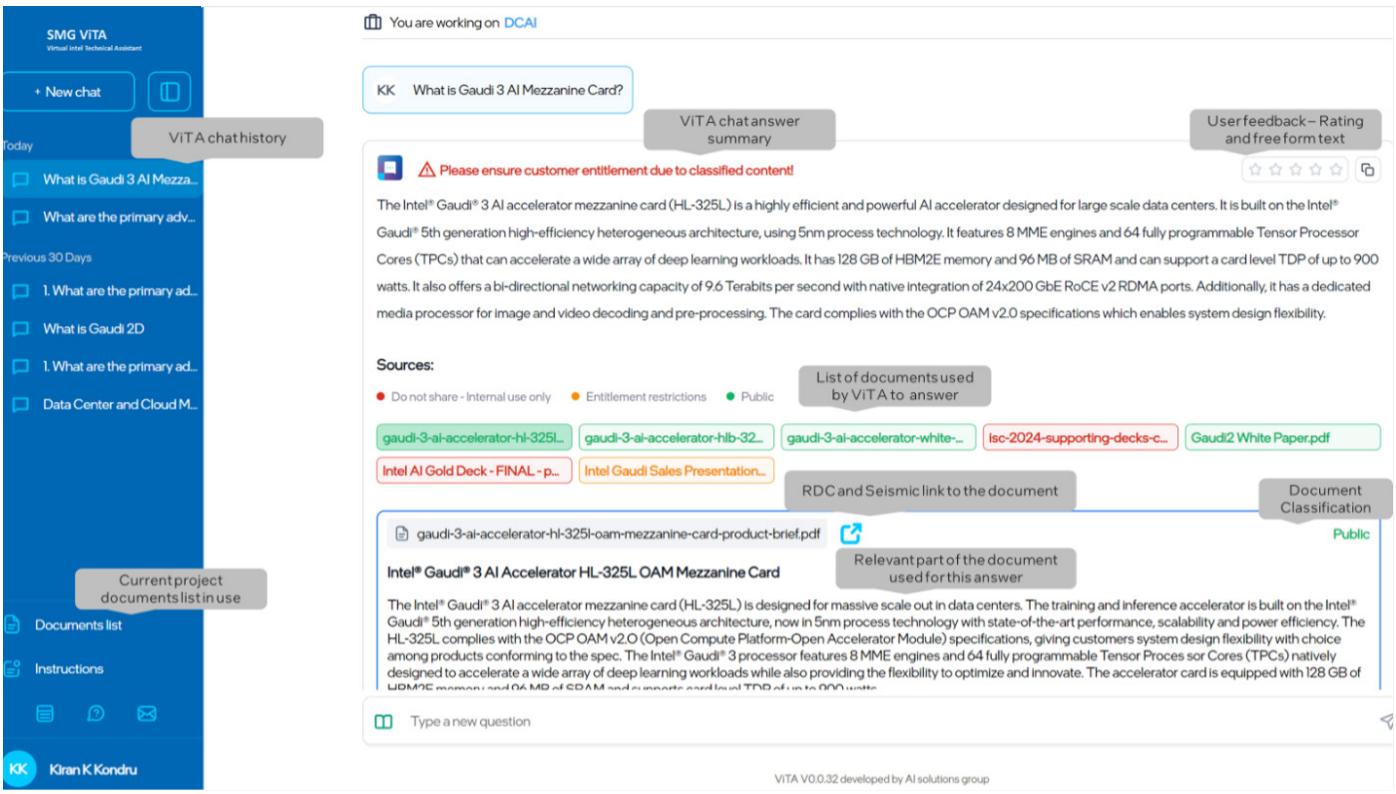

Figure 3. The VTA user interface prioritizes simplicity and access to relevant sources.

Improving VTA's Answers

The VTA UI strives to maintain a simplicity that matches the public ChatGPT interface. This should make it both easier to use and increase adoption by new users. Users can input any question into the UI. Answers appear, as well as attribution—a list of the relevant content sections the results utilized. It is very simple for users to read original documents and references to evaluate the quality of the GenAI-produced answers.

Users are then able to rate the quality of answers by using the Star based scoring options, allowing continual improvement of VTA's generated answers by evaluating user feedback and fine-tuning the model behind the scenes. The goal is to identify high-quality answers to help build a gold standard of GenAI-produced content, thus allowing even better fine-tuning.

The Future of VTA

VTA started supporting a few hundred users, but, in the six months since its launch, has grown to about 5,000 SMG and BU users. Adoption continues to increase. Thanks to the platform's scalability, it has the capacity to serve many more users. The current goal is to grow usage to 20,000 users.

Future enhancements are geared towards integrating VTA with a broader array of data sources and supporting a wider range of products and technologies, thereby enriching the context and depth of the solutions and answers it can provide. By tapping into ticketing systems, CRM platforms, and product design roadmaps, VTA will be able to offer a more comprehensive understanding of each customer's unique situation and needs. VTA will be able to anticipate issues and provide proactive support, potentially even before the seller recognizes the need to ask a question. Furthermore, the integration of VTA within business process applications promises a seamless and intuitive experience for sellers, allowing them to access critical information with unparalleled speed and efficiency.

Business Impact

VTA has made a substantial business impact just by supporting its existing user base. It has already shown the following positive results:

- **High Adoption:** With over 1,000 users, the chatbot has become a go-to resource for technical queries.
- **Instant Answers:** It has reduced the time to respond to technical questions from hours or days to instantaneously.
- **Faster Design-In:** VTA has accelerated sales cycles by enabling quicker support during the selection and integration process.
- **Enhanced Customer Engagement:** VTA has improved satisfaction by providing timely and accurate information, with attribution to give the user confidence.
- **Increased Seller Productivity:** Sellers can now address complex issues more efficiently, enhancing overall productivity.
- **Knowledge Democratization:** By breaking down silos of expertise and revealing pockets of knowledge, the project has enabled wider access to shared information.
- **Easier Access to Data:** The VTA platform has streamlined the process of obtaining necessary technical details.
- **Content Creator Feedback:** By highlighting content gaps for creators based on unanswered queries, VTA is self- improving through a constant feedback loop.
- **Visibility into Design Challenges:** VTA offers insights into common questions and issues with specific products.

In addition, VTA has begun to shift the focus of technical support offered by technical sellers. Technical sellers are able to more easily find deep and rich technical answers, enabling them to handle dramatically more complicated queries from their customers with increased confidence.

With VTA, sales will be able to address every question arising from systems, software, industry, use-cases, workloads, ecosystem, and real customer problems. This streamlined and enhanced approach not only elevates the customer experience but also informs content development and product support strategies, driving Intel's growth and competitive edge.

Conclusion

VTA exemplifies Intel's commitment to innovation and customer-centric solutions, maximizing Intel technologies. By harnessing the power of generative AI, Intel equips its sellers with a powerful tool to better serve customers and accelerate sales cycles. As the GenAI journey continues, Intel remains at the forefront of AI-driven transformation.

Learn More

- Read "Building Blocks of RAG with Intel"
- Explore more resources about generative AI