## Case Study

Edge AI
Intel® Core™ Processors, Intel® Core™ Ultra Processors, Intel Atom® Processors
Intel® Arc™ Graphics
Intel® Extensions for PyTorch (IPEX)
Intel® Distribution of OpenVINO™ Toolkit

# Revolutionizing Pharmacy Retail Operations with DFI's AI-Driven Kiosk

**DFI's AI-powered kiosk leverages large language models (LLMs) and chatbots to enhance customer experience and streamline manpower utilization in pharmacy retail environments.**

The retail industry is undergoing a profound shift driven by rapid technological advancements and changing consumer expectations. With global retail sales projected to exceed $30 trillion in 2024[1], consumers are increasingly seeking faster, more convenient, and personalized shopping experiences both online and offline. Customer experience has emerged as the primary differentiator and a significant driver of growth, with experience-led businesses enjoying higher brand awareness[2], increased customer satisfaction rates[2], and stronger customer retention[2].

This transformation is equally critical within the retail industry, particularly the healthcare retail sector, particularly pharmacies. Pharmacies play a vital role in dispensing medications and offering healthcare advice but now face heightened pressure to innovate and enhance their services. Customers are more willing than ever to switch brands after a single bad experience, compelling pharmacies to reevaluate how they operate, serve customers and manage their supply chains to stay competitive.

However, pharmacies face significant challenges:

- **Staffing Shortages:** A global shortage of healthcare professionals makes it difficult for pharmacies to maintain adequate staffing levels, especially in rural and underserved areas.
- **Rising Customer Demand:** An aging population and the increasing prevalence of chronic diseases have led to a significant rise in prescription volumes. According to the World Health Organization (WHO)[3], the global burden of chronic diseases is expected to continue rising, putting more pressure on pharmacies to handle larger volumes while maintaining accuracy and timeliness.
- **Operational Inefficiencies:** Reliance on outdated technology and manual processes makes many traditional pharmacy systems inefficient and prone to errors, leading to long wait times and service delays during peak hours.

To meet elevated customer expectations and address these challenges, pharmacies are increasingly turning to automation and AI-driven technologies. These solutions have the potential to revolutionize pharmacy operations by reducing manual workloads, improving accuracy, and enhancing the overall customer experience. From automated dispensing systems to AI-powered chatbots, such technologies provide real-time assistance and help pharmacies operate more efficiently, positioning them to thrive in the evolving retail landscape. One such pharmacy retail store is BenQ Healthcare Tin Tin Smart Pharmacy in Taiwan.

## DFI's AI-driven Service Kiosk Solution in Action

As BenQ Healthcare Tin Tin Smart Pharmacy seeks to expand its mission of providing high-quality health management services—particularly focusing on the elderly—they are turning to technology-based solutions to enhance the customer experience at scale. Tin Tin Pharmacy, now part of the Qisda Group, operates 92 stores nationwide and aims to expand to 100 stores by the end of 2024[4]. By integrating advanced technologies such as digital service platforms, the Internet of Things (IoT), big data, and AI, BenQ Healthcare Tin Tin Smart Pharmacy is actively transforming to seize opportunities in the healthcare market.

DFI's AI-driven Service Kiosk offers a seamless, automated platform that allows customers at BenQ Healthcare Tin Tin Smart Pharmacy to access essential pharmacy services 24/7*, eliminating the need for staff intervention for routine inquiries and product recommendations. This kiosk brings the convenience of digital technology to the forefront of pharmacy services, enhancing customer experience by leveraging advanced AI and real-time processing capabilities.

The kiosk's AI Chatbot, powered by an advanced language model, is capable of engaging in natural conversations with customers, guiding them through the pharmacy's services, offering tailored over-the-counter medication recommendations, and assisting with order placement. This enhances the customer experience and frees up staff to focus on more complex, high-value tasks, ensuring smooth operations in busy pharmacies. Moreover, the kiosk is equipped with a secured prescription management and dispensing system, ensuring that customers can receive their medications safely and accurately. The secure dispensing system maintains the highest standards of privacy and security, which are critical for healthcare services.

### Key Components of the AI-Driven Service Kiosk

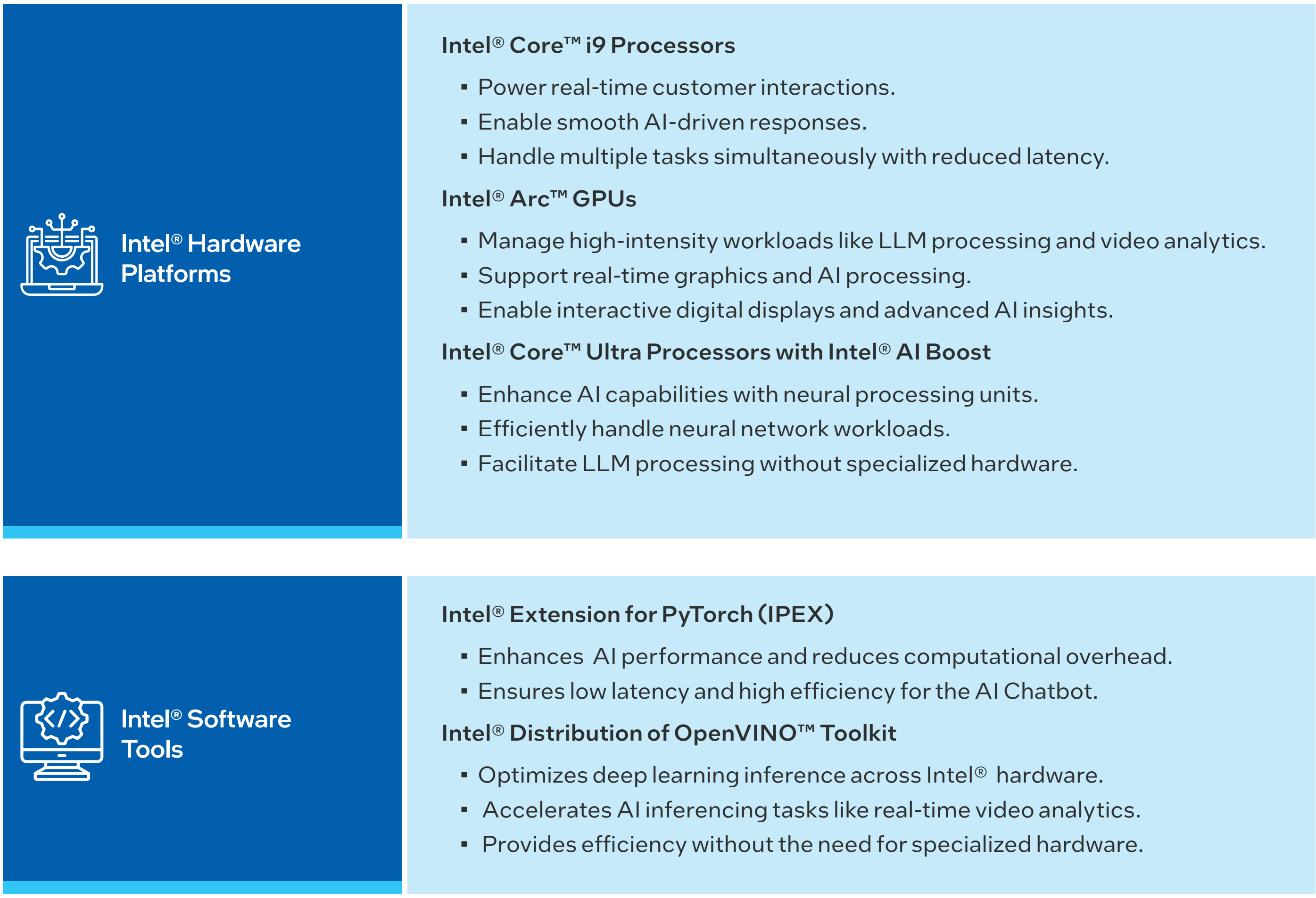
- **AI Chatbot:** Provides personalized customer interaction by processing requests, answering queries, and managing prescriptions and product recommendations through a user-friendly interface.
- **Dashboard & Smart Ads:** Offers AI-driven insights for operators to optimize services and uses smart advertisements to deliver personalized promotions based on customer interactions.
- **Remote Management:** Enables real-time monitoring and troubleshooting to reduce downtime and maintenance costs, allowing efficient management without on-site staff.
- **DFI In-House Management Module:** Seamlessly integrates hardware and software management for robust performance and easy maintenance, streamlining pharmacy operations.
- **Payment System:** Features an AI-enhanced, cashless system supporting multiple payment methods for swift, secure, and convenient transactions.
- **24/7 Operation*:** Designed for continuous service availability, providing access beyond standard pharmacy hours.

By adopting DFI's AI-driven Service Kiosk, BenQ Healthcare Tin Tin Smart Pharmacy is poised to enhance its service efficiency and quality. This technological integration allows them to provide personalized and convenient medication purchasing experiences, aligning with their commitment to improving the quality of life for the elderly and offering high-quality health management services.

(*Varies from region to region)

## Solution Architecture

DFI's AI-driven service kiosk leverages a powerful combination of Intel's cutting-edge hardware and software technologies, ensuring a seamless, real-time customer experience while maintaining flexibility, scalability, and security.



**Intel® Hardware Platforms**

**Intel® Core™ i9 Processors**

- Power real-time customer interactions.
- Enable smooth AI-driven responses.
- Handle multiple tasks simultaneously with reduced latency.

**Intel® Arc™ GPUs**

- Manage high-intensity workloads like LLM processing and video analytics.
- Support real-time graphics and AI processing.
- Enable interactive digital displays and advanced AI insights.

**Intel® Core™ Ultra Processors with Intel® AI Boost**

- Enhance AI capabilities with neural processing units.
- Efficiently handle neural network workloads.
- Facilitate LLM processing without specialized hardware.

**Intel® Software Tools**

**Intel® Extension for PyTorch (IPEX)**

- Enhances AI performance and reduces computational overhead.
- Ensures low latency and high efficiency for the AI Chatbot.

**Intel® Distribution of OpenVINO™ Toolkit**

- Optimizes deep learning inference across Intel® hardware.
- Accelerates AI inferencing tasks like real-time video analytics.
- Provides efficiency without the need for specialized hardware.

**Figure 1.** The system architecture of DFI's AI-driven service kiosk at a glance

## Performance and Testing

DFI's AI-driven service kiosk has undergone rigorous testing to ensure optimal performance, particularly in high-traffic environments. The system is designed to manage multiple customer interactions simultaneously, a necessity for busy retail pharmacies where efficiency and speed are paramount.

### LLM Inference Efficiency

One of the key components of the AI-driven service kiosk is the AI-powered chatbot, built on Large Language Models (LLMs) that require efficient inference for real-time interactions. Traditionally, LLMs face challenges due to their complex model structures and massive computational demands, which can introduce latency during inference. However, through advancements in LLM architecture and optimization, DFI's AI-driven service kiosk has overcome these obstacles.

Using a simplified LLM decoder layer and optimized data movement strategies, the solution has significantly reduced memory access frequency[5], lowering system latency[5]. The integration of a segment KV cache policy ensures efficient memory management, separating request and response tokens into distinct memory spaces, which increases the runtime batch size and improves throughput. By deploying these techniques on Intel® hardware, particularly Intel® Arc™ GPUs, the system delivers high-speed performance while handling complex interactions.

### LLM Performance Benchmarks on Intel® Core™ Ultra Processors

Tests were conducted on Windows 11 with Intel® Core™ Ultra processors, utilizing LLMs ranging from 6 billion to 13 billion parameters. With 1,024 input tokens and a batch size of 1, the optimized solution achieved notable reductions[6] in next-token latency. This performance enables real-time LLM inference on client hardware such as laptops and desktops, making it suitable for the kiosk's requirements.

### LLM Performance Benchmarks on Intel® Arc™ GPUs

Similarly, tests on Intel® Arc™ GPUs demonstrated significant reductions[7] in next-token latency for LLM inference. Conducted on Ubuntu 22.04 with 1,024 input tokens and a batch size of 1, the system effectively handled models ranging from 6 billion to 13 billion parameters. This ensures the kiosk can manage complex customer interactions seamlessly.

| Model Category | Model Name | Precision | Batch Size | Latency (ms) |
|---|---|---|---|---|
| **Text Prediction** | GPT-2 | FP16 | 1 | **210** |
| **Question Answering** | BERT-Large-Uncased-Whole-Word-Masking-SQuAD | FP16 | 1 | **45** |

Table 1: Inference Performance on Intel® Arc™ GPUs[7]

By leveraging the IPEX-LLM library—a PyTorch library optimized for running LLMs on Intel® CPUs and GPUs—the kiosk achieves state-of-the-art performance in LLM inference. The library includes optimizations such as low-bit (INT4, FP4, INT8, and FP8) LLM accelerations and seamless integration with community libraries like Hugging Face, LangChain, LlamaIndex, and vLLM.

These benchmarking results confirm that DFI's AI-driven service kiosk can provide seamless, real-time AI-powered interactions with customers. The integration of optimized hardware and software solutions ensures high efficiency and responsiveness, enhancing the overall user experience while maintaining the performance required for complex AI tasks.

### Real-Time Interaction & Latency

The solution's low latency and high throughput were achieved by combining Intel® Arc™ GPUs and Intel® Core™ i9 processors with a customized Scaled-Dot-Product-Attention kernel designed to reduce computational overhead. When compared with standard implementations like HuggingFace, the kiosk's optimized solution achieved up to 7x lower token latency[5] and 27x higher throughput[5] on popular LLMs.

### Peak Performance with Intel® Hardware

Powered by Intel® Core™ i9 processors, an integrated neural processing unit (NPU) called Intel® AI Boost, and supported by Intel® Extension for PyTorch (IPEX), the AI-driven kiosk has consistently maintained high performance during real-time AI inference tasks.

The system demonstrated the ability to handle simultaneous customer interactions, managing complex prescription requests, customer queries, and medication dispensing without any performance bottlenecks. This reliability is crucial in real-world applications, where pharmacies may experience bursts of customer demand.

## Business Impact

### Improved Customer Experience

DFI's AI-driven service kiosk has been designed to enhance the customer journey at BenQ Tin Tin Smart Drugstore by automating routine interactions, which significantly reduces wait times and improves the overall customer experience. The AI Chatbot provides personalized product recommendations based on the customer's history and needs, offering a tailored experience that keeps customers engaged. This seamless interaction leads to greater customer satisfaction, increasing loyalty, repeat visits, and boosting customer retention. The self-service nature of the kiosk ensures that customers receive real-time assistance, even during high-traffic periods when human staff might be unavailable, thus creating a smoother, more efficient pharmacy experience.

### Cost Savings

The automated nature of DFI's AI-driven service kiosk reduces the need for additional staff during peak operating hours, minimizing operational costs for BenQ Tin Tin Smart Drugstore. Pharmacies no longer have to worry about hiring additional team members to handle routine inquiries or prescription orders, allowing them to focus resources on more critical tasks such as complex medication consultations or inventory management. The kiosk's capability to handle multiple customers simultaneously means pharmacies can serve more people without needing to physically expand their facilities. This leads to optimized resource utilization, enhanced staff productivity, and overall cost reduction.

### Revenue Generation

The AI-driven operations of DFI's AI-driven service kiosk at BenQ Tin Tin Smart Drugstore offer opportunities for increased revenue through upselling and cross-selling. By analyzing customer queries and purchase histories, the kiosk can recommend related or supplementary products, thus boosting average transaction value. For example, a customer picking up cold medicine may also receive recommendations for vitamins or personal care items. This personalized product recommendation not only increases the likelihood of additional sales but also encourages customers to explore other services or products offered by the pharmacy, creating a more comprehensive shopping experience that benefits both the customer and the business.

## Future Applications

DFI's AI-driven service kiosks are not limited to pharmacy settings—they represent scalable and versatile solutions with the potential to transform various sectors. For example, DFI is also focusing on revolutionizing electric vehicle (EV) charging stations by transforming them into multifunctional hubs. By integrating Intel® Arc™ GPUs, DFI's EV chargers are equipped with advanced digital interfaces and AI-driven features, effectively turning charging stations into intelligent kiosks. This enhancement not only increases the functionality of the charging stations but also improves accessibility and user interaction, making the charging experience more engaging and efficient.

As AI and machine learning technologies evolve, the functionality of the kiosk can be further expanded to include advanced healthcare services. This could involve telemedicine consultations, allowing customers to speak directly with licensed healthcare providers from the kiosk, or remote prescription renewals, enabling customers to renew medications without an in-person visit. Furthermore, kiosks could be integrated into healthcare facilities for automated health check-ups, giving patients easy access to vital services like blood pressure monitoring, diabetes management, or routine health screenings. In retail environments ranging from supermarkets and convenience stores to department stores and specialty shops, the AI capabilities of the kiosk could be leveraged for personalized marketing. Stores can deliver targeted promotions and advertisements based on individual customer preferences and behavior patterns, opening new avenues for revenue generation and customer engagement.

## Conclusion

DFI's AI-driven service kiosk, powered by Intel's advanced computing platforms and AI chatbot capabilities, marks a significant advancement in retail automation. By streamlining customer interactions—such as product inquiries, personalized recommendations, and transaction processing—the kiosk enhances service delivery, reduces wait times, and elevates the overall customer experience. It also lowers operational costs and boosts revenue through personalized upselling and cross-selling, making it a cost-effective modernization solution for retailers seeking to stay competitive.

## Learn More

Intel® Core™ Processors

Intel® Core™ Ultra Processors

Intel Atom® Processors

Intel® Arc™ Graphics

Inte® AI Accelerators

Intel® Industry Solution Builders

**Sources:**
1: https://www.invespcro.com/blog/online-retail-statistics-and-trends-for-2024/
2: https://www.forrester.com/blogs/cx-index-2019-results/
3: https://www.who.int/news-room/fact-sheets/detail/diabetes
4: https://www.taiwan-healthcare.org/en/news-detail?id=0sa08ovebffaiyzb
5: https://arxiv.org/pdf/2401.05391
6: https://www.intel.com/content/www/us/en/developer/articles/technical/accelerating-language-model-inference-on-your-pc.html
7: https://www.intel.com/content/www/us/en/content-details/817734/unlocking-the-ai-power-of-intel-arc-gpu-for-the-edge-a-deep-dive-into-hardware-and-software-enablement.html

intel®

Intel is committed to respecting human rights and avoiding complicity in human rights abuses. 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 a violation of an internationally recognized human right.

Performance varies by use, configuration, and other factors. Learn more at www.Intel.com/PerformanceIndex

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

Your costs and results may vary.

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

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

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