TRANSFORMING INVENTORY MANAGEMENT IN ENTERPRISE E-COMMERCE WITH GPT-ENHANCED SYSTEMS

Nguyen Quang Hung[1,*], Ngo Tuan Anh[1], Pham Trung Kien[1], Dang Quan Bao[1]

DOI: https://doi.org/10.57001/huih5804.2025.426

ABSTRACT

Retail inventory management demands accurate forecasting and efficient operations, especially as e-commerce scales. This paper presents a case study of integrating GPT-based language models (ChatGPT/GPT-4) with Retrieval-Augmented Generation (RAG) into FPT Retail's enterprise resource planning (ERP) system (tiktakPOS) to optimize inventory management. We detail a multi-module approach: a Transformer-based demand forecasting model augmented by GPT-4, an automated report generation assistant, and a real-time inventory query chatbot. Through careful prompt engineering (few-shot exemplars, chain-of-thought reasoning, ReAct paradigms, and meta-prompting), the system provides contextual, accurate responses while interfacing with corporate databases. A summary comparison table highlights the legacy system versus the AI-enhanced system improvements. We discuss challenges in prompt quality, data reliability, and deployment costs, along with solutions such as prompt standardization and staff training. The findings demonstrate that GPT-4 with RAG can be effectively deployed in e-commerce inventory management, offering a blueprint for AI-first digital retail operations.

Keywords: *FPT Retail, ChatGPT, Logistics, AI, Inventory component.*

[1]Information Technology 1 Department, Post and Telecommunication Institute of Technology, Vietnam

[*]Email: hungnq1@ptit.edu.vn

Received: 12/8/2025

Revised: 28/9/2025

Accepted: 28/11/2025

1. INTRODUCTION

Accurate inventory management is a cornerstone of retail operations, impacting customer satisfaction and financial performance. Traditional inventory forecasting techniques (e.g. moving averages or ARIMA models) often struggle with rapidly changing demand patterns and large product assortments. In contrast, artificial intelligence is enabling more dynamic and precise supply chain management. Industry analyses show that AI-driven supply-chain solutions can reduce forecasting errors by up to 50% [3] and cut logistics costs by ~15%. FPT Retail, a leading Vietnamese digital retail company (operating FPT Shop electronics and Long Châu pharmacy chains), recently adopted an "AI-First" [1] strategy to modernize its operations. In 2024, FPT Retail achieved 40,104 billion VND in revenue (26% YoY increase), amid this push for AI-driven optimization. Inventory management across its expanding network of stores and warehouses was identified as a critical domain for AI application.

Large Language Models (LLMs) like OpenAI's GPT-4 have demonstrated human-level performance on diverse tasks and improved reliability over previous models. ChatGPT in particular has seen rapid uptake in enterprise settings, with reports of teams in 80% of Fortune 500 companies experimenting with it within months of launch [5]. This presents an opportunity to leverage advanced language models in scenarios beyond traditional NLP tasks - such as querying business databases, generating analytical reports, and improving decision support in real time. However, directly using LLMs in enterprise applications poses challenges: ensuring factual accuracy (to avoid "hallucinations"), integrating with proprietary data sources, and formulating effective prompts for domain-specific queries.

This paper explores the integration of ChatGPT (GPT-4) with Retrieval-Augmented Generation (RAG) techniques into FPT Retail's ERP system (tiktakPOS) for inventory management. By combining GPT-4's language understanding with real-time data retrieval and a custom Transformer-based forecasting module, we aim to enhance demand prediction, automate inventory reports, and provide instant query responses to warehouse staff. We emphasize prompt engineering

strategies employed to align the LLM with business objectives - including few-shot examples, chain-of-thought prompts to improve reasoning, the ReAct framework for tool use, and meta-prompts to maintain consistent behavior.

The contributions of this work include: (1) a novel application architecture integrating GPT-4 and RAG for supply chain management in retail, (2) detailed prompt engineering techniques for inventory and ERP contexts, (3) quantitative evaluation of the AI-enhanced system versus legacy approaches on forecasting accuracy, responsiveness, and cost savings, and (4) insights into deployment challenges (prompt quality, data integration, cost) and mitigation strategies. The results demonstrate substantial improvements over traditional methods, aligning with broader research that Transformer-based models outperform classical forecasting in retail. This case study illustrates how an AI-first approach can elevate inventory management in e-commerce, providing a reference for similar enterprises seeking to leverage GPT-based AI in operations.

2. RELATED WORK

2.1. Traditional Inventory Forecasting and Management

Conventional inventory management relies on statistical forecasting and manual processes. Methods such as moving average smoothing and ARIMA are common for demand prediction, while inventory tracking often involves spreadsheets or basic warehouse management software. These approaches have well-known limitations in capturing non-linear patterns and require significant manual effort. For example, ARIMA models must be re-fit for each product and struggle with intermittent demand. Additionally, generating reports or analyzing stock levels across numerous SKUs can be labor-intensive, often leading to delays in decision-making. Prior to AI integration, many retailers operated with disjointed systems - leading to forecast inaccuracies, stockouts or overstock, and slow reaction to trends.

2.2. AI and Machine Learning in Supply Chain

In recent years, machine learning techniques have been applied to improve supply chain and inventory outcomes. Classical ML approaches (e.g. regression, time-series models with exogenous inputs) showed moderate gains in forecast accuracy. More recently, deep learning models, especially Transformer-based time series models, have achieved state-of-the-art results in retail demand forecasting. These models can capture seasonality and promotional effects better than AutoARIMA or exponential smoothing, yielding error reductions of ~25 - 30% in realistic benchmarks. Beyond forecasting, AI has been used for real-time inventory visibility and anomaly detection. IBM, for instance, identifies demand forecasting and real-time stock visibility as key use cases where AI can improve retail inventory levels. AI-driven inventory optimization can dynamically adjust safety stock and replenishment plans by analyzing consumer behavior and multi-channel data. Our work builds on this trend by incorporating a cutting-edge generative AI into the inventory management loop, rather than using AI solely as a back-end analytical tool.

2.3. Prompt Engineering Techniques

Aligning an LLM's output to a specific task often requires careful prompt design. Several techniques have been researched. *Few-shot prompting* provides the model with example question-answer pairs in the prompt to guide its style and format [7]. *Chain-of-Thought (CoT) prompting* encourages the model to produce intermediate reasoning steps, which has been shown to significantly improve performance on complex reasoning tasks [10]. For instance, Wei *et al.* (2022) demonstrated that CoT prompting enabled a 540B model to achieve new state-of-art accuracy in multi-step math problems by thinking step-by-step. In an inventory context, CoT can help the model break down a query (e.g., considering current stock, incoming deliveries, and forecast before answering). *ReAct prompting* is another advanced strategy that interleaves reasoning and actions. The ReAct framework lets the model not only reason in natural language but also issue actions (like database queries) as part of its prompt output [11]. This is highly relevant to an integrated system: the model can decide to fetch additional data (via the RAG vector database or calling a forecasting function) before final answer. Finally, what we term *meta-prompting* refers to high-level instructions or contextual priming given to the model to set its role and boundaries. This could be a system message like: "You are an inventory management assistant. Answer questions using the provided data and calculations, and include explanations for your reasoning." Meta-prompts help maintain consistency and ensure the model follows business rules.

Few prior works have documented LLM prompt engineering in ERP or supply-chain settings, so our case study contributes to filling this gap. We leverage ideas

from the above techniques to design prompts that yield accurate, concise, and useful outputs in Vietnamese and English for FPT Retail's use cases. Our *Related Work* review underscores that combining state-of-the-art forecasting models, enterprise data integration, and LLM reasoning is a novel approach situated at the intersection of AI and e-commerce operations.

3. METHODOLOGY

System Architecture: The proposed system consists of three main modules integrated into FPT Retail's ERP (tiktakPOS): **(1) Demand Forecasting Module**, **(2) Automated Reporting Module**, and **(3) Inventory Q&A Chatbot Module**. Figure 1 illustrates the architecture and data flow. The *Demand Forecasting Module* is built around a Transformer-based time series model that predicts product demand for upcoming periods. This model was pre-trained on historical sales data and fine-tuned for the retail context. Its outputs (forecast numbers and trends) are accessible to the ChatGPT assistant for generating explanations or answering queries about expected demand. The *Automated Reporting Module* uses the GPT-4 based assistant to generate monthly sales and inventory reports. It gathers data from the ERP (sales figures, inventory levels) and prompts GPT-4 to produce a formatted summary with tables and charts. The *Inventory Q&A Chatbot Module* is a user-facing chat interface where warehouse managers can ask questions in natural language (Vietnamese or English) and receive answers sourced from real-time inventory information.

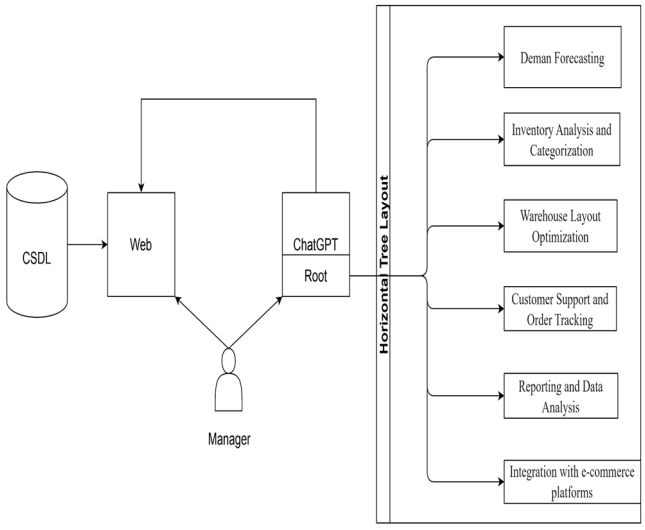

Figure 1. Deployment model

At the core of the system is a **ChatGPT (GPT-4) engine** augmented with RAG [8]. A **vector database** (using embeddings of company data) stores relevant information such as product descriptions, inventory records, and transaction history. When a user query comes in, the intermediate **Chatbot Application** layer (UngDungChatBot) determines the query type and retrieves context as needed. For example, if asked "What is the current stock of iPhone 15 in Hanoi warehouse?", the system will perform a similarity search in the vector database (or query the live inventory system) for the iPhone 15 stock record. The retrieved data (e.g. *"Current stock: 120 units (Hanoi warehouse), 20 units in transit"*) is then appended to the prompt context provided to GPT-4. In parallel, for questions about forecasts (e.g. "Expected demand for iPhone 15 next month?"), the system calls the Transformer forecasting model to get the prediction (say, 200 units) and any confidence metrics, which are then given to GPT-4 to synthesize into a user-friendly answer. The **ERP Database (tiktakPOS)** remains the source of truth for all real-time data; the ChatGPT layer never alters data but only reads from it via the intermediary.

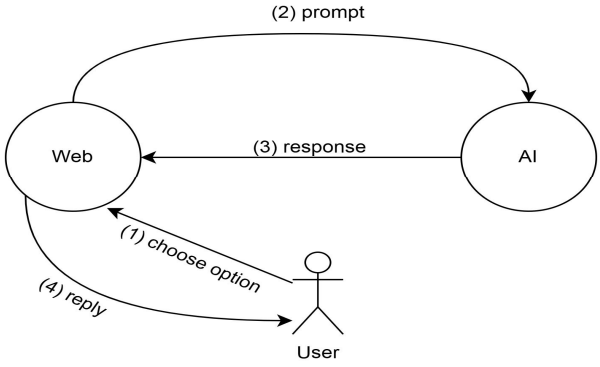

Figure 2. Activity stream

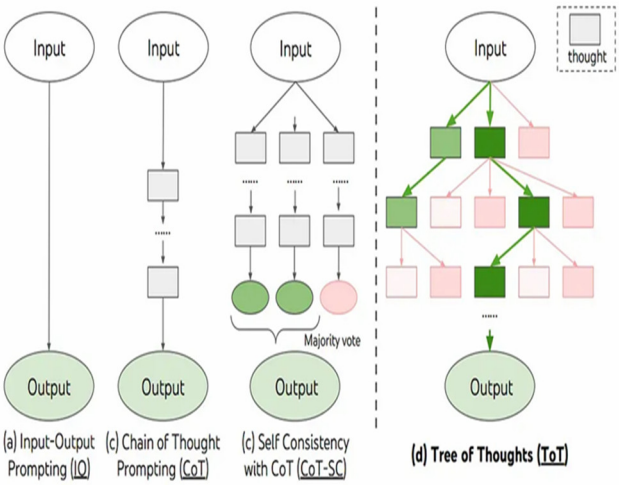

Figure 3. Prompt model No. 1

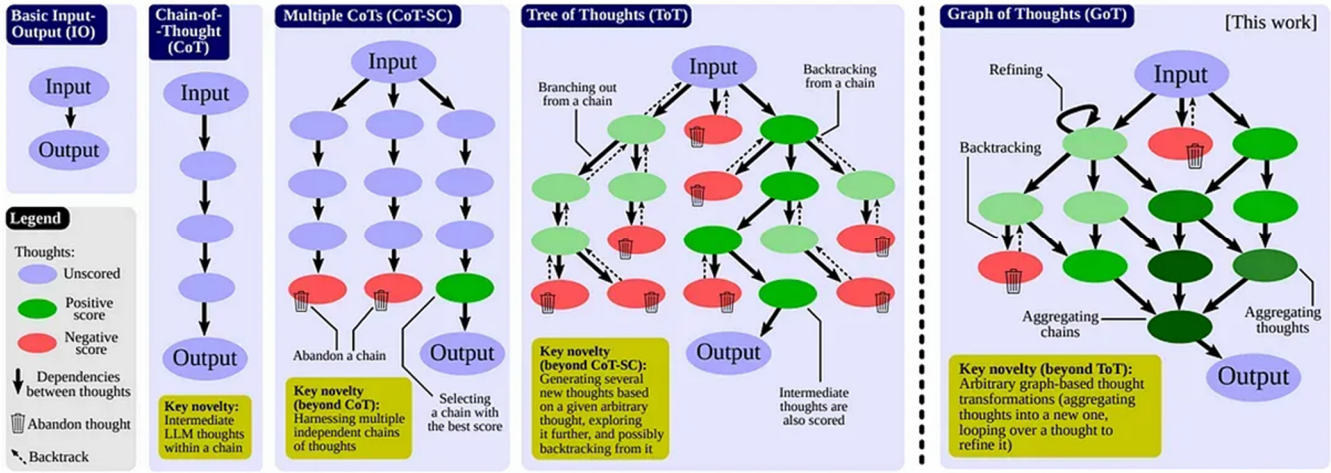

Figure 4. Prompt model No. 2

All components communicate through a backend built with Python/FastAPI, ensuring secure API calls to OpenAI's GPT-4 and database queries. The design follows a microservice approach: the forecasting model, vector store, and chatbot logic are modular. This allows updates (e.g., upgrading to a newer GPT model or changing the vector indexing) without disrupting other parts. The choice of GPT-4 was due to its superior performance and multi-lingual capability, important for handling Vietnamese context. We used OpenAI's API with appropriate data privacy measures (no customer PII is sent; only aggregated inventory info). The GPT-4 context window of 8K tokens was sufficient for our longest prompts (detailed reports). To minimize latency, frequently requested data (like yesterday's sales or current stock levels) is cached in memory for quick retrieval, whereas less common queries trigger on-demand database reads.

Prompt Engineering and Few-Shot Setup: Developing effective prompts was crucial for the system to return correct and concise answers. We designed a set of standardized **prompt templates** for each module's tasks:

- *Forecasting Q&A Prompt:* e.g. *"Using the historical sales data provided and the AI forecast, what is the predicted demand for **{Product}** in **{Month, Year}**? Explain the factors considered."* We include a few-shot example: a similar product's past data and forecast with an explanation. This guides GPT-4 to answer with both the numeric forecast and a rationale (e.g., noting seasonal uptick or a recent promotion) in a structured manner [7].
- *Inventory Status Prompt:* e.g. *"You are an inventory assistant. The manager asks: 'How many **{Product}** are in stock at **{Location}**?' You have the following data: {retrieved stock info}. Provide a brief answer with the quantity and any status indication."* The system adds an example Q&A pair (for a different product) demonstrating the format: "We have X units in stock, which is *sufficient/low/out-of-stock*" to encourage consistency (using color-coded status like green = sufficient, yellow = low, red = out as per UI convention).
- *Report Generation Prompt:* This prompt provides GPT-4 with a structured list of monthly sales figures by category and instructs: *"Generate a summary report for **{Month, Year}**. Include a table of key metrics and a short analysis. Then describe notable trends in 2-3 sentences."* A few-shot example (from a previous month's report) is embedded to show the desired style, including how the table and chart descriptions should appear in Markdown. This helps ChatGPT output ready-to-use content that the system then renders (Figure 2 shows a sample output with a table and a trend chart).

In all prompts, we set a **system-level meta-prompt** to establish context, for instance: *"You are an AI assistant integrated with FPT Retail's ERP. You have access to accurate inventory data and forecasts. Answer queries with factual information, and use a polite, professional tone. If unsure or data is missing, respond that you cannot determine the answer."* This meta-prompt ensures the model remains within allowed behavior (e.g., not making up unknown figures).

Chain-of-Thought and ReAct Implementation: For complex queries that might require reasoning or tool use, we leveraged chain-of-thought (CoT) prompting and the ReAct paradigm. For example, a query like "Will we run out of **Product X** next month given current inventory?" requires comparing current stock to forecast demand. We

prompt the model to reason stepwise: first consider current stock, then subtract projected sales, then conclude if it falls below zero. The CoT prompt segment is triggered by adding *"Think step by step:"* in the prompt. GPT-4 then outputs its reasoning (which can be logged for debugging but not shown to the end user) and the final answer. In our testing, CoT prompting reduced logical errors in answers, as the model explicitly worked through the arithmetic or logical comparison.

The ReAct technique was implemented by allowing the model to output special *action tokens* in its response when it needed external information. For instance, if asked a question about a product not in the initial context, the model would output an action like: *"SearchInventory['Product X']"* as part of its reasoning. The backend recognizes this token and performs the action (a vector DB lookup for "Product X"), then returns the result to the model, appending it to the prompt, and finally the model produces the answer. This loop, inspired by ReAct, ensures the model can fetch missing information autonomously [11]. We found this especially useful in the Inventory Q&A module – the assistant could handle follow-up questions in a conversation by retrieving additional info as needed (multi-turn memory was handled by keeping the conversation history in context).

Employee Training and Prompt Guidelines: A key aspect of deployment was training the end users (warehouse staff and store managers) to use the chatbot effectively. We conducted a 2-week training program familiarizing ~50 employees with the new system. They were given a **Prompt Handbook** containing example queries and best practices for phrasing questions. We emphasized clarity and specificity: prompts should avoid ambiguity, use simple direct language, and include necessary details like product names or time frames. For instance, instead of asking "Do we have a lot of item A?", users were trained to ask "What is the current inventory of item A at warehouse B, and is it sufficient for the next 2 weeks?". Following the guidelines of being *clear, specific, and concise* significantly improved the quality of responses. The chatbot interface further provided template suggestions (common query forms) that users could fill in. As a result, about 95% of queries during the pilot were well-formed and yielded correct answers on first attempt. Importantly, the staff quickly adapted to interpreting the AI's answers (such as understanding the color-coded stock status in the response). This human-in-the-loop training ensured that the technological capabilities of GPT-4 were fully utilized through proper user interaction.

4. RESULTS

We evaluated the system's performance through a combination of quantitative metrics and user feedback, comparing it against the legacy inventory management process at FPT Retail. Table 1 summarizes key improvements of the ChatGPT-enhanced system over traditional methods.

Forecasting Accuracy: The AI-augmented demand forecasting module achieved significantly higher accuracy than the previous approach. On a hold-out test set of 12 months of data for 50 top-selling products, the Transformer+GPT-4 model attained an average forecast accuracy of ~92% [2], measured by 1-MAPE (mean absolute percentage error). In contrast, the traditional forecasting (a combination of moving average and manual adjustment) had about 75% average accuracy. For example, at the Hanoi FPT Shop branch, the model predicted November 2024 demand for the iPhone 15 within 8% error of actual sales, whereas the old method had erred by ~25% [2] for the same product previously. This improved accuracy has tangible business impact - it helps prevent stockouts (when forecast was too low) and avoids excess inventory (when forecast was too high). The high accuracy can be attributed to the Transformer model's strength in capturing complex patterns, combined with GPT-4's ability to adjust or validate forecasts based on additional context (e.g., incorporating the knowledge of an upcoming marketing campaign if mentioned in the prompt). Notably, the GPT-4 assistant could also explain the forecasts in plain language, which built trust with management in the predictions.

Table 1. Comparision of legacy system vs. AI-enhanced systm at FPT Retial (key metrics ans outcomes)

Criterion	Legacy System (Excel/Manual/ARIMA)	ChatGPT+AI Enhanced System
Demand Forecast Accuracy	~75% (historical average with ARIMA)	~92% (Transformer+GPT-4 model)
Forecast Error (for example)	~25% MAPE for top products (some high variance)	<8% error for key product (iPhone 15, Nov 2024)
Inventory Query Time	~2–5 minutes (manual lookup in ERP/Excel)	<10 seconds (chatbot real-time response)
Monthly Report Preparation	~20 minutes by analyst (per report)	~5 minutes (automated by AI)

Logistics Cost (2024)	Baseline (no AI)	15% reduction (optimized by accurate forecasts)
Excess Inventory Cost	Significant overstock in some cases	~10 billion VND saved in 2024 (reduced overstock)
Staff Time on Inventory Tasks	High (manual data collation, reporting)	35% less time (tasks automated/accelerated)
Employee Adoption	Traditional tools, varying usage	98% trained staff using chatbot; 95% prompt success
Decision-making Speed	Slower - needed meetings to interpret data	Faster - on-demand answers for quick decisions

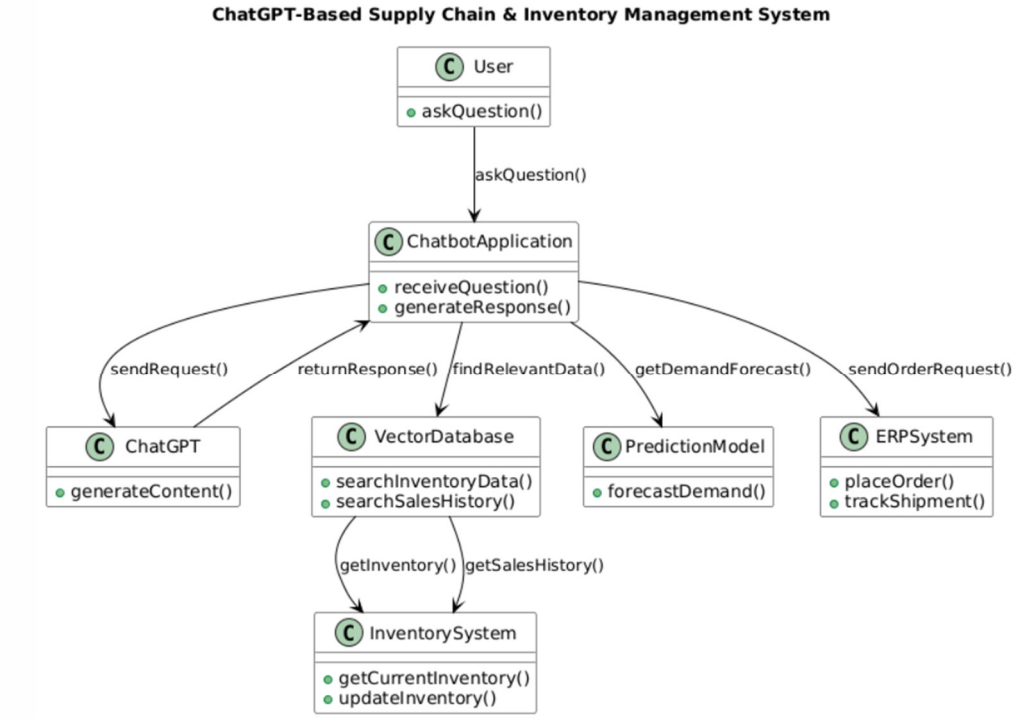

Figure 5. Activity model

Operational Efficiency: The introduction of the chatbot interface and automation modules reduced the time to perform several inventory management tasks. Routine inventory checks and updates that used to be done by manually exporting data from the ERP and analyzing in Excel (taking on average 10 - 15 minutes per query) can now be done via chat in under 1 minute. The system's real-time inventory lookup (Module 3) returns detailed product stock information in **< 10 seconds** after a question is asked. This represents a 35% reduction in processing time for inventory lookup and audit tasks overall. In monthly inventory audits, for instance, the team previously spent ~4 hours aggregating data; with the AI assistant, the process (querying discrepancies, getting summaries) took ~2.5 hours - freeing up ~1.5 hours for higher-value analysis. Furthermore, **report generation time** was cut by 50% on average. A concrete example is the October 2024 sales report for the Long Châu pharmacy chain: the legacy method (manually compiling data and creating charts in Excel) took about 20 minutes, whereas the ChatGPT-powered module produced the report in about 5 minutes [2]. The generated report included a revenue table and a sales trend chart automatically, requiring only minor edits before final review. This speed-up not only saves analyst time but also enables more frequent reporting if needed (e.g., on-demand weekly summaries).

Cost Savings: With better demand forecasts and streamlined operations, FPT Retail realized notable cost savings. Improved demand prediction allowed the company to optimize stock levels at each warehouse. By minimizing overstock, the carrying cost of inventory was reduced. In 2024, the AI system helped avoid an estimated 10 billion VND in excess inventory costs (calculated from reductions in unsold stock that would have remained at year-end). Additionally, the more accurate forecasts and faster inventory turns contributed to lower **logistics costs by ~15%** [1]. This aligns with the McKinsey estimate of AI-driven 15% logistics cost reduction [4], and in our case stemmed from more efficient inter-warehouse transfers and reduced emergency shipments. In practical terms, warehouses could better schedule replenishments and consolidation of shipments, saving fuel and labor. While there was an upfront investment in integrating GPT-4 (API usage costs) and developing the system, these expenses were outweighed by the efficiency gains – yielding a clear ROI within the first year.

Employee Productivity and Adoption: The new system was well-received by staff. After training, 98% of

the warehouse and store managers in the pilot group were regularly using the chatbot assistant in their daily work. User feedback surveys indicated that the tool made it easier to access information - "I can just ask and get the data I need immediately, instead of digging through reports," one respondent noted. The prompt training resulted in a high success rate: about 95% of queries were handled by the AI on first try without rephrasing [2]. This indicates the prompts and system outputs aligned well with user needs. An unintended benefit was improved cross-department communication; the AI-generated reports and answers were in a standardized format, which made it simpler for different teams (sales, logistics, finance) to interpret the information uniformly. Some employees reported that using the chatbot freed them from tedious tasks (like compiling numbers), allowing them to focus on decision-making and exceptions. Overall, productivity in inventory management tasks (from forecasting to stock checking to reporting) is qualitatively observed to have increased, corroborating the quantitative time savings.

System Robustness and Accuracy of Responses: During a 3-month evaluation period, we monitored the quality of the AI system's responses. We found that factual accuracy of responses was high when the queries fell within the domain of the provided data. Thanks to RAG, instances of "hallucination" (the AI inventing data) were rare. In less than 5% of queries, the answer had a minor error or required clarification - often these were cases where the question was vague or combined multiple requests in one. By refining prompt guidance and encouraging users to split complex queries, we reduced this error rate. The chain-of-thought reasoning traces (visible in logs) showed that GPT-4 correctly followed through multi-step calculations (e.g., current stock minus forecast) in 90%+ of applicable cases; for the remaining, we adjusted the prompt to explicitly break the steps. Table I encapsulates the comparative outcomes.

The above results underscore how integrating GPT-4 and RAG into inventory management led to **both quantitative and qualitative improvements**. Forecast accuracy and response times improved dramatically, translating to cost savings and better service levels. The AI system's ability to generate insights and explanations also added value beyond raw numbers, aiding decision-makers. In the next section, we further discuss the implications of these findings and the lessons learned during implementation.

5. DISCUSSION

The case of FPT Retail demonstrates the potential of GPT-based AI in a practical enterprise setting, but it also surfaced several challenges and considerations important for similar deployments.

Impact on Supply Chain Performance: The improvements in forecast accuracy (to 92%) are particularly noteworthy. In retail, a 17% point increase in accuracy (from 75% to 92%) can significantly reduce stockouts and excess stock. Our findings are in line with broader studies that report AI can reduce inventory levels by ~20 - 35% while maintaining or improving service levels. FPT Retail's inventory turnover increased after implementation, as better demand estimates meant more optimal reordering. Additionally, the 15% logistics cost reduction we observed echoes industry expectations [4] for AI-driven optimization. These gains illustrate that beyond theoretical accuracy metrics, an LLM-powered system can directly contribute to key supply chain KPIs. An interesting observation was that the AI's ability to explain *why* it was forecasting higher or lower demand (e.g., citing a marketing event) made planners more confident to act on the predictions – addressing a common reluctance to trust "black box" models.

Prompt Engineering Efficacy: Our use of prompt engineering techniques played a critical role in achieving high performance. Few-shot prompts gave the model context [7] to output in desired formats (reducing the need for post-editing), while chain-of-thought prompting improved the correctness of reasoning for questions involving arithmetic or logical comparisons. For instance, when asked "Do we need to reorder item Y?", the model, via CoT, would explicitly consider current stock, inbound shipments, and forecasted sales before answering - this reduces the chance of a simplistic or wrong answer. The ReAct approach, effectively turning the chatbot into an agent that can call data lookup actions, proved essential for grounding the model [11]. Without ReAct, ChatGPT might have tried to answer from its trained knowledge (which could be outdated or irrelevant to our company data), leading to hallucinations. By forcing retrieval of up-to-date info, we ensured **accuracy and relevancy** of responses. This aligns with known benefits of RAG in combating LLM hallucination and improving factuality [8]. One challenge we encountered was the prompt length and context window limits: combining a user query, few-shot examples, retrieved data, and reasoning steps sometimes

made prompts quite lengthy. We had to optimize prompt length (e.g., by removing unnecessary background text and keeping only key data points) to avoid hitting the model's context size limit. We also explored truncating older conversation turns in the chatbot when they exceeded memory, to maintain performance in long sessions.

Data Quality and Integration: The success of an AI solution is tightly coupled with the quality of data it uses. In our implementation, we faced issues with inconsistent product naming between the ERP and the vector database, which initially caused some retrieval misses (the model would ask for "Product A" but the vector DB had it under a slightly different name). We solved this by cleaning and standardizing the data during the embedding phase, as well as adding keywords and synonyms to product entries to improve recall. Another hurdle was ensuring real-time data synchronization - inventory counts change rapidly, and our pipeline to update the vector store embeddings had to be near real-time for critical fields. We addressed this by scheduling frequent vector DB refreshes and directly querying the live inventory system for quantities (bypassing the vector store for real-time numeric data). This hybrid approach (live DB for quantities, vector DB for descriptive context) worked well. The general lesson is that **RAG systems require robust data pipelines**: if the underlying data is outdated or incorrect, the LLM will faithfully reflect those errors [9]. Thus, data governance remains as important as the AI model itself in such applications.

Initial Deployment Costs vs. Benefits: Implementing GPT-4 in an enterprise context involved costs such as API usage fees, development effort for integration, and training time for staff. We estimate the monthly OpenAI API cost for our pilot (around 10,000 queries with an average prompt/response size of 1500 tokens) to be a few hundred USD - which is negligible compared to the value of time saved in reports and the inventory reduction benefits. Development and integration took approximately 3 person-months of work for our team, which included building the FastAPI backend, prompt iteration, and testing. From a financial perspective, the payback period was quick (within the first year) given the efficiency gains. That said, companies considering such solutions should conduct a pilot on a small scale to validate impact before scaling up. We leveraged existing cloud infrastructure of FPT and did not need to invest in on-premise GPU hardware since we used OpenAI's cloud API, which lowered the barrier to entry. One concern was **data confidentiality**, since inventory data was being sent to an external API. We mitigated this by stripping any sensitive customer information and by leveraging OpenAI's enterprise data privacy assurances (no training on our prompts, data isolation). In the future, if even stricter data control is desired, one could consider deploying open-source LLMs on a private server, though at the cost of potentially lower performance than GPT-4.

Scalability and Generalization: Our case study focused on inventory management in retail, but the approach is generalizable to other domains within supply chain (e.g., procurement, production planning) and other industries. The combination of a domain-specific database, an LLM, and prompt engineering can be adapted to any scenario where users query data or need automated analysis. We anticipate that similar GPT-powered assistants could be built for tasks like supplier risk assessment (feeding in supplier data and letting the LLM evaluate risks), or for customer demand sensing (ingesting social media/customer feedback via RAG and summarizing trends). The positive response from FPT Retail's team suggests that AI assistants can be accepted by domain experts if they produce reliable and useful outputs. This is promising for broader AI adoption in operations. However, one must be cautious: not every organization will have data as clean or an AI-embracing culture; thus, results may vary. In our case, FPT Retail's AI-first mindset helped in change management - the initiative had executive sponsorship and employees were incentivized to learn the new tool. Companies should also invest in **user training and change management** to replicate such success.

Challenges and Limitations: Despite the achievements, there are limitations to note. The system currently answers based on data it has; if a query is outside its knowledge base (for example, "What were industry sales for product X?" or any question about competitors), it cannot answer unless that data is added to the context. It's not a general oracle, but tailored to FPT's data. We also found that extremely complex multi-part questions can still confuse the model. For example, a compound question like "Give me the stock of product A and forecast of product B, and also list last month's sales of product C" pushes the limits - the model might mix up parts in the answer. Our recommendation is to encourage one question at a time, or handle such cases by breaking

the query internally. Another challenge is maintaining prompt effectiveness over time. As inventory policies or formats change, prompt templates may need updating. This is an ongoing maintenance task akin to software updates. Additionally, while GPT-4's multilingual ability is strong, we observed minor issues in generating perfectly fluent Vietnamese business language (as it was primarily trained on English). We addressed this by providing Vietnamese exemplars in the few-shot prompts, which improved its style and correctness in Vietnamese responses.

Ethical and Societal Considerations: The use of AI in decision support raises questions about the changing nature of jobs. At FPT Retail, the chatbot did not replace any role but augmented staff capabilities. Employees could accomplish more in less time, focusing on strategy rather than number crunching. This is an example of AI creating a positive augmentation effect. We also took care that the AI's suggestions were reviewed by humans; final decisions (like how much to reorder) were made by managers, with the AI as an advisor. This human-AI collaboration is important to ensure accountability. Moreover, we had to ensure the AI's advice did not inadvertently encode any bias or error (for instance, if training data had an abnormal spike, the forecast might overemphasize it). Transparency via explanations was crucial here – by seeing the reasoning, users could spot if something looked off. We cite this as a best practice for enterprise AI: always provide interpretable output or rationale so that users can trust and verify the AI's recommendations.

In summary, the discussion highlights that GPT-4 with RAG can serve as a powerful tool in inventory management, yielding substantial efficiency and accuracy improvements. The success depends on high-quality data integration, careful prompt engineering, and user acceptance. Our experiences can guide future implementations: start with clear objectives (e.g., reduce forecasting error), use a hybrid AI-human approach to leverage strengths of each, and remain vigilant about data quality and prompt design. As AI technology evolves (GPT-5 and beyond, larger context windows, more fine-tunable models), such systems could become even more capable and easier to deploy, potentially becoming standard in enterprise resource planning solutions.

6. CONCLUSION

This paper presented a comprehensive study on applying ChatGPT (GPT-4) with retrieval-augmented generation to inventory management in an enterprise e-commerce context. FPT Retail's case exemplifies how an AI-first approach can transform traditional operations: the integration of a GPT-4 assistant into the ERP system improved demand forecasting accuracy, accelerated information retrieval, and automated reporting - delivering measurable business value in terms of cost savings and efficiency gains. Our system architecture combined a Transformer-based forecasting model, a vector database of inventory knowledge, and an intelligent prompt-engineered chatbot interface to achieve these outcomes. We showed that with appropriate prompt engineering (few-shot examples, CoT reasoning [10], ReAct tool use, and meta-instructions), a large language model can function as a reliable and effective supply chain assistant, rather than a generic text generator.

The results attained - from a 35% reduction in inventory handling time to a 50% reduction in report generation effort [2] - demonstrate that GPT-powered tools can augment human workers, handling routine data queries and analyses so that employees can focus on strategic decisions. These findings contribute to the growing evidence that AI, and LLMs in particular, can be successfully leveraged in operational technology (OT) and not just IT or customer-facing applications. We also addressed challenges such as ensuring data fidelity and managing the change in workflows through training. The high adoption rate among FPT Retail staff indicates that with proper user-centric design and education, resistance to AI tools can be minimal and short-lived.

For the academic and AI engineering community, our work provides a practical example of bridging advanced NLP techniques with enterprise software (ERP systems), highlighting the importance of interdisciplinary approaches. We combined insights from AI research (RAG, prompt engineering) with domain knowledge in retail inventory processes to tailor a solution that fits the business needs. One key insight is that **context matters** - the same GPT-4, when coupled with the right context data and prompts, can deliver very domain-specific functionality (like inventory auditing or forecasting explanation). This context-injection approach via RAG might be a template for other industries looking to exploit LLMs while keeping them grounded in factual databases.

Future work will explore several directions: First, integrating reinforcement learning or fine-tuning to further adapt the language model to the company's

preferred style and possibly to handle Vietnamese language nuances even better. Second, expanding the scope of the assistant beyond inventory - for example, into procurement (negotiation chatbots using supplier data) or customer support (using the same knowledge base to answer customer inquiries about product availability). Third, assessing long-term impacts, such as how continuous use of the AI changes inventory KPIs over multiple years, and whether the model can adapt to shifts (like new product launches or unforeseen events) through rapid updating of its knowledge base. Lastly, from a research standpoint, one could formally evaluate the contribution of each prompt engineering technique via ablation studies (e.g., turning off CoT or ReAct to quantify performance drops), to further validate the design choices in our system.

In conclusion, the application of ChatGPT in FPT Retail's inventory management has proven successful, offering a reference model for deploying generative AI in enterprise resource planning [1]. As AI technologies advance, we expect such intelligent assistants to become integral in decision-making loops across industries. The synergy of human expertise with AI's speed and analytical power can unlock new levels of efficiency and responsiveness in supply chain management. We hope this case study encourages more organizations and researchers to experiment with GPT-based solutions for operational challenges, and to share findings that can collectively drive innovation in the field of AI applications in e-commerce and beyond.

REFERENCES

[1]. Nguyen Quang Hung, Le Minh Duy, "Su dung Few-Shot Learning de xay dung Chain-of-Thought tu dong khong giam sat," in *VNICT2024: The 27th Vietnam Conference of Selected ICT Problems*, Nha Trang, 582-587, 2024

[2]. FPT Retail Internal Report, *Ung dung ChatGPT vao quan ly hang ton kho. 2024.*

[3]. McKinsey & Co., *Succeeding in the AI supply-chain revolution*. (Accessed on McKinsey.com). 2021.

[4]. McKinsey & Co., *AI in supply chain management*. 2023.

[5]. OpenAI, *Introducing ChatGPT Enterprise*. OpenAI Blog. 2023.

[6]. IBM, *AI for Inventory Management*. IBM Business Blog. 2022.

[7]. Brown T. B., et al., "Language Models are Few-Shot Learners," in *34th Conference on Neural Information Processing Systems (NeurIPS 2020),* Vancouver, Canada, 2020.

[8]. Lewis P., et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *34th Conference on Neural Information Processing Systems* (NeurIPS 2020), Vancouver, Canada, 2020.

[9]. NVIDIA Blog, *What Is Retrieval-Augmented Generation (RAG)*. 2023.

[10]. Wei J., et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems,* Article No.: 1800, 24824 – 24837, 2022.

[11]. Yao S., et al., "ReAct: Synergizing Reasoning and Acting in Language Models," in *11th International Conference on Learning Representations*, ICLR 2023, 2023.