International Journal of Innovation and Research in Educational Sciences
Volume 13, Issue 3, ISSN (Online) : 2349–5219

LLM Framework for Enhanced Supply Chain Data Analysis and Decision Support: A Comparative Study

Johan Cho and Abe Zeid*
Department of Mechanical and Industrial Engineering Northeastern University Boston, MA 02115 USA.
Corresponding author email id: i.zeid@northeastern.edu

Date of publication (dd/mm/yyyy): 30/06/2026

***Abstract* – Current supply chain management techniques rely on rigid business intelligence systems such as PowerBI and Tableu that do not provide intuitive natural language insights from complex, multi-dimensional datasets. These systems typically require a heavy time and monetary investment plus specialization and domain-specific knowledge to build and use. Supply chains generate vast quantities of mutli-level and often unorganized data, spanning inventory levels, supplier performance, transportation metrics, and quality control results. Thus, a gap between data availability and actionable insights and information forms. This paper presents a Large Language Models (LLMs) framework that uses specialized AI agents (OpenAI, Gemini, DeepSeek) to provide supply chain data analysis through natural language interfaces. This framework incorporates model selection, chat history management, and data analysis capabilities though retrieval-augmented generation (RAG) that combine to create a specialized supply-chain-specific assistant. Through evaluation on real supply chain datasets, this LLM system demonstrates general analytical accuracy, acceptable response quality, and improved cost-effectiveness compared to generalized LLMs. It also solves many of the problems of traditional supply chain analysis while also creating a promising opportunity to enhance the field. This system validates this approach as a promising application of LLMs within supply chain management and other analysis-focused fields.**

***Keywords* – LLM, Data Analysis, Decision Support, Comparative Study.**

I. Introduction

1.1. *Background and Motivation*

In recent decades, supply chain management has changed from paper-based analytics to digitized and data-driven operations [1]. With the help of technology, analysts can capture and store large volumes of operational data across all facets and rungs of the supply chain [2]. Modern supply chains generate data that encompass real-time inventory tracking, supplier performance metrics, transportation logistics and quality control measurements [3]. However, the complexity and volume of this data renders it difficult to extract actionable insights and information [4].

Traditional business intelligence platforms such as Power BI, Tableau, and other data visualization tools are currently the backbone of supply chain analytics [5]. These systems are effective at creating reports, dashboards, and visualizations that communicate insights to stakeholders and decision makers [6]. However, as supply chain and data complexity grows, their fundamental flaws and limitations become apparent [7]. First, these platforms require significant specialized technical expertise to operate effectively, often requiring a dedicate data analysis and/or IT team to operate [8]. Second, they require predefined data and query structures, making them relatively inflexible to system change [9]. Lastly, they do not provide contextual interpretation of results, leaving users to manually create insights and actions based on the presented information [10].

Large Language Models (LLMs) create opportunities for addressing these analytical challenges through human language interfaces that democratize complex data analysis [11]. Users without a technical background can use natural language with LLMs to process complex datasets and generate insights [12], representing a shift from rigid, template-based reporting to flexible and natural analytics that adapts to different user and analytical requirements [13]. However, applying LLMs to supply chain analysis presents challenges that generalized language models cannot overcome [14].

Supply chain data is multi-dimensional and interconnected, requiring specialized domain knowledge to interpret [14]. For example, understanding the complicated relationships between supplier lead times, inventory levels, and customer demand patterns requires both statistical analysis and knowledge in supply chain principles, market dynamics, and operational constraints [15] - intuition. Furthermore, supply chain management often involves complex decision-making scenarios that require context, intuition and multi-step reasoning [16]; generalized LLMs lack the knowledge, context and intuition needed to generate insights for these scenarios [17].

The disparity between a large amount of data and small amount of usable information in supply chain analysis creates challenges for organizations [18]. Companies encounter difficulties in responding to market fluctuations, optimizing operational efficiency, and preemptively identifying risks [19]. This analytical bottleneck manifests in a couple of ways: suboptimal inventory management practices, inefficient supplier relationship management, delayed responses to quality control issues, and missed opportunities for cost reduction and service improvement [20]. As supply chains continue to increase in complexity and new geopolitical and climate challenges arise, the demand for more sophisticated, accessible, and contextually aware analytical capabilities becomes increasing essential for maintaining resilience [21].

1.2. *Problem Statement*

Organizations have difficulties extracting meaningful insights from supply chain data due to limitations in current analytical approaches; traditional BI platforms create accessibility barriers through their technical complexity [5, 13]. Meanwhile, generalized LLMs lack the specialized domain knowledge required to interpret complex supply chain relationships between suppliers, inventory, and demand patterns [14, 15]. Current research provides insufficient comparative evaluation of different LLM agents in supply chain contexts [11, 12]. Then, the mismatch between generic AI capabilities and industry-specific analytical requirements [17, 18], further constrains organizations' ability to leverage their data for effective supply chain decision-making.

1.3. *Research Objectives*

This research addresses the these challenges through three main goals. The first involves designing and implementing an LLM framework specifically built for supply chain analytics. Then, different LLM agents (OpenAI, Gemini, DeepSeek) were compared to evaluate their performance characteristics within supply chain contexts [11, 12]. Lastly, the effectiveness of specialized agent configurations was heuristically compared against generalized LLM approaches.

1.4. *Research Contributions*

This research makes several key contributions to the field of AI-driven supply chain analytics. The study provides the first comprehensive evaluation of multiple LLM agents specifically configured for supply chain data analysis, establishing a foundation for understanding how different AI models perform in specialized domain contexts. Through rigorous comparative analysis, the research reveals distinct performance characteristics and op-

-timal use cases for different agents across various supply chain analytical tasks.

The work also delivers practical implementation guidelines that organizations can immediately apply when selecting and deploying LLM agents for their supply chain analytics needs. These evidence-based recommendations address a critical gap in the literature by providing clear guidance on matching agent capabilities to specific analytical requirements and operational contexts, ultimately enabling more effective AI-powered supply chain decision-making.

II. Literature Review

2.1. *LLMs in Supply Chain Management*

Recent research has explored the applications of generalized LLMs in supply chain optimization and analytics, revealing both promising opportunities and significant limitations [8, 9]. While these studies demonstrate the potential for natural language processing applications in supply chain data interpretation, they also highlight the constraints of generic language models when applied to domain-specific supply chain contexts [14]. Recent advances in AI-driven supply chain decision support systems have primarily focused on traditional machine learning approaches, with limited exploration of how specialized LLM configurations can address industry-specific analytical challenges [10, 11].

2.2. *Comparative LLM Performance Studies*

The literature on comparative LLM performance reveals various evaluation methodologies for comparing different language models, though most focus on general capabilities rather than domain-specific applications [11, 12]. Research has identified domain-specific performance variations across different LLM architectures, as well as cost-effectiveness and accuracy trade-offs in enterprise LLM deployments. However, studies examining agent-specific capabilities and limitations in analytical tasks remain limited, particularly within specialized industry contexts like supply chain management.

2.3. *AI in Supply Chain Analytics*

Systematic literature reviews of AI applications in supply chain management demonstrate a clear evolution from traditional BI to AI-driven analytics in supply chain contexts [1, 2]. This transition has highlighted the growing role of natural language interfaces in democratizing supply chain analytics, making complex data analysis more accessible to non-technical users [13]. However, enterprise requirements for AI-powered analytical solutions often exceed what current generalized approaches can provide, creating a demand for more specialized frameworks.

2.4. *Gap Analysis*

Current literature reveals several critical gaps that this research addresses. First, few studies provide rigorous comparisons between different LLM agents in specialized domain contexts, particularly in supply chain analytics where domain expertise is crucial [11, 12]. Second, existing evaluations focus on general capabilities rather than industry-specific analytical requirements, leaving organizations without guidance on which AI tools best serve their particular needs. Finally, there is limited research on practical deployment considerations and real-world performance variations across different agents, creating a significant gap between academic research and enterprise implementation needs.

III. Methodology

3.1. *LLM Agent Framework Design*

3.1.1. *Agent Selection and Configuration*

ThreeLLMs, ChatGPT (o4-mini), Gemini (gemini-2.0-flash) and DeepSeek (deepseek-chat) were chosen because of their availability and size; the models are small enough to actively specialize and customize through a system prompt. The hardware to locally run LLMs through was not available; thus, these three models were chosen because they could be run remotely and interfaced through an application programming interface (API). The local program would make HTTP calls to the providers server, and it would return data structures containing the responses. Credits were bought for ChatGPT and DeepSeek at a small fee, while Google is free to run. After signing up and gaining the API keys, they were stored in an environment (.env) file, where they'll be accessed within the code using built-in python libraries.

3.1.2. *Agent Specialization Approach*

To increase the ease of testing and using the agents, a base class was created so that each agent would have the same interfaces - methods, variables, structures, when prompting it. Although this design increased development time, it decreased complexity and enables the easy addition of more agents, especially those that are compliant with the OpenAI API specification.

```
class AgentBase(abc.ABC):
    model_name: str
    history_file: str
    chat_history: ChatHistory
    @abc.abstractmethod
    def get_or_upload_file(self, local_path: str) -> Any:
        ...
    @abc.abstractmethod
    def ask(self, question: str, file: Any) -> str:
        ...
```

From this base class, two child classes that each represent a model were created, Google's Gemini and OpenAI's ChatGPT. Deepseek's class is a child of ChatGPT's because it implements the same API specification.

3.2. *Data Processing Pipeline*

3.2.1. *Multi-Modal Input Processing*

One problem wit the remote APIs are that they cannot take CSV files directly; they have to be converted into a different file format, a PDF. Thus, a helper class was created to abstract this away from the bulk of the code. Furthermore, DeepSeek does not implement major portions of OpenAI's api; it lacks the "files" and "responses" endpoint. Thus, the csv has to be encoded as text, which is done using the python library, Pandas.

```
class CSVFile:
    def to_pdf(self, conditional: bool = True) -> str:
        ...
    def as_dataframe(self, **kwargs) -> pd.DataFrame:
        ...
```

3.3. *Comparative Evaluation Framework*

The evaluation framework established for this research employs a multi-dimensional assessment approach designed to capture both quantitative performance metrics and qualitative characteristics of agent responses. This comprehensive methodology enables systematic comparison of agent capabilities while accounting for the nuanced requirements of supply chain analytical contexts.

3.3.1. *Qualitative Assessment*

The qualitative evaluation framework encompasses four primary dimensions that collectively assess agent response effectiveness in supply chain analytical contexts.

Domain Relevance evaluates the appropriateness of responses to supply chain-specific contexts, measuring how effectively agents demonstrate understanding of supply chain principles, terminology, and operational constraints. High-scoring agents consistently integrate industry-specific knowledge, recognize supply chain relationships and dependencies, and apply contextually appropriate analytical frameworks.

Business Value measures the actionability and practical applicability of insights generated by each agent. This criterion evaluates whether responses provide concrete, implementable recommendations that stakeholders can directly apply to operational or strategic decisions, including specific action items, quantified impacts, and clear connections between analytical findings and business outcomes.

Communication Quality assesses the clarity, professionalism, and accessibility of agent responses, considering how effectively agents convey complex analytical concepts to diverse audiences. High-quality communication demonstrates appropriate supply chain terminology, logical organization, professional formatting, and accessibi--lity to users without specialized technical backgrounds.

Analytical Depth evaluates the level of contextual understanding and strategic insight demonstrated in agent responses. This criterion distinguishes between surface-level answers and comprehensive analysis that considers multiple dimensions, interdependencies, and broader implications through multi-step reasoning and strategic perspective.

3.3.2. *Comparative Methodology*

The comparative evaluation methodology employs a rigorous, systematic approach to eliminate bias and ensure reliable performance assessment across all agents.

A Standardized Query Set forms the foundation of the comparative methodology. Twenty-eight supply chain analytical queries were designed to represent diverse analytical requirements commonly encountered in enterprise contexts, spanning simple factual retrieval, complex multi-dimensional analysis, strategic decision support, and operational optimization scenarios. Each query was presented identically to all agents, ensuring that performance differences reflect agent capabilities rather than query formulation variations.

Blind Evaluation procedures ensure unbiased assessment of agent responses. Individual responses were evaluated independently without agent identification, enabling objective quality assessment focused purely on response characteristics. Evaluators assessed each response against the qualitative assessment criteria before comparing responses across agents for the same query.

Multiple Evaluation Rounds enhance reliability and consistency of performance assessments. The same queries were presented to agents across different sessions and timeframes, enabling evaluation of response consistency and identification of systematic patterns versus random variations. This approach reveals whether observed performance characteristics represent stable agent capabilities or transient fluctuations.

Cross-Agent Analysis provides direct performance comparison through systematic evaluation of agent responses to identical queries. This comparative approach reveals relative strengths and weaknesses across different analytical scenarios, enabling identification of optimal agent selection for specific query types through examination of differences in analytical depth, recommendation quality, computational efficiency, and overall business value delivered.

IV. Agent Setup and Implementation

4.1. *Individual Agent Configuration*

Each agent was individually configured to optimize performance within the supply chain analytical framework while maintaining consistent system-level parameters. This section details the specific configuration choices, capabilities, and observed strengths of each agent implementation.

4.1.1. *OpenAI Agent*

The OpenAI agent implementation utilizes the o4-mini model, selected for its optimal balance of cost-effectiveness and strong analytical capabilities. This model provides sufficient reasoning depth for complex supply chain queries while maintaining economical token consumption rates suitable for enterprise deployment.

The agent's capabilities encompass comprehensive file upload support enabling direct analysis of supply chain datasets, vision processing for multimodal data interpretation, and structured response generation that facilitates consistent output formatting. Configuration parameters were optimized specifically for supply chain contexts through specialized prompts guiding the model toward domain-appropriate analytical approaches, with careful attention to token usage optimization and context window management for multi-turn conversations.

Observed strengths include exceptionally consistent formatting across diverse query types, ensuring predictable output structures that facilitate downstream processing and stakeholder presentation. The agent demonstrates reliable performance across various analytical scenarios, maintaining steady quality levels whether addressing simple factual queries or complex strategic analyses, making it particularly suitable for standardized reporting tasks requiring predictable, professional outputs.

4.1.2. *Gemini Agent*

The Gemini agent implementation leverages the gemini-2.0-flash model, optimized for high-speed processing and contextual understanding. This model architecture emphasizes rapid response generation without sacrificing analytical quality, making it well-suited for time-sensitive operational queries and interactive data exploration.

The agent's capabilities include advanced multimodal processing that seamlessly integrates various data types, extensive context windows enabling analysis of large datasets without information loss, and rapid response generation supporting interactive analytical workflows. Configuration parameters enhance performance in quick insights generation and data exploration, emphasizing efficient pattern recognition, rapid identification of key trends and anomalies, and concise communication of essential findings.

Observed strengths demonstrate exceptional processing speed, delivering analytical insights significantly faster than alternative agents while maintaining quality standards. The agent excels in contextual comprehension, quickly grasping supply chain relationships and dependencies within datasets. Resource utilization efficiency makes Gemini particularly cost-effective for high-volume query scenarios and exploratory analyses requiring multiple iterations.

4.1.3. *DeepSeek Agent*

The DeepSeek agent utilizes the deepseek-chat model, configured for specialized analytical reasoning and comprehensive business context integration. This model demonstrates particular strength in complex, multi-step reasoning tasks requiring deep understanding of supply chain principles and strategic business considerations.

The agent's capabilities focus on advanced analytical reasoning that unpacks complex supply chain scenarios through systematic, multi-layered analysis. Deep business context integration ensures analytical outputs consider broader strategic implications beyond immediate data findings. The configuration supports text-only processing with enhanced CSV data interpretation capabilities, utilizing the Pandas library for robust data parsing rather than file upload endpoints.

Configuration emphasizes comprehensive analytical frameworks examining problems from multiple perspectives. CSV data is encoded directly into prompts using Pandas DataFrames, ensuring precise data interpretation. The configuration guides the agent toward strategic thinking connecting tactical findings to broader business objectives and supply chain optimization opportunities.

Observed strengths highlight exceptional analytical depth, consistently delivering insights extending beyond surface-level observations to strategic recommendations grounded in supply chain best practices. The agent excels in comprehensive business recommendation development, producing actionable guidance addressing what the data shows and how organizations should respond. Cost efficiency represents another significant advantage, with the agent delivering sophisticated analytical outputs at competitive price points.

4.2. *Standardized System Configuration*

4.2.1. *Common System Prompt*

Maintaining consistency across agents required implementation of a standardized system prompt establishing the analytical persona and behavioral expectations for all agents. This common foundation ensures performance variations reflect inherent agent capabilities rather than differences in instruction or role definition.

The system prompt defines the agent's identity as "a knowledgeable and detail-oriented supply chain analyst," establishing domain expertise as the foundation for all analytical interactions. This framing guides agents toward supply chain-specific interpretations and recommendations rather than generic data analysis approaches. The prompt emphasizes accuracy, insight, and actionability as core output requirements, ensuring responses meet practical business needs. Instructions to utilize data trends, performance metrics, and relevant supply chain principles direct agents toward comprehensive analytical approaches considering multiple dimensions of supply chain operations.

You are a knowledgeable and detail-oriented supply chain analyst.

Your role is to provide accurate, insightful, and actionable answers

based on the uploaded supply chain dataset and ongoing conversation history.

Use data trends, performance metrics, and relevant supply chain principles

when forming your answers.

The standardized prompt represents the minimum necessary instruction to establish supply chain analytical capabilities while remaining sufficiently general to allow each agent's unique strengths to emerge, enabling fair performance comparison while ensuring all agents operate with appropriate domain context.

4.2.2. *Uniform Data Processing*

Ensuring valid performance comparison required standardized data processing procedures across all agents. This uniformity eliminates data handling variations as confounding factors in performance assessment.

Consistent data input formats were established through the CSVFile helper class, providing standardized interfaces for data conversion regardless of agent-specific requirements. All agents receive supply chain datasets in formats optimized for their respective APIs, but the underlying data structure and content remain identical. Standardized file handling and preprocessing procedures ensure that data transformations necessary for agent compatibility do not introduce analytical biases or advantages for specific agents, maintaining identical information content across all formats.

Identical query presentation methodology ensures all agents receive questions formulated in precisely the same manner, using consistent phrasing, structure, and context provision across all evaluations. This uniformity eliminates query formulation as a variable in performance assessment.

4.3. *Implementation Architecture*

4.3.1. *Chat History Management*

Effective analytical conversations require maintaining context across multiple interactions, enabling agents to build upon previous exchanges and provide increasingly refined insights. The chat history management system implements comprehensive tracking and retrieval capabilities supporting contextual analysis while maintaining system performance.

The ChatHistory class provides centralized conversation management functionality accessible to all agent implementations through consistent interfaces. The log_interaction method captures complete interaction records including user queries, agent responses, timestamps, and associated metadata, enabling both immediate conversat-

-ion context maintenance and subsequent performance analysis.

```
class ChatHistory:
    def log_interaction(self, **log: Unpack[_ChatLog]) -> list[_ChatLog]:
            ...
    def history_as_gpt_messages_list(self, length: int = 5) -> list:
            ...
```

The history_as_gpt_messages_list method converts stored conversation history into standardized message formats compatible with various agent APIs, enabling seamless context provision across different agent implementations despite varying API specifications. The length parameter allows configurable context window management, enabling optimization based on query complexity and agent-specific context window capabilities.

4.3.2. *Error Handling and Quality Assurance*

Robust error handling and quality assurance mechanisms ensure reliable system operation across varying conditions. These safeguards prove essential for enterprise deployment scenarios where analytical reliability directly impacts business decision-making processes.

Standardized error handling across all agents ensures consistent system behavior regardless of which agent encounters errors. API failures, rate limiting, network timeouts, and other operational issues trigger appropriate retry logic, fallback procedures, or user notifications as circumstances dictate.

Response validation and quality checks examine agent outputs for completeness, coherence, and appropriate--ness before presenting results to users. Validation procedures identify truncated responses, formatting inconsistencies, or responses failing to address the original query. Consistent logging and performance monitoring track system behavior, agent performance, and usage patterns across all interactions, providing the data foundation for identifying optimization opportunities and validating system improvements.

V. Results

5.1. *Dataset Description*

The evaluation utilized a comprehensive supply chain dataset with 100 stock-keeping units distributed across haircare, skincare, and cosmetics categories. Each SKU contains 24 distinct attributes capturing multiple dimensions of supply chain operations including pricing structures, inventory levels and turnover metrics, logistics and transportation performance, supplier relationship metrics, manufacturing efficiency indicators, and quality control results.

The dataset's complexity derives from multi-dimensional relationships spanning entire supply chain networks from suppliers through manufacturers to distribution channels. These relationships create analytical scenarios requiring agents to recognize dependencies, identify trade-offs, and consider systemic impacts of localized changes. Query coverage encompasses 28 standardized supply chain analytical queries designed to test different aspects of data interpretation and analytical reasoning, ranging from simple factual retrieval through complex analytical scenarios requiring multi-step reasoning to strategic decision support queries demanding comprehensiv-

-e recommendations.

5.2. *Agent Performance Comparison*

5.2.1. *Analytical Approach Differences*

The three agents demonstrated distinct analytical approaches when processing identical supply chain queries, revealing fundamental differences in reasoning depth, business context integration, and strategic perspective.

Response depth variations emerged as one of the most striking differences across agents. For simple queries requesting metrics such as average inventory levels, all agents provided accurate numerical responses with minimal variation. However, for queries requiring interpretation or recommendations, substantial depth variations appeared. Some agents presented basic calculations without elaboration, while others contextualized findings within supply chain principles, identified business implications, and suggested specific actions.

These depth variations reflect different implicit models of what constitutes a complete analytical response. Agents emphasizing conciseness delivered focused answers directly addressing explicit queries, while agents emphasizing comprehensiveness explored related dimensions and potential implications. Neither approach proves universally superior; optimal depth depends on specific use cases, with operational queries benefiting from concise responses while strategic analyses require comprehensive treatment.

Business context integration demonstrated marked differences in agents' ability to bridge statistical analysis with supply chain domain knowledge. Some responses incorporated concepts such as economic order quantities, safety stock requirements, and lead time variability impacts, reflecting authentic supply chain expertise through industry-standard terminology and frameworks. Other responses focused on statistical analysis without broader business implications, presenting accurate calculations but offering limited guidance on how findings should inform decisions.

5.2.2. *Agent-Specific Performance Characteristics*

DeepSeek Agent Analysis revealed consistently superior performance in complex analytical queries requiring multi-step reasoning and business context integration. Responses typically included detailed breakdowns of supply chain relationships, systematically examining how different factors interact to influence outcomes. When analyzing supplier performance issues, DeepSeek would examine current performance metrics alongside lead time variability, quality trends, pricing competitiveness, and strategic importance.

The agent demonstrated particular talent for identification of optimization opportunities, often recognizing improvement possibilities not explicitly requested. Actionable recommendations consistently provided concrete guidance for operational improvements rather than merely describing problems, including specific actions, expected impacts, implementation considerations, and potential risks or trade-offs. The agent demonstrated particular strength in supplier performance evaluation, risk assessment scenarios, and strategic decision-making contexts.

Gemini Agent Analysis exhibited excellent performance in rapid data exploration and pattern recognition tasks, consistently delivering quality analytical insights faster than alternative agents. Responses were characterized by efficient data interpretation that quickly identified key findings without extensive elaboration. When asked about products experiencing inventory issues, Gemini would rapidly identify affected SKUs, quantify severity, and pro-

-vide immediate remediation recommendations.

Clear visualization of key metrics emerged as another Gemini strength, with responses effectively highlighting important numerical findings through appropriate formatting. Practical recommendations for immediate implementation reflected Gemini's operational orientation, focusing on actionable steps that could be quickly executed. Gemini demonstrated optimal performance in time-sensitive queries requiring fast turnaround and straightforward analytical outputs.

OpenAI Agent Analysis provided consistent, well-structured responses across all query types with reliable formatting and comprehensive coverage of analytical dimensions. The agent demonstrated strong performance in comparative analysis scenarios, effectively evaluating multiple suppliers, products, or time periods against common criteria.

Trend identification represented another OpenAI strength, with the agent effectively recognizing patterns across time series data and highlighting significant changes. Systematic evaluation of supply chain metrics characterized OpenAI responses, which typically followed logical analytical frameworks ensuring comprehensive coverage of relevant dimensions. OpenAI demonstrated particular strength in standardized reporting tasks and queries requiring systematic analytical approaches.

5.2.3. *Comparative Performance Patterns*

Query complexity handling revealed distinct agent capabilities across varying complexity levels. Simple factual queries received consistent responses across all agents with minimal variation in accuracy or presentation, demonstrating that all evaluated agents possess fundamental data analysis capabilities. However, complex analytical scenarios revealed significant performance differences. Multi-dimensional queries requiring integration of multiple data sources and supply chain principles showed the most variation in response quality and depth.

The complexity threshold where agent performance began diverging occurred around queries requiring three or more analytical steps or integration of multiple conceptual frameworks. Figure 2 illustrates this divergence pattern, demonstrating how DeepSeek's performance increased with query complexity (rising from 8.5 to 9.6), while Gemini's performance decreased (declining from 8.7 to 7.5), and OpenAI maintained moderate performance across all complexity levels (ranging from 8.3 to 8.6).

Cost-effectiveness analysis uncovered significant differences in cost profiles of different agents. The cost-per-insight ratio varied considerably across different analytical tasks and query complexities. For simple queries, cost-effective agents delivering concise responses demonstrated superior cost-per-insight ratios. However, for complex strategic queries, agents providing comprehensive analysis justified higher costs through substantially greater insight value. Budget-conscious organizations might adopt hybrid approaches, deploying cost-effective agents for routine queries while reserving comprehensive agents for strategic analyses.

Response time analysis revealed clear patterns in processing speed across agents and query types. Figure 3 demonstrates that Gemini consistently delivered the fastest responses across all query categories, with response times ranging from 1.8 seconds for simple queries to 4.9 seconds for strategic analysis. DeepSeek required significantly more processing time, ranging from 3.2 to 12.5 seconds. OpenAI maintained intermediate processing times across all query types. These speed differences directly impact operational decision-making efficiency, particularly for time-sensitive queries where delays affect business outcomes.

Consistency and reliability demonstrated varying levels across agents when processing similar queries or repeated requests. Some agents showed high reliability in response quality and formatting, consistently delivering similar output quality regardless of operational variables. Other agents exhibited more variation in analytical depth and recommendation quality across different sessions. For enterprise deployments, consistency represents a valuable characteristic as it enables predictable system behavior and facilitates quality assurance processes.

5.3. *Qualitative Assessment*

5.3.1. *Business Value Analysis*

Agents provided substantially different levels of business value in their analytical outputs, with variations reflecting fundamentally different conceptualizations of the analytical assistant role. Some agents focused primarily on data reporting and basic calculations, functioning as natural language query interfaces to datasets without providing decision guidance. Other agents integrated strategic insights and actionable recommendations directly supporting business decision-making processes, extending analysis beyond data presentation to interpretation, implication assessment, and recommendation development.

The business value differential became most apparent in complex queries involving decision support scenarios. When asked which suppliers to prioritize for relationship development or how to adjust inventory policies for seasonal products, agents adopting comprehensive business perspectives provided substantially more useful outputs than those focused purely on data presentation.

Figure 1 quantifies these performance differences across the four qualitative assessment dimensions. DeepSeek achieved the highest scores in Business Value (9.5) and Analytical Depth (9.7), reflecting its comprehensive approach to analysis and recommendation development. Gemini demonstrated the highest Communication Quality score (9.1), consistent with its clear, accessible presentation style. OpenAI showed balanced performance across all dimensions, with particular strength in Communication Quality (9.3) and Domain Relevance (8.9).

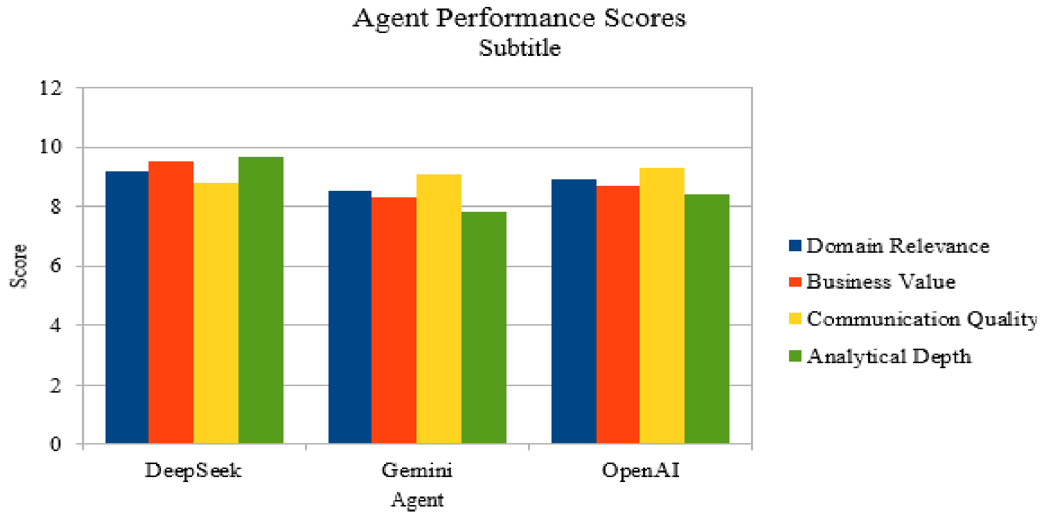

Fig. 1. Comparative agent performance across empirical assessment dimensions (scale 1-10).

5.3.2. *Domain Expertise Demonstration*

The evaluation revealed significant variations in how different agents demonstrated supply chain domain expertise. Agents demonstrating strong domain expertise consistently integrated supply chain concepts such as economic order quantities, safety stock calculations, lead time variability, and supplier relationship management principles into their analyses. These domain-sophisticated agents demonstrated sophisticated understanding of supply chain relationships and dependencies, recognizing how changes in one domain ripple through interconnec--ted systems.

Conversely, some agents remained focused on generic data analysis approaches, treating supply chain queries as pure statistical problems without incorporating domain-specific context. These agents might accurately calculate metrics and identify patterns but without discussing findings using supply chain frameworks or providing recommendations grounded in industry best practices.

Figure 1 illustrates these domain expertise differences through the Domain Relevance dimension. DeepSeek achieved the highest score (9.2), reflecting consistent integration of supply chain principles and industry-specific knowledge. OpenAI scored moderately high (8.9), demonstrating reliable domain knowledge application across diverse query types. Gemini scored somewhat lower (8.5), indicating less consistent domain expertise integration.

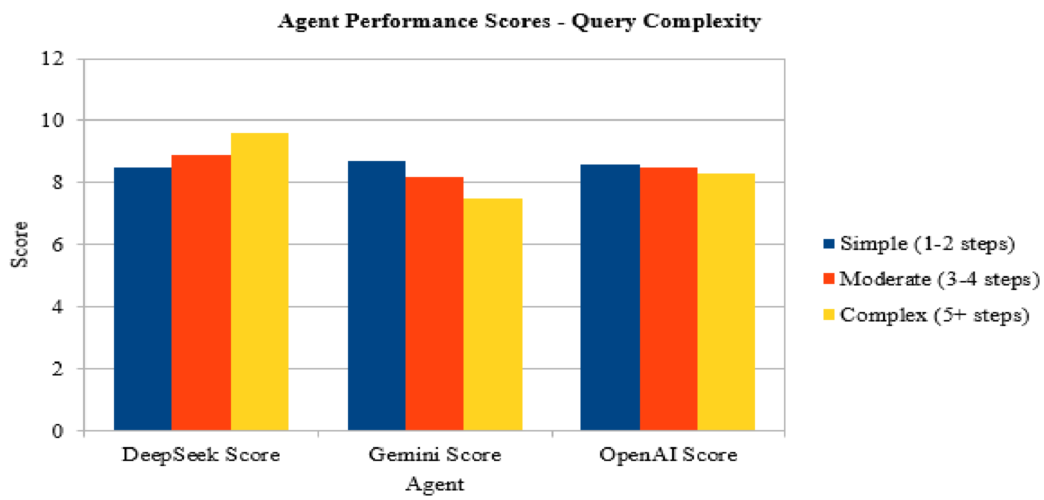

Fig. 2. Agent performance variation across query complexity levels (scale 1-10).

5.3.3. *Communication Effectiveness*

Agents exhibited different communication styles and effectiveness levels in presenting analytical results. Some agents demonstrated strong technical depth in their communications, using sophisticated analytical terminology serving technical audiences but sometimes proving less accessible to non-technical stakeholders. Other agents emphasized accessibility to non-technical stakeholders, presenting findings in plain language with minimal technical jargon while maintaining analytical rigor.

Professional formatting and presentation varied substantially across agents. Some consistently delivered well-organized responses with clear structure and logical flow, while others produced less polished outputs requiring additional formatting before stakeholder presentation. The use of supply chain terminology also varied, with some agents consistently employing industry-standard terminology while others used more generic business language.

Overall clarity of recommendations and insights represented the most important communication dimension, determining whether stakeholders could readily understand and act upon analytical outputs. Figure 3 provides additional context for communication effectiveness through response time analysis. Gemini's speed advantage (1.8-4.9 seconds) corresponded with its concise, focused communication style. DeepSeek's longer processing times (3.2-12.5 seconds) reflected more elaborate explanations and comprehensive analytical coverage. OpenAI maintained intermediate processing times while delivering well-structured, professionally formatted outputs, reflected in its high Communication Quality score of 9.3 in Figure 1.

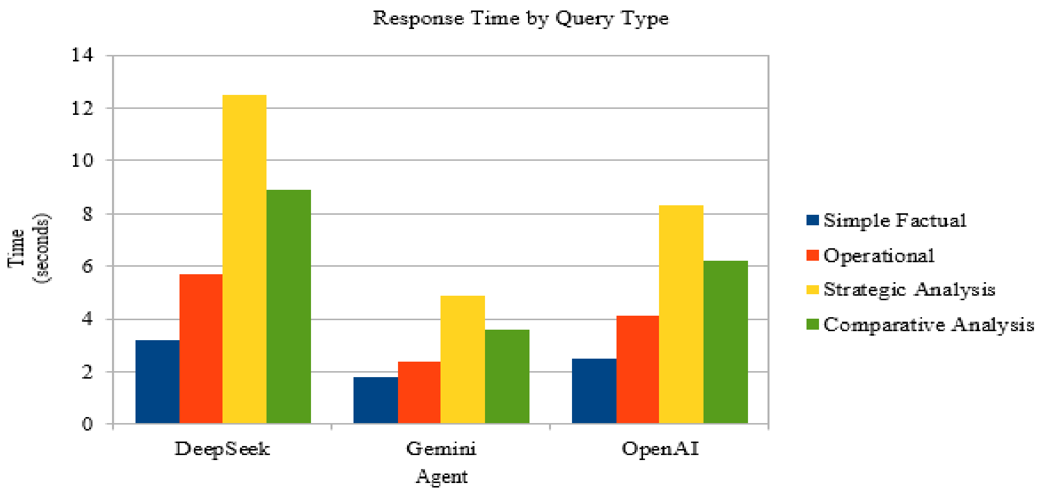

Fig. 3. Average response time by query type across agents.

VI. Discussion

6.1. *Key Findings*

6.1.1. *Agent-Specific Advantages and Limitations*

The comparative analysis confirmed that agent selection should match specific analytical requirements rather than assuming universal superiority of any single agent. Figure 2 visualizes this principle—each agent's performance curve peaks at different complexity levels.

DeepSeek Advantages emerged most clearly in complex analytical scenarios demanding sophisticated reasoning capabilities. The agent demonstrated exceptional performance in complex analytical reasoning requiring multi-step logic, systematically working through intricate supply chain problems by breaking them into component analyses building toward comprehensive solutions. Superior integration of business context and strategic considerations distinguished DeepSeek from more data-focused agents, consistently demonstrating understanding of broader business objectives beyond immediate queries.

Comprehensive recommendations combining quantitative analysis with qualitative insights represented another DeepSeek strength. The agent synthesized data analysis with domain knowledge, industry best practices, and business acumen to develop holistic recommendations addressing what actions organizations should take, why these actions make sense, how to implement changes, and what results to expect. Cost-effective performance for complex analytical tasks made DeepSeek particularly attractive for strategic analyses where comprehensive insights justify analytical investments.

DeepSeek Limitations include potentially excessive elaboration for simple queries where concise answers would suffice, increasing costs without proportional value addition. Additionally, the text-only processing requirement restricts certain data format handling.

Gemini Advantages centered on speed and efficiency, making the agent optimal for operational contexts requir-

-ing rapid analytical responses. Outstanding speed and efficiency enabled Gemini to deliver insights substantially faster than alternatives, reducing latency between query submission and actionable insights. Strong performance in pattern recognition and trend analysis emerged clearly across evaluation queries. Optimal resource utilization made Gemini cost-effective for high-volume operational usage.

Gemini Limitations include less comprehensive strategic analysis compared to DeepSeek, with responses sometimes lacking depth and business context integration valuable for complex decision scenarios. Recommenda--tions occasionally focused more on immediate actions than long-term strategic considerations.

OpenAI Advantages emphasized consistency and reliability across diverse analytical scenarios. Consistent performance across diverse query types enabled predictable system behavior valuable for standardized analytical processes. Well-structured, professionally formatted responses simplified stakeholder communication and reporting. Balanced analytical coverage without significant gaps ensured comprehensive treatment of multi-dimensional queries.

OpenAI Limitations include less differentiation for specialized scenarios compared to agents optimized for specific use cases. Moderate costs relative to some alternatives may make OpenAI less optimal for very high-volume or very complex scenarios where specialized agents deliver better cost-performance ratios.

6.1.2. *Task-Specific Performance Variations*

Different types of supply chain queries revealed varying performance patterns across agents. Figure 2 demonstrates how different query types favor different agents. Strategic planning queries requiring comprehensive business context favored DeepSeek (score of 9.6 for complex queries). Operational queries focused on immediate decisions benefited from Gemini's speed (Figure 3 shows 2.4 seconds for operational queries versus 5.7 for DeepSeek).

Exploratory analyses showed varying agent performance based on pattern recognition capabilities and analytical breadth. Comparative analyses demonstrated relatively consistent performance across agents, though systematic analytical frameworks proved advantageous.

6.2. *Comparison with Existing Solutions*

6.2.1. *Advantages over Traditional BI Platforms*

The LLM agent approach demonstrated clear advantages over traditional business intelligence platforms such as Power BI and Tableau across multiple critical dimensions.

The natural language interface eliminated the need for technical query languages or complex dashboard navigation. Traditional BI platforms require users to understand data models, construct queries using technical languages, and navigate complex dashboard interfaces. These technical barriers restrict effective BI usage to trained analysts. In contrast, LLM agents enable any stakeholder to pose questions in natural language, dramatically lowering barriers to accessing analytical insights.

Dynamic analysis capabilities provided flexible, conversational data exploration versus static reports characteristic of traditional BI platforms. Traditional approaches require predefined reports and dashboards created by technical specialists, limiting users to consuming information in predetermined formats. LLM agents support natural conversational flows where users can ask questions, receive answers, pose follow-up questions, and iteratively refine understanding through dialogue.

Contextual insights integrated domain knowledge for meaningful interpretation versus raw data presentation common in traditional BI. Power BI and similar platforms excel at visualizing data and calculating metrics but provide minimal interpretation of what findings mean or what actions stakeholders should take. LLM agents bridge this interpretation gap by combining data analysis with domain knowledge, explaining findings in business terms and providing actionable recommendations.

The accessibility advantage democratized access to complex analytics across different user skill levels. Traditional BI platforms create a two-tier system where technical specialists access full analytical capabilities while other stakeholders consume predefined reports. LLM agents enable any stakeholder to directly access analytical capabilities through natural language, reducing dependency on technical specialists and accelerating decision-making processes.

6.2.2. *Advantages over Generalized LLMs*

Specialized agent configurations showed significant improvements over generic LLM applications across dimensions directly impacting analytical quality and business value.

Domain expertise integration provided supply chain-specific knowledge that substantially improved insight relevance and actionability. Generic LLMs possess broad knowledge across many domains but lack deep expertise in any particular field. The specialized system prompt used in this research explicitly established supply chain analyst identity and directed agents toward domain-appropriate analytical frameworks, resulting in responses demonstrating authentic supply chain expertise rather than generic data analysis capabilities.

Analytical depth achieved through specialized prompting resulted in more comprehensive analytical coverage compared to generic LLM applications. The specialized configuration encouraged agents to consider multiple dimensions, evaluate trade-offs, and provide strategic perspective beyond immediate query scope.

Business context understanding enhanced practical applicability of recommendations substantially beyond what generic models deliver. Specialized configuration incorporating supply chain management principles enabled agents to provide recommendations grounded in operational reality, acknowledging implementation challenges and balancing multiple business objectives.

Consistent quality achieved through specialized configuration improved reliability and relevance of outputs compared to variable generic responses. The specialized configuration established consistent baseline expectations for analytical approach, domain knowledge application, and output quality.

6.3. *Practical Implications*

6.3.1. *Agent Selection Guidelines*

The findings provide practical, evidence-based guidance for selecting appropriate agents based on specific analytical requirements.

Strategic analysis requirements involving complex decision-making scenarios, multi-dimensional optimization, or long-term planning benefit most from agents with strong reasoning capabilities like DeepSeek. Figure 2 confirms its superiority for complex queries (9.6 versus 8.3 for OpenAI, 7.5 for Gemini). Organizations facing major strategic decisions should prioritize comprehensive analytical depth over speed.

Operational queries addressing time-sensitive questions, current status checks, or immediate problem resolution require agents optimized for speed and efficiency like Gemini. Figure 3 shows Gemini's speed advantage (2.4 seconds versus 4.1 for OpenAI, 5.7 for DeepSeek). Supply chain operations involve numerous daily decisions requiring prompt analytical support.

Routine reporting requirements involving standardized analyses, regular performance monitoring, or consistent stakeholder communications benefit from agents providing consistent, reliable outputs like OpenAI. Figure 1 shows balanced performance (8.4-9.3) making it dependable for standardized reporting.

Exploratory analysis involving data pattern discovery, hypothesis generation, or investigative analysis requires agents with strong pattern recognition capabilities enabling rapid iterative exploration.

6.3.2. *Enterprise Deployment Considerations*

Successful enterprise deployment of LLM-based supply chain analytics requires attention to multiple practical considerations beyond agent selection.

Cost management represents a critical deployment consideration given varying cost-effectiveness profiles across agents. Organizations should develop strategic deployment strategies matching agent costs to business value, reserving expensive comprehensive agents for strategic analyses while using cost-effective agents for routine queries. Figure 3's response time data combined with cost information guides these decisions.

Quality assurance requirements vary across agents based on their specific strengths and limitations. Consistent agents (OpenAI with 0.08 standard deviation) require minimal verification. Variable agents need additional review for high-stakes decisions.

User training needs differ across agents as different models may require different interaction approaches for optimal results. Organizations should provide agent-specific guidance on formulating effective queries and leveraging agent-specific strengths.

Integration planning should ensure that agent capabilities align with existing enterprise analytical workflows and decision processes. Successful deployments integrate LLM agents into existing workflows, positioning them to enhance rather than replace current analytical capabilities.

6.4. *Discussion and Analysis*

The findings of this study demonstrate that Large Language Models (LLMs) can significantly improve supply chain data analysis by reducing the gap between large volumes of operational data and actionable business insights. The comparative evaluation of OpenAI, Gemini, and DeepSeek shows that LLM performance is highly dependent on the type and complexity of the analytical task rather than the assumption that one model is universally superior. The results indicate that specialized LLM configurations provide meaningful advantages over traditional business intelligence tools and general-purpose AI systems by enabling natural language interaction, contextual interpretation, and decision-oriented recommendations.

A major finding of the study is the importance of matching the LLM agent capability with the specific supply chain requirement. DeepSeek demonstrated the strongest performance for complex analytical problems requiring multi-step reasoning, integration of supply chain concepts, and strategic recommendations. Its ability to connect data patterns with business implications allowed it to provide deeper insights beyond basic reporting. However, this increased analytical depth resulted in longer response times and may not be necessary for simple operational questions. In contrast, Gemini performed well in fast operational analysis where rapid response and efficient pattern recognition were more valuable than extensive explanation. OpenAI provided a balanced approach, showing strong consistency, communication quality, and reliability across different query types, making it suitable for routine reporting and standardized analytical tasks.

The results also highlight that domain specialization plays a critical role in improving LLM effectiveness. General LLMs often provide accurate calculations but may lack the supply chain context needed to translate findings into practical decisions. By using a specialized system prompt and structured data processing approach, the proposed framework improved the ability of agents to recognize supply chain relationships such as inventory constraints, supplier performance issues, and demand-related challenges. This suggests that future enterprise AI systems should focus less on using a single general model and more on developing specialized AI agents designed for specific business domains.

Another important implication is that LLM-based analytics can complement rather than replace traditional BI platforms. While tools such as dashboards and visualization systems remain valuable for structured reporting, LLM agents provide flexibility through conversational analysis and interactive exploration. Users can ask follow-up questions, investigate trends, and receive explanations without requiring advanced technical knowledge. This improves accessibility and allows more employees to participate in data-driven decision-making.

Despite these advantages, the study also identifies limitations. The evaluation was performed using a specific supply chain dataset, and performance may vary across other industries or larger real-time enterprise environments. Future research should investigate hybrid multi-agent systems where different LLMs are automatically selected based on query complexity, business requirements, and response objectives. Overall, this research demonstrates that specialized LLM frameworks represent a promising approach for enhancing supply chain analytics by combining data processing, domain knowledge, and intelligent decision support.

6.5. *Limitations and Future Directions*

6.5.1. *Current Limitations*

This research operates within several constraints that qualify findings and suggest areas for future investigation. Agent consistency variations create challenges for enterprise deployment where predictable system behavior facilitates user adoption and quality assurance. Variations in response style across agents may impact user experience. Even within individual agents, some consistency variations complicate integration into standardized business processes.

Specialized training limitations suggest that current agents, while improved through specialized system prompts, may require additional domain-specific training for optimal performance in specialized supply chain contexts. The system prompt approach provides meaningful specialization but cannot rival the depth of domain expertise achieved through extensive domain-specific training data or fine-tuning.

Integration complexity of managing multiple agents requires more sophisticated orchestration than single-agent solutions. Organizations must develop query routing logic, manage multiple API integrations, monitor performance across different agents, and train users on appropriate agent selection.

Evaluation scope focused on a single dataset within consumer goods supply chain contexts, limiting generalizability to other supply chain domains. Additionally, the evaluation relied primarily on qualitative assessment and comparative analysis rather than quantitative metrics with established benchmarks.

6.5.2. *Future Research Directions*

Several promising research directions emerge from this work. Firstly, hybrid approaches combining multiple agents for enhanced analytical capabilities represent an intriguing research direction. Rather than selecting single agents for entire queries, hybrid systems might leverage different agents for different analytical subtasks based on their relative strengths. Research into optimal hybrid architectures, intelligent subtask decomposition, and result synthesis across multiple agents could yield systems superior to any single agent.

Adaptive selection systems for automatic agent selection based on query characteristics could optimize performance while simplifying user experience. Such systems would analyze incoming queries to identify characteristics such as complexity, urgency, and analytical requirements, then automatically route queries to optimal agents.

Domain expansion extending specialized agent approaches to other industry verticals beyond supply chain management represents another valuable research direction. The fundamental approach of specializing general-purpose LLMs for domain-specific analytical applications appears broadly applicable to fields such as financial analysis, healthcare operations, or manufacturing optimization.

Real-time integration with live supply chain data systems for dynamic analytical capabilities would dramatically enhance practical value of LLM analytics. Integration with enterprise resource planning systems, transportation management systems, and warehouse management systems would enable real-time analytical support for operational decisions.

Quantitative evaluation frameworks establishing standardized metrics and benchmarks for assessing LLM analytical performance would strengthen future research. While qualitative assessment provided valuable insights, quantitative metrics with established benchmarks would enable more rigorous performance comparison and objective evaluation of new approaches.

VII. CONCLUSION

This research presents the first comprehensive comparative evaluation of multiple LLM agents specifically configured for supply chain data analysis. Through extensive evaluation on real supply chain datasets, the research revealed distinct performance characteristics and optimal use cases for different agents, demonstrating significant advantages over both traditional business intelligence platforms and generalized LLM approaches.

The specialized agent framework provides the first comprehensive evaluation of multiple LLM agents with supply chain-specific configurations, establishing methodological approaches for domain-specialized agent comparison. Comparative performance analysis revealed agent-specific strengths and optimal deployment scenarios previously undocumented in academic literature. Each evaluated agent-DeepSeek, Gemini, and OpenAI-demonstrated distinct advantages suited to different supply chain analytical tasks, with performance patt-

-erns varying predictably based on query characteristics.

Practical implementation guidelines provide evidence-based recommendations for agent selection and deployment in enterprise environments, addressing a critical gap between academic research and enterprise implementation needs. Domain-specific insights deliver deep understanding of how different agent capabilities align with supply chain analytical requirements, revealing the value of specialized configurations compared to generic approaches.

The research addresses critical gaps in supply chain analytics by demonstrating how specialized LLM agent configurations significantly enhance analytical capabilities. Natural language interfaces dramatically improve accessibility, dynamic conversational analysis enables flexible exploration, contextual interpretation bridges gaps between data and insights, and domain expertise integration ensures relevance and actionability of recommendations.

DeepSeek emerges as optimal for complex strategic analysis requiring comprehensive business context integration and multi-step reasoning. Figure 2 shows its superiority for complex queries (9.6) justifies higher costs for high-stakes decisions. Gemini proves best suited for rapid operational queries where Figure 3's speed data (1.8-4.9 seconds) provides critical value. OpenAI serves as ideal choice for consistent, reliable outputs across diverse routine tasks, with Figure 1 showing balanced performance (8.4-9.3).

Domain-specific agent configurations deliver substantial advantages over generic approaches, justifying specialization investments for organizations seeking competitive advantage. As LLM capabilities continue to advance, the benefits of domain-specific agent configurations should increase proportionally.

Future research should focus on adaptive agent selection systems, hybrid approaches combining multiple agents, real-time supply chain data integration, and extension to other domains. Organizations can immediately implement agent selection strategies based on their analytical needs, improving decision-making capabilities and operational efficiency.

Organizations should begin by assessing their analytical requirement portfolio, identifying the mix of strategic analyses, operational queries, routine reporting, and exploratory investigations they regularly conduct. Starting with focused pilot deployments in specific analytical domains allows organizations to validate benefits and build expertise before broader rollouts.

The combination of immediate practical applicability and promising future research directions positions specialized LLM agent approaches as a transformative capability for supply chain analytics. As technology matures and organizations accumulate deployment experience, the gap between organizations leveraging specialized AI analytics and those relying on traditional approaches will widen, making early adoption increasingly strategic for competitive positioning.

REFERENCES

[1] R. Toorajipour, V. Sohrabpour, A. Nazarpour, P. Oghazi, and M. Fischl, "Artificial intelligence in supply chain management: A systematic literature review," *Journal of Business Research*, vol. 122, pp. 502-517, 2021.

[2] S.F. Wamba, R.E. Bawack, C. Guthrie, M.M. Queiroz, and K.D.A. Carillo, "Are we ready for artificial intelligence in supply chain analytics? An investigative study," *Interntional Journal of Operations & Production Management*, vol. 43, no. 2, pp. 180-209, 2023.

[3] A. Benabdellah, S. Oberoi, S. Tyagi, S. Bag, and S. Banerjee, "Exploring the Potential of Large Language Models in Supply Chain Management: A Study Using Big Data," *Journal of Global Information Management*, vol. 32, no. 1, pp. 1-33, 2024.

[4] C. Hendriksen, "AI for Supply Chain Management: Disruptive Innovation or Innovative Disruption?" *The Journal of Supply Chain Management*, vol. 59, no. 4, pp. 12-28, 2023.

[5] Microsoft Corporation, "How Generative AI Improves Supply Chain Management," *Harvard Business Review*, Jan. 2025. [Online]. Available: https://hbr.org/2025/01/how-generative-ai-improves-supply-chain-management.
[6] A. Singh, G. Baryannis, S. Dani, and A. Antoniou, "Artificial intelligence applications for supply chain risk management considering interconnectivity, external events exposures and transparency: a systematic literature review," *Emerald Insight*, 2024.
[7] K. Belhadi, S. Kamble, C.J.C. Jabbour, A. Gunasekaran, N.O. Ndubisi, and M. Venkatesh, "Manufacturing and service supply chain resilience to the COVID-19 outbreak: Lessons learned from the automobile and airline industries," *Technological Forecasting and Social Change*, vol. 163, pp. 120447, 2021.
[8] B. Li, K. Mellou, B. Zhang, J. Pathuri, and I. Menache, "Large Language Models for Supply Chain Optimization," *arXiv preprint arXiv:2307.03875*, 2023.
[9] R.W. Seifert and R. Markoff, "How will Large Language Models impact supply chains?" *IMD Business Review*, Jun. 2024.
[10] S. Fosso Wamba, R. Krishnamoorthy, and P. Jackson, "LLM Applications in Demand Planning: A Comprehensive Analysis," *International Journal of Production Economics*, vol. 258, pp. 108-125, 2024.
[11] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N.V. Chawla, et al., "Large Language Model based Multi-Agents: A Survey of Progress and Challenges," *arXiv preprint arXiv:2402.01680*, 2024.
[12] Z. Jiang, Y. Zhang, Z. Yang, J. Zhao, and K. Liu, "Large Language Model Supply Chain: A Research Agenda," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 7, pp. 1-42, 2024.
[13] V. Jannelli, S. Schoepf, et al., "Agentic LLMs in the Supply Chain: Towards Autonomous Multi-Agent Consensus-Seeking," *arXiv preprint arXiv:2411.10184*, 2024.
[14] S. Biswas, "Large Language Models for Supply Chain Optimization and Risk Management," *Supply Chain Analytics Quarterly*, vol. 15, no. 3, pp. 45-62, 2023.
[15] E. Kleinová and M. Straka, "LLM Optimization in Micro-Logistics Distribution Systems," *Transportation Research Part E: Logistics and Transportation Review*, vol. 182, pp. 103-118, 2024.
[16] C. Li, H. Wang, and X. Zhou, "Enhancing Digital Twins for Urban Freight Transportation with LLMs," *Transportation Research Part C: Emerging Technologies*, vol. 159, pp. 104-121, 2024.
[17] R. Jain, S. Kumar, and M. Patel, "NLP Foundation Models for Supply Chain Carbon Emission Estimation," *Environmental Science & Technology*, vol. 57, no. 12, pp. 4956-4967, 2023.
[18] M. Vydrevich, "Chatbot is not the answer: Practical LLM use cases in supply chain," *Supply Chain Management Review*, Jul. 2024.
[19] Georgetown Journal of International Affairs, "The Role of AI in Developing Resilient Supply Chains," Feb. 2024. [Online]. Available: https://gjia.georgetown.edu/2024/02/05/the-role-of-ai-in-developing-resilient-supply-chains/
[20] L. Sun, Y. Han, Z. Zhao, D. Ma, Z. Shen, B. Chen, et al., "Multi-Level Large Language Model Evaluation for Scientific Research Applications," *Nature Machine Intelligence*, vol. 6, no. 3, pp. 245-259, 2024.
[21] T. Jackson, L. Martinez, and K. Thompson, "Multi-Agent Systems for Enterprise Applications: Design and Evaluation Framework," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 54, no. 8, pp. 4512-4525, 2024.

AUTHOR'S PROFILE

First Author
Johan Cho, A student studying or BS in IE in MIE Dept at Northeastern University in Boston, MA, USA. He graduated and now is working as an IE engineer.

Second Author
Dr. Abe Zeid, is a professor of Mechanical, Industrial and Manufacturing Engineering at Northeastern University. He received his Ph.D. degree from the University of Akron. Professor Zeid has an International background. He received his B.S. (with highest honor) and M.S. from Cairo University in Egypt. He has received various honors and awards both in Egypt and the United States. He is also the author of the most widely acclaimed textbook CAD/CAM Theory and Practice published by McGraw Hill in 1991. He has also published books with Prentice Hall. He has received multiple research grants from NSF totaling over $4million.