

Raman, Raghu; Sreenivasan, Aswathy; Suresh, M.; Gunasekaran, Angappa;
Nedungadi, Prema

Article

AI-driven education: a comparative study on ChatGPT and Bard in supply chain management contexts

Cogent Business & Management

Provided in Cooperation with:

Taylor & Francis Group

Suggested Citation: Raman, Raghu; Sreenivasan, Aswathy; Suresh, M.; Gunasekaran, Angappa; Nedungadi, Prema (2024) : AI-driven education: a comparative study on ChatGPT and Bard in supply chain management contexts, Cogent Business & Management, ISSN 2331-1975, Taylor & Francis, Abingdon, Vol. 11, Iss. 1, pp. 1-24,
<https://doi.org/10.1080/23311975.2024.2412742>

This Version is available at:

<https://hdl.handle.net/10419/326601>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<https://creativecommons.org/licenses/by/4.0/>

AI-driven education: a comparative study on ChatGPT and Bard in supply chain management contexts

Raghu Raman, Aswathy Sreenivasan, M. Suresh, Angappa Gunasekaran & Prema Nedungadi

To cite this article: Raghu Raman, Aswathy Sreenivasan, M. Suresh, Angappa Gunasekaran & Prema Nedungadi (2024) AI-driven education: a comparative study on ChatGPT and Bard in supply chain management contexts, Cogent Business & Management, 11:1, 2412742, DOI: 10.1080/23311975.2024.2412742

To link to this article: <https://doi.org/10.1080/23311975.2024.2412742>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 18 Oct 2024.



Submit your article to this journal [↗](#)



Article views: 1596



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 5 View citing articles [↗](#)

AI-driven education: a comparative study on ChatGPT and Bard in supply chain management contexts

Raghu Raman^a, Aswathy Sreenivasan^b, M. Suresh^b, Angappa Gunasekaran^c and Prema Nedungadi^d

^aAmrita School of Business, Amritapuri, Amrita Vishwa Vidyapeetham, Kollam, Kerala, India; ^bAmrita School of Business, Amrita Vishwa Vidyapeetham, Coimbatore, India; ^cSchool of Business Administration, Penn State Harrisburg, Middletown, PA, USA; ^dAmrita School of Computing, Amritapuri, Amrita Vishwa Vidyapeetham, Kollam, Kerala, India

ABSTRACT

This study conducts a comparative analysis of two prominent generative artificial intelligence (GAI) tools, ChatGPT and Bard, specifically in the context of supply chain management. Using a dataset of 150 certified supply chain professional questions, the models are evaluated on the basis of accuracy, relevance, and clarity, and *t* tests are employed to assess differences between the tools. ChatGPT outperforms Bard in both accuracy and relevance, with statistically significant results, whereas Bard demonstrated a slight edge in readability, scoring higher on the Flesch readability ease scale. Both models exhibited moderate to high cosine similarity for the majority of the questions, indicating closely aligned outputs. However, variations in their performance arose from differences in their underlying architectures – ChatGPT's iterative improvement process balances utility and safety, whereas Bard is designed with stricter safeguards to minimize misuse. These findings have important implications for the integration of GAI tools in educational settings, such as developing supply chain curricula and training materials requiring high accuracy and relevance. Additionally, the results suggest broader applications of the GAI in supply chain decision-making, operational efficiency improvements, and enhanced stakeholder communication. The study also highlights the importance of continuous model adaptation to ensure the ethical, safe, and effective use of AI technologies in professional settings. Future research could explore how real-time feedback loops impact AI performance and how diverse training datasets influence model accuracy and relevance across different industries, further advancing the role of AI in complex domains such as supply chain management.

ARTICLE HISTORY

Received 28 November 2023
Revised 19 September 2024
Accepted 1 October 2024

KEYWORDS

Supply chain; logistics; LLM; ChatGPT; Bard; readability; cosine similarity; information management; text mining; ethics

SUBJECTS

Continuing Professional Development; Educational Research; Information Technology; Artificial Intelligence

1. Introduction

Artificial intelligence (AI) has quickly emerged as a disruptive force in many sectors of the economy, including education. The capacity of generative AI (GAI) tools such as ChatGPT and Bard to produce language that resembles that of a human has made them popular in recent years. This means that there is a great deal of promise for improving learning experiences and making it easier to acquire complex knowledge (Dwivedi et al., 2023). These advanced language model-powered tools are increasingly being incorporated into curricula, enabling automated assessment, individualized learning, and content production help. Although artificial intelligence (AI) has many advantages for education, it is crucial to identify and resolve any potential biases and limits in these technologies. The fact that GAI techniques rely on large datasets with potential biases that could unintentionally be reinforced in the created content is one of their main limitations. This may result in the spread of false information or cultural insensitivity, which is especially problematic in educational contexts where inclusivity and truth are crucial. Teachers may find it challenging to fully accept or comprehend AI-generated outputs because of the opaque nature of AI decision-making processes,

sometimes known as the ‘black box’ problem. This makes it difficult to grasp how these tools arrive at their findings.

Generative artificial intelligence (GAI) tools such as ChatGPT and Bard, which are based on large language models (LLMs), use deep learning to process and comprehend human language (Shen et al., 2023). These models can understand syntax and context, and they can even produce text responses resembling those of humans because they are trained on enormous volumes of textual data. They are now more crucial than ever for various NLP activities, including chatbots, sentiment analysis, text summarization, and language translation. GAI tools’ ability to produce coherent and contextually appropriate text makes them essential for automating language-based tasks and improving human–computer interactions (Abd-Alrazaq et al., 2023). They can continuously advance and change to meet new problems because of their capacity to learn from data. LLMs use their text-generating capabilities to enable generative AI systems (Feuerriegel et al., 2023). These technologies have a wide range of applications in fields such as marketing, customer service, content generation, and creative writing since they can generate original and contextually appropriate responses, mimic conversations, produce text that sounds like a real person wrote it and help with content creation.

LLMs with unique characteristics and applications include ChatGPT and Bard (Floridi, 2023). The GPT-3.5-based ChatGPT system, created by OpenAI, is intended for interactive text-based dialogs (Zhao et al., 2023). It is perfect for chatbots, customer assistance, content generation, and creative writing since it can produce coherent and contextually appropriate responses (Javaid et al., 2023). Google created Bard, which aims to comprehend and produce words in a more conversational and lively way (De Ruiter et al., 2006). It seeks to engage users in more dynamic and open-ended conversations to make interactions feel more human.

ChatGPT and Bard have demonstrated promise in various applications, from fact-checking, content creation, and human-like interactions to medical education and specialized industries (Mohamadi et al., 2023). They are valuable tools in many disciplines since they may produce contextually appropriate replies, but researchers are still investigating their advantages and disadvantages in many situations.

For several reasons, LLMs’ grasp of particular subjects is highly pertinent and valuable. First, having expertise in specific disciplines enables LLMs to provide users with more accurate and relevant information on the basis of their context. The quality of the responses is improved, and there is more confidence in the model’s outputs because of the higher accuracy. LLMs with domain-specific understanding have a wide range of uses. Individuals can use them to carry out specialized tasks such as financial modeling, legal research, and medical diagnosis. It is crucial to be aware of the consequences of such comprehension. It is essential to guarantee the accuracy and objectivity of the data used to train the models to avoid perpetuating any biases already present in the data. Open communication regarding the boundaries of their knowledge is also necessary to prevent potential abuse and excessive reliance on LLM outcomes in situations requiring crucial decision-making.

Supply chains were chosen as the research topic because they significantly impact many different organizations and economies. The study of supply chains has practical applications because of its crucial role in guaranteeing effective operations, cost reduction, and customer satisfaction. The capacity of GAI tools, such as ChatGPT and Bard, to grasp and produce precise responses. Additionally, supply chains offer a perfect setting for objectively comparing and assessing ChatGPT and Bard. Evaluating the clarity and accuracy of GAI tools’ responses across many models and versions is simpler since supply chain processes entail standardized procedures and well-defined key performance indicators (KPIs).

This study postulates that because of its extensive training on a variety of datasets and incremental improvements, ChatGPT outperforms Bard in terms of accuracy and relevance when answering queries related to supply chain management. Although this hypothesis recognizes the possibility for biases and the limitations of AI tools in educational environments, it is based on the hope that the domain-specific understanding of LLMs improves response quality.

The following research questions are answered in this study:

1. RQ1: How do ChatGPT and Bard compare in terms of accuracy when responding to supply chain certification exam questions?

2. RQ2: Does posing the confirmation query, 'Are you sure?' lead to improved response accuracy for ChatGPT or Bard?
3. RQ3: How do the explanations provided by ChatGPT and Bard rank on relevance and clarity?
4. RQ4: Is there a correlation between the accuracy of the AI response and the clarity or relevance scores assigned to the explanation?
5. RQ5: Is there a statistically significant difference in the accuracy of responses between ChatGPT and Bard when addressing MCQs from the supply chain certification exam?
6. RQ6: When comparing the relevance scores of ChatGPT and Bard, is there a significant distinction in how pertinent their answers are to the supply chain certification MCQs?
7. RQ7: Considering the clarity of explanations provided, is there a discernible difference in scores between ChatGPT and Bard's responses to the MCQs?

With respect to their capabilities, this study compares ChatGPT and Bard, two recent GAI tools. The study intends to evaluate their performance in comprehending and answering supply chain-related questions to determine their correctness and comprehensibility in a challenging and real-world setting. The comparison will provide researchers, companies, and industries with insight into the benefits and drawbacks of each GAI tool, assisting them in understanding prospective applications and making defensible choices regarding their use in supply chain management and related contexts.

2. Literature review

Large language models (LLMs) are a class of artificial intelligence (AI) models that process and produce human-like language via deep learning techniques (Bonner et al., 2023). These models stand out because of their vast scale, which enables them to detect complex patterns in language data. They frequently include tens of billions of parameters. The advent of LLMs, such as OpenAI's GPT-4, has transformed natural language processing tasks, which allow machines to comprehend and produce human language with astounding fluency and contextuality (Khan et al., 2023). The transformer architecture used by LLMs enables them to analyze data sequences and comprehend the links between words and phrases in a text (Luitse & Denkena, 2021). Their capacity to provide coherent and contextually relevant responses depends on this design.

LLMs have considerably improved generative AI tools by offering cutting-edge language generation capabilities (Luitse & Denkena, 2021). Two well-known LLMs that have drawn much interest in the AI research community are ChatGPT and Bard. The capacity of Bard to use and incorporate information from the internet in real-time when creating responses sets it apart from ChatGPT and GPT-4. However, efforts to directly compare the two models are still in their early stages (Ali et al., 2023). Notably, there have not been any direct contrasts between Bard and ChatGPT for supply chain-related certification exams.

Researchers have examined both ChatGPT and Bard in a variety of contexts. The effectiveness of LLMs, including ChatGPT and Bard, in medical education and specialized medical professions, including neurosurgery, radiation oncology physics, and nephrology, was evaluated in studies by Ali et al. (2023), Holmes et al. (2023), and Noda et al. (2023). They looked at how LLMs could help with patient-clinician interactions, the development of practice question banks, and the assessment of medical expertise. When 82.6% of the questions were answered correctly, GPT-4 outperformed GPT-3.5 (62.4%) and Bard (44.2%), according to Ali et al. (2023). The imaging and accuracy of GPT-4 were superior to those of the other methods ($P=0.05$). Holmes et al. (2023) reported that ChatGPT (GPT-4) performed consistently and deductively better than LLMs and medical physicists did. A majority vote did not result in a better ChatGPT (GPT-4) outcome. Noda et al. (2023): GPT-4 demonstrated 54.5% accuracy, significantly outperforming GPT-3.5 (31.3%) and Bard (32.3%). The GPT-4 scores were comparable to those of seasoned residents. Caramancion (2023) study includes the use of a variety of LLMs, such as ChatGPT and Bard, to evaluate the accuracy of news reports. This demonstrates their potential for debunking false information and fact-checking. The capacities of LLMs in logical reasoning and cognitive inference, which are essential to human intelligence, were investigated by Xu et al. in 2023. Patnaik and Hoffmann (2023) compared ChatGPT and Bard's responses to patient-focused anesthesia queries. The remarks from ChatGPT were comprehensive,

wise, and valuable. Doshi et al. (2023) investigated the use of LLMs, such as ChatGPT and Bard, to increase patient comprehension of radiological findings and make them easier to understand. The study used LLMs to analyze 254 radiological reports using three prompts. With more context, ChatGPT performed better and streamlined reporting. Patient outcomes, communication, and health literacy might all improve. The readability and clinical accuracy, however, still need improvement. The study by Raman, Mandal, et al. (2024) analyzes ChatGPT adoption among higher education students, using a mixed-method approach to assess five key innovation attributes shaping their intentions and perceptions. Agarwal et al. (2023) assessed the applicability of ChatGPT and Bard in developing reasoning-based multiple-choice questions for medical students studying physiology. Khademi (2023) evaluated how writing prompts produced by ChatGPT and Bard were perceived by people and how complex they were. The OpenAI ChatGPT and the Google Bard both had low interreliability compared with the industry standard of human evaluations, according to the performance parameter known as intraclass correlation (ICC). More recently ChatGPT and Bard have been compared across various domains, including digital governance (Raman, Nair, et al., 2024), the UN Sustainable Development Goals (Raman, Lathabai, et al., 2024), ethical hacking (Raman, Calyam et al., 2024) and human resource management (Raman, Venugopalan, et al., 2024).

Table 1 demonstrates the synthesis of the past literature on both ChatGPT and Bard in a variety of contexts:

Having one's skills and knowledge in supply chain management officially recognized and validated is known as supply chain certification (Centobelli et al., 2022). Supply chain management is essential in today's digital world for guaranteeing the smooth movement of products, services, and information from raw material suppliers to final consumers (Gunasekaran & Ngai, 2004). It includes several phases: logistics, distribution, production, and procurement. A well-managed supply chain can result in cost savings, greater customer satisfaction, and increased corporate competitiveness (Chou et al., 2004). Globalization, rising customer expectations, and technological improvements have all contributed to an exponential rise in the relevance of supply chain management (Ahmed et al., 2018). Professionals require specialized knowledge and abilities to address issues, including demand volatility, inventory optimization, risk management, and sustainability, as supply chains become more complex and linked.

It is advantageous for LLMs to comprehend intricate and specialized sectors such as supply chain management. With respect to jobs involving supply chain planning, optimization, and decision-making, LLMs can better understand the context and produce more accurate and pertinent solutions by knowing supply chain concepts (Ansari & Kant, 2017). With this knowledge, LLMs may assist in creating backup plans for firms, resolving supply chain issues, and recommending best practices.

The lack of research on how LLMs affect specialized domains such as supply chain management constitutes a research gap. Although LLMs demonstrate remarkable performance in natural language processing tasks, more studies are needed to determine how well they understand the nuances of a given domain and how well they can apply them to real supply chain management problems. By examining ways to use LLMs to improve supply chain decision-making, optimization, and overall efficiency in the context of the modern digital world, this study seeks to close this gap.

2.1. What is the APICS-certified supply chain professional (CSCP) program?

The role of the supply chain has assumed unprecedented significance in today's increasingly interconnected world, operating as the backbone of company operations, ensuring seamless access to critical items for consumers, and constituting a pillar for social functionality. The APICS Certified Supply Chain Professional (CSCP) certification stands out in this context as a veritable playbook, giving people the knowledge required to comprehend and successfully coordinate this crucial process. The all-inclusive CSCP programme acts as a thoroughfare that navigates the vast terrain of the global supply chain, plotting a course from suppliers to end users and back again. The APICS exam is the only global supply chain certification. It equips the individual with the knowledge and ability to manage operations efficiently along the worldwide supply chain, including interactions with suppliers, factories, distributors, and customers worldwide. Achieving prestigious CSCP accreditation denotes more than experience; it denotes managing risks and anticipating and avoiding interruptions. Additionally, it represents the capacity to evaluate, accept, and integrate cutting-edge supply chain technologies that characterize the

Table 1. Synthesis of past literature comparing ChatGPT and Bard.

Sl no	Reference	Aim	Findings	Limitations
1	Ali et al. (2023)	To evaluate how well three LLMs (GPT 3.5, GPT 4, and Google Bard) performed on a question bank created especially for practice for the neurosurgery oral boards exam.	The answer rates for the GPT-3.5 and GPT-4 were 62.4% (CI: 54.1–70.1%) and 82.6% (CI: 95.2–88.1%). Bard received a score of 44.2% (confidence interval [CI]: 36.2–52.6%), ($P=0.023$ and $P=0.01$) GPT-4 outperformed GPT-3.5 and Bard. Not GPT-4, but Bard's accuracy (OR = 0.76, $P=0.014$) and GPT-3.5 (OR = 0.80, $P=0.042$) were negatively impacted by problem-solving. GPT-4 performed better in imaging (68.6% vs. GPT-3.5's 47.1%, $P=0.044$; tied with Bard at 66.7%, $P=1.000$) and had fewer hallucinations (2.3% vs. GPT-3.5's 57.1%, $P0.001$; Bard's 2.09, $P=0.002$). Without image descriptions, GPT-3.5 and Bard had increased hallucination probabilities (OR = 1.45, $P=0.012$ and OR = 2.09, $P0.001$, respectively).	MCQs might not replicate the open format of neurosurgical exams. Open-ended testing will be investigated in the future. Future studies will examine this despite minor changes (under 5%) in evaluator judgments of question qualities. Future studies will evaluate the validity of incorporating image data into LLMs, which has not yet occurred.
2	Caramancion (2023)	The project aimed to assess how well-known LLMs – OpenAI's ChatGPT 3.5 and 4.0, Google's Bard/LaMDA, and Microsoft's Bing AI – performed black box testing to determine the veracity of news stories.	Researchers evaluate the efficacy of LLMs by comparing LLM classifications with confirmed data from independent agencies. Average rating of 65.25/100, which indicates fair competency. GPT-4.0 scored 71, demonstrating the advantage of more recent LLMs in detecting fact-deception. However, AI must improve human fact-checkers understanding of complex news circumstances. The results show ongoing AI development and human cognitive abilities, emphasizing AI's potential for fact-checking.	The study emphasizes the possibilities and limitations of LLMs in eradicating false information – score on a scale of 100, 65.25. Analysis of ChatGPT 3.5, 4.0, Bard/LaMDA, and Bing AI. GPT-4.0 took the lead with 71 points, although it lacks the nuances of human fact-checkers like PolitiFact and Snopes. AI assists but cannot replace human inspection.
3	Rahsepar et al. (2023)	To assess and compare the consistency and correctness of answers provided by publicly accessible ChatGPT and Google Bard in response to questions from laypeople about lung cancer screening, prevention, and terminology used in radiology reports.	120 Questions, 70.8% Correct, 11.7% Partial, 17.5% Wrong on ChatGPT-3.5. Bard: 51.7% accuracy, 19.2% missed answers, and 80.8% right – Bing: 27.5% incorrect, 10.8% incomplete. Google searches are 55% accurate, 22.5% incomplete, and 22.5% incorrect. Compared to Bard, ChatGPT-3.5 is 1.5 times more precise/partial (OR = 1.55, $P=0.004$). When compared to Bard, Google search is 28.83 ($P=.002$) times more trustworthy than ChatGPT-3.5 (6.65 ($P=0.002$)).	Responses are produced by ChatGPT and Google Bard using sophisticated algorithms and training data. However, depending on the input wording and the particular model version employed, their responses can differ significantly, resulting in inconsistent answers.
4	Holmes et al. (2023)	The study suggests evaluating LLMs in the specialized field of radiation oncology physics. This is crucial for scientific and medical communities and provides a valuable benchmark. Utilizing a Mayo Clinic-created exam of 100 questions, LLMs like ChatGPT (GPT-3.5), ChatGPT (GPT-4), Bard (LaMDA), and BLOOMZ were tested against medical physicists and laypeople.	On average, ChatGPT (GPT-4) performed better than all other LLMs and medical physicists. It excelled even more when given explanations. Contrary to humans, ChatGPT (GPT-3.5 and GPT-4) consistently chose answers throughout trials, correctness or not. ChatGPT (GPT-4) fared well in novel deductive reasoning, indicating a growing skill. Despite excellent results, a majority vote could not further enhance ChatGPT (GPT-4). Collective decision-making may significantly enable medical physicists to surpass ChatGPT (GPT-4) in performance.	The complex clinical practice of radiation oncologists and medical physicists may have needed to be better represented by the exam utilized to evaluate LLMs in this study. Therefore, comparing LLMs and medical physicists based on equal performance may be inaccurate. The achievement of GPT-4 on a test that resembled a specialized certification raises questions about the breadth of its understanding. A medical physicist's duties are unlikely to be replaced by GPT-4 unless we consider a very unusual situation.
5	Xu et al. (2023)	Logic continues to be essential in knowledge engineering and AI. NLP's LLMs, a recent development, perform exceptionally well on typical NLP tasks. However, whether they possess human intelligence-like reasoning skills is still being determined. With in-depth studies, this work fills this gap.	The study categorizes datasets, tests LLMs in diverse scenarios, and evaluates them objectively and subjectively. Researchers construct a six-dimensional universal evaluation system for logical reasoning proficiency using a fresh, unbiased dataset. This work discusses LLMs' strengths, weaknesses, and future research directions.	Overall, observers may see that all LLMs have significant limitations in their logical thinking, showing more competency in deductive reasoning while facing obvious difficulties in inductive circumstances. Additionally, the present evaluation benchmarks, which mostly rely on objective criteria, must be revised to assess LLMs fully.

(Continued)

Table 1. Continued.

Sl no	Reference	Aim	Findings	Limitations
6	Patnaik and Hoffmann (2023)	Before surgery, they looked into how well LLMs could respond to patient-oriented questions on anesthesia. In the study, they explored 'hallucinations' phenomena, created textual data evaluation metrics, and assessed the viability of employing LLMs at the patient-clinician interface. Compared to Bard, ChatGPT was deemed more thorough, insightful, and valuable in its response.	ChatGPT showed no 'hallucination' issues, while Bard had a 30.3% error rate. Bard's comments were more casual and 8th-grade level, while ChatGPT's were challenging (college-level). ChatGPT had 19.7% higher linguistic quality than Bard, irrespective of response length (66.16 vs. 55.27; $P=0.0037$). Bard's sentiment was more favorable (0.16 vs. 0.11; $P=0.0323$), with similar subjectivity (0.54 vs. 0.50; $P=0.3030$). LLM responses are suitable as a flexible healthcare aid, not a replacement for vital counseling.	Access to the previous version of ChatGPT was restricted owing to software upgrades, so they only used the accessible version of ChatGPT (GPT-3.5 based) in this study. They could have detected more variations in the quantifiable metrics between the two versions of ChatGPT. It is necessary to address user variation in ChatGPT and Bard answers.
7	Doshi et al. (2023)	The study explores using LLMs, specifically OpenAI's ChatGPT3.5, ChatGPT4.0, Google Bard, and Microsoft Bing, to streamline radiology reports and potentially improve patient comprehension.	The study used LLMs to analyze 254 radiological reports using three prompts. With more context, ChatGPT performed better and streamlined reporting. Patient outcomes, communication, and health literacy might all improve. Readability and clinical accuracy, however, still need improvement.	Since the readability criteria utilized in this study are language- and structure-focused, they have comparable limitations. They only sometimes accurately reflect relevance or comprehensibility from a medical standpoint.
8	Agarwal et al. (2023)	To create reasoning-based multiple-choice questions (MCQs) for MBBS (Bachelor of Medicine, Bachelor of Surgery) undergraduate students studying physiology, they set out to evaluate and compare the applicability of ChatGPT, Bard, and Bing.	For ChatGPT and Bard, 110 MCQs were generated; however, Bing only generated 100 because of the absence of two skills questions. ChatGPT scored 3 (1.5–3), Bard scored 3 (1.5–3), and Bing scored 1 (0–1). MCQ validity revealed significant differences ($P=0.001$). The difference in MCQ difficulty across ChatGPT 1, Bard 1, and Bing 1 was statistically significant ($P=0.006$). ChatGPT, Bard, and Bing scored 1 (1–2) on the required reasoning ability ($P=0.235$). K values for all AI models were less than 0.8 for all parameters.	The study was restricted from generalization because it only focused on AI-generated medical physiology MCQs. When using ChatGPT, relying on a single user increases the chance of inconsistent response times. Question rephrasing can change responses, which might impact the evaluation. Despite efforts, subjective assessment by human evaluators could still result in prejudice. Analysis of the MCQ items involving MBBS students is still in progress.
9	Noda et al. (2023)	Large language models (LLMs) trained on massive data are essential for new AI advancements. GPT-4 performs exceptionally well in general medicine, although it could perform better in nephrology. In the trial, researchers assess ChatGPT and Bard for their potential use in clinical nephrology settings.	GPT-3.5 had an overall correct answer percentage of 31.3%, GPT-4 had a 54.5%, and Bard had a 32.3%. GPT-4 performed better than GPT-3.5 and Bard ($P=0.01$). Within three years, GPT-4 passed all tests. Compared to GPT-3.5 and Bard, it performed exceptionally well on the clinical and nonimage questions. Between third and fourth-year nephrology, residents had the highest rate of GPT-4.	LLMs need to improve their image questions to maintain accuracy. GPT-4 updates in the future will improve this. Multilingual evaluations are required but are only available to Japanese people. We investigate coauthor biases and question leaking. Due to potential biases, researchers are assessing the accuracy of the correct response percentages for resident participants.
10	Khademi (2023)	The study evaluates how well-educated and skilled humans perceive and rate the complexity of writing prompts compared to OpenAI ChatGPT and Google Bard.	The OpenAI ChatGPT and the Google Bard both had low interreliability compared to the industry standard of human evaluations, according to the performance parameter known as intraclass correlation (ICC).	

contemporary environment. The CSCP's core curriculum covers many essential topics, such as the complexity of supply chains, demand management, forecasting, and the dynamics of global supply chain networks. It explores the nuances of acquiring goods and services, simplifying internal processes and inventory control, coordinating forward and reverse logistics, building strong supply chain relationships, and negotiating the intricate world of supply chain risk. The exam also explores the fields of technology, sustainability, and optimization, combining vital strategic components that will influence supply chain management in the future. The candidate will have 3.5 hours to finish the 150-question exam (130 operational and 20 pretest). Candidates should answer all exam questions. The CSCP test has a 200–350-point scale. A failing grade has 299 or fewer points, whereas a passing grade has 300 or more points.

Our mission was to determine the capabilities of AI-driven systems such as ChatGPT and Bard, so we needed to choose a reliable and trustworthy dataset. The dataset of 150 Certified Supply Chain Professional (CSCP) questions was carefully selected because it mirrors complex, real-world supply chain management scenarios that professionals face. These questions represent a comprehensive and widely recognized certification, ensuring that the dataset is both rigorous and representative of typical challenges in the field. This allows for a more realistic evaluation of generative AI performance in professional contexts.

3. Methods

3.1. Study design

The research employs a comparative analysis methodology to evaluate the performance of ChatGPT and Bard in the context of supply chain management. A dataset was constructed via multiple-choice questions (MCQs) from the APICS CSCP exam, covering topics such as logistics, procurement, and related domains. Each MCQ featured multiple answer options, with one correct answer and multiple incorrect choices. The subsequent procedures delineate the specific steps taken to assess the capabilities of ChatGPT and Bard in answering these certification exam questions.

1. **Question input:** ChatGPT and Bard received each supply chain certification exam question. They received the initial responses from the AI models, which involved giving them the wording of the inquiry.
2. **Initial response:** ChatGPT and Bard produced recorded responses to the input questions. The initial responses of the AI models are their first attempt to respond to the inquiries.
3. **Confirmation query:** An additional question was put to the appropriate model if the original response from either ChatGPT or Bard did not match the accurate response. The AI model was asked, 'Are you sure?' to get it to rethink or reevaluate its earlier response.
4. **Second response capture:** The AI model produces a response after posing the confirmation query and recording it. This process tracks any adjustments or changes the AI model makes in response to the confirmation request.
5. **Explanation capture:** We record the explanation for each response, regardless of whether the second response to a question was accurate (we recorded and analyzed responses for three measures: correctness, relevance, and clarity). This defense explained why the AI model produced a particular response.

3.2. Evaluation metrics

The given assessment metrics provide a formal method for classifying and rating the comments produced by ChatGPT and Bard for the study. Among the most popular generative AI language models currently available are ChatGPT and Bard. Because of their propensity to produce writing that resembles that of a human, they are well suited to assess the potential of artificial intelligence in professional settings such as supply chain management. Bard and ChatGPT are trained on various datasets and have distinct underlying architectures. While Google developed Bard via its proprietary architecture, ChatGPT is built on the OpenAI-derived GPT-4 framework. The study's ability to evaluate and contrast the models' performances is made possible by their diversity in architecture and training data, which offers insights into how various methods for evaluating the AI development effect results in particular domains. Professional applications such as customer service, content production, and decision assistance have made use of both ChatGPT and Bard. Through an assessment of their responses to questions on supply chain certification, the study seeks to determine their suitability and efficacy in a particular field of expertise.

We rate the metrics on a 5-point Likert scale, with thorough explanations for each level, emphasizing accuracy, relevance, and clarity. The scoring system is designed to comprehensively analyze the performance of AI models by capturing the subtleties of their explanations and responses. Additionally, the

inclusion of an adjudicator and the participation of subject matter experts both strengthen and increase the trustworthiness of the review process.

- **Accuracy:** This metric considers how closely the given response matches the correct response to assess its accuracy.
- **Relevance:** This measure evaluates how well the answer corresponds to the context of the question. It focuses on how well the response fits the purpose of the query.
- **Clarity:** This measure assesses how simple it is to comprehend the response. It considers the explanation's coherence and understandability.

This study intends to provide a thorough assessment of model performance in a professional setting by selecting a generative AI model that is widely used, diverse, and capable of handling complicated linguistic problems, as well as a dataset that is typical of real-world supply chain management scenarios. With this method, experts may utilize generative AI to improve their decision-making and problem-solving skills in real-world scenarios, ensuring that the findings are applicable.

3.2.1. Scoring scale

Each metric is graded on a Likert scale ranging from 1 to 5, with 5 being the highest score, with clearly stated criteria for each level. This methodical technique helps reduce subjectivity and ensures consistency in ratings.

a. **Accuracy**

1. Completely inaccurate: The response is entirely incorrect or false.
2. Mostly inaccurate: The response contains some truth but is largely incorrect.
3. Somewhat accurate: The response is a mixture of correct and incorrect information.
4. Mostly accurate: The response is correct but may contain minor inaccuracies.
5. Completely accurate: The response is entirely correct and truthful.

b. **Relevance**

1. Completely irrelevant: The response does not address the question or prompt.
2. Mostly irrelevant: The response somewhat addresses the question but is mainly off-topic.
3. Somewhat relevant: The response addresses the question somewhat but could be more focused.
4. Mostly relevant: The response is on-topic and closely aligns with the question, with minor deviations.
5. Completely relevant: The response addresses the question or prompts and remains on-topic.

c. **Clarity**

1. Very confusing: The response is incoherent and difficult to follow.
2. Mostly confusing: The response has some clear elements but is overall hard to understand.
3. Moderately clear: The response is understandable but could be more precise or concise.
4. Mostly clear: The response is clear, with only minor ambiguities or complexities.
5. Completely clear: The response is articulated clearly and coherently.

3.3. Analysis of responses

Two independent, qualified supply chain experts read each explanation text from ChatGPT and Bard, and it was then scored on a scale of 1 to 5 for each of the three measures. Some explanations might be more straightforward and require a small quantity of subjective interpretation.

We used a third qualified supply chain expert as an adjudicator to confirm the scoring procedure's validity and uniformity. When the first two experts could not agree on a score, the expert was consulted.

The third expert's job was to evaluate the explanatory text and determine which should be regarded as the final version.

3.3.1. Statistical analysis

A statistical method to evaluate the variations in scores between ChatGPT and Bard for each of the three measures – accuracy, relevance, and clarity – is provided by using paired sample t tests in your study (Roscoe et al., 2023). When two sets of observations are contrasted, where each observation in one set is paired with a corresponding observation in the other, paired sample t tests are acceptable (Rouder et al., 2009). The observations in this instance are the marks given to ChatGPT and Bard for each MCQ question.

- a. **Accuracy t test:** *To compare the accuracy ratings of ChatGPT and Bard for each MCQ, we used a paired t test.*
- b. **Relevance t test:** *To compare the relevance scores of the two AI models for each MCQ, we conducted another paired t test.*
- c. **Clarity t test:** *To compare the clarity ratings of the justifications offered for each MCQ response between ChatGPT and Bard, we undertook a third paired t test.*

3.3.2. Readability

Undoubtedly, when assessing the responses produced by ChatGPT and Bard, taking into account readability metrics such as the Flesch readability ease score and the Flesch–Kincaid grade level score is a valuable approach, particularly when the target audience includes nonexperts or when the text is accessible to a wide range of readers.

- a. **Flesch readability ease score:** *The Flesch readability ease score is a metric that expresses how simple it is to comprehend a text (Eleyan et al., 2020). The greater the score is, the easier the text is to grasp. This metric considers elements such as the number of syllables in each word and the length of the sentences. To determine which model's responses are more approachable and reader friendly, the Flesch readability ease scores of ChatGPT and Bard should be compared (Haver et al., 2023).*
- b. **Flesch–Kincaid Grade Level Score:** *The Flesch–Kincaid Grade Level score determines what grade level of education in the United States is necessary to understand a particular text (Williamson & Martin, 2010). A lower score indicates that the material is simple to read and understandable to a larger audience. A higher score, on the other hand, denotes more challenging material appropriate for advanced readers. By comparing the Flesch–Kincaid Grade Level scores, researchers can ascertain the extent to which each model adjusts its replies to the required level of complexity (Seth et al., 2023).*

We use a free tool to determine the readability ratings (<https://goodcalculators.com/flesch-kincaid-calculator/>). We copied each ChatGPT and Bard's response to a question into the tool and noted the outcomes.

3.3.3. Cosine similarity

We may learn much about how ChatGPT and Bard think and handle problems by comparing their answers to questions on the CERH exam via cosine similarity. The cosine similarity measure is often employed in natural language processing to compare the similarity between two vectors (in this case, text representations). Researchers often use the well-known cosine and Jaccard similarity algorithms. Huang (2008) discovered that Jaccard provided better pure clustering for text clustering than did cosine. Jaccard is a famous tool, although it is a lexical tool that does not perform well for semantic analysis, according to Qurashi et al. (2020). They discovered that cosine similarity provided better semantic analysis than Jaccard similarity did. A high cosine similarity shows that both models produce conceptually and contextually comparable replies. This indicates that they are likely to use similar phrases, concepts, and ideas to comprehend the question and formulate solutions. A low cosine similarity, on the other

hand, suggests that ChatGPT and Bard treat the same problem in different ways. This might indicate differences in the interpretation, logic, or focus of the two models.

By computing and examining cosine similarity, one can learn more about how closely or differently ChatGPT and Bard respond to the same queries (Xia et al., 2015). This strategy deepens your comparative study and helps you comprehend the behavior of the models at a deeper level.

A publicly available tool was used to calculate the cosine similarity values (<https://tilores.io/cosine-similarity-online-tool>). The response for each question by ChatGPT and Bard was copied into the tool, and the resulting similarity value was noted.

4. Results

4.1. Descriptive analysis

On the basis of their responses to a dataset of 150 CSCP questions, ChatGPT and Bard, two general artificial intelligence (GAI) tools, are compared for accuracy in Table 2. Seventy-nine questions were correctly answered by Bard when ChatGPT also provided the correct response, whereas 14 questions were answered correctly by Bard but incorrectly by ChatGPT. On the other hand, Bard was wrong in 21 cases, ChatGPT was right, and both tools answered 36 questions incorrectly. The subtle distinctions between instances where one model was accurate while the other was not accurate and those where both models made the same errors add an intriguing dimension to the comparison. This identifies the advantages and disadvantages of each model and points to underlying variations in their construction or training set that might affect their performance in particular areas.

For ChatGPT and Bard, Table 3 shows the average score, standard error, and standard deviation for correctness, relevance, and clarity.

4.1.1. Average score

Bard (4.873) has a lower average comprehensive score than ChatGPT (4.947). Questions about the supply chain on ChatGPT tend to be more pertinent and leave out less information. Compared with Bard, ChatGPT has a slightly higher average relevance score (4.993). This suggests that, compared with Bard's comments, ChatGPT's responses are more concise and direct. Bard's remarks might have more superfluous details or be less relevant than ChatGPT's. Bard (4.5) has a lower average clarity score than ChatGPT does (4.793). This suggests that those with the necessary supply chain background knowledge find ChatGPT's comments easier to grasp. Bard falls short of ChatGPT in terms of accuracy, relevance, and clarity. It offers answers that address more parts of the issues, are simpler to learn, and go straight to the point with less superfluous information.

4.1.2. Standard error

The Chat GPT's accuracy's deficient standard error (0.018) denotes a great degree of accuracy in calculating the average accuracy score. Conversely, Bard has a more significant standard error (0.030), indicating that its average accuracy score is more variable. The Chat GPT's relevance standard error (0.007) is minimal, showing little variability in its average score estimate of 4.993. The average relevance score of

Table 2. Accuracy analysis of the ChatGPT and BARD responses.

	ChatGPT Correct	ChatGPT Incorrect
Bard Correct	79	14
Bard Incorrect	21	36

Table 3. Average score, standard error, and standard deviation of accuracy, relevance, and clarity.

	Average Score		Standard Error		Standard Deviation	
	Bard	ChatGPT	Bard	ChatGPT	Bard	ChatGPT
Accuracy	4.873	4.947	0.030	0.018	0.372	0.225
Relevance	4.953	4.993	0.017	0.007	0.212	0.082
Clarity	4.5	4.793	0.041	0.033	0.502	0.406

4.953 has more uncertainty according to Bard's standard error (0.017), which is somewhat larger. The Chat GPT's standard error for clarity (0.033) is low, similar to that for correctness and relevance, showing great accuracy in predicting its average clarity score of 4.793. Although significantly more extensive than Chat GPTs, Bard's standard error (0.041) indicates that the average clarity score of 4.5 may be subject to more significant variability. In conclusion, the standard error values provide information about the accuracy and dependability of the average scores for each criterion. Standard errors for Chat GPT are typically lower than those for Bard, indicating more accurate estimations of its typical performance in terms of accuracy, relevance, and clarity.

4.1.3. Standard deviation

To show how the scores for each criterion changed between ChatGPT and Bard, we present the numbers for the standard deviation. The accuracy standard deviation for ChatGPT is relatively low (0.225), indicating that the results are more reliable and more in line with the average score of 4.947. Although Bard's replies are less consistent and spread out around the mean score of 4.873, this is indicated by the group's more significant standard deviation (0.372). Similarly, Chat GPT's standard deviation (0.082) is relatively low, indicating that replies around the mean score of 4.993 were not particularly variable. Bard's standard deviation (0.212) is greater than its mean score of 4.953, indicating a wider spread and variability in relevance scores. The standard deviation for clarity on ChatGPT is also lower (0.406), indicating more consistent responses around the mean score of 4.793. Compared with its mean score of 4.5, Bard's standard deviation (0.502) is greater, indicating more significant variability in clarity scores. Compared with Bard, Chat GPT typically has fewer standard deviations, indicating higher accuracy, relevance, and clarity consistency. Owing to its replies' higher standard deviations, Bards may not have been as consistent in meeting these criteria.

4.1.4. *t* Test

We compare each AI model's average accuracy, relevance, and clarity ratings via *t* tests, as shown in Table 4. These tests assume that the differences between paired observations are normally distributed, which we verified via the Shapiro–Wilk test. The results confirmed that the assumption of normality was met.

For each criterion (Accuracy, Relevance, and Clarity), the offered *t* test values inform us of the statistical significance of the differences between the average scores of ChatGPT and Bard. We use a *t* test to determine whether there is a statistically significant difference between the means of the two groups. The accuracy of the *t* test result is 0.033808604 (<0.05). The difference between the average accuracy scores of ChatGPT and Bard is statistically significant, according to a simple *t* test. The relevance *t* test result is 0.033437127 (<0.05). The difference in significance between the two AI models is statistically significant, as shown by a small *t* test. The Clarity *t* test result is $1.31013\text{E}-08$ (>0.05). The average clarity scores of ChatGPT and Bard do not differ statistically in this instance, according to the rather significant *t* test value. The average ratings for accuracy and relevance differ significantly between ChatGPT and Bard. ChatGPT performed better than Bard in both of these areas. The average clarity scores of ChatGPT and Bard do not, however, differ significantly. In regard to providing brief responses to supply chain-related inquiries, both AI models perform similarly.

4.2. Confirmation queries impact

Table 5 displays the effects of confirmation inquiries on the 150 responses provided by the GAI tools Bard and ChatGPT. Initially, Bard offered 58 wrong responses, whereas ChatGPT offered 50 wrong responses. One of Bard's responses was incorrect, whereas eleven of the ChatGPTs were incorrect until the confirmation question 'Are you sure?' was included. However, even after the confirmation question,

Table 4. Average scores via a *t* test.

	Accuracy	Relevance	Clarity
<i>t</i> test ChatGPT – BARD	0.033808604	0.033437127	1.31013E-08

Table 5. Confirmation queries.

	Incorrect	Correct after Are you sure	Incorrect after Are you sure
Bard	58	1	57
ChatGPT	50	11	39

Table 6. Flesch–Kincaid Grade Level score and Flesch Readability Ease score of ChatGPT and Bard.

GenAI tool	Flesch–Kincaid Grade Level score	Flesch Readability Ease score	Estimated reading level
ChatGPT	11.2	34.6	College (difficult to read)
Bard	11.3	39.9	College (difficult to read)

39 of ChatGPT's answers were still wrong, whereas 57 of Bard's answers were wrong. This information suggests that the confirmation query increased the accuracy of both systems. However, the effectiveness differed, with Bard displaying a higher correction rate.

4.3. Cosine similarity analysis

After analyzing the dataset, we found that the 150 items under consideration fell into a distinct band of cosine similarity values. A similarity rating of 60–80 was found for most of the 115 questions, indicating moderate-to-significant alignment or similarity. Notably, 35 questions distinguished themselves with a similarity value of more than 80, indicating a very high match or almost identical nature between ChatGPT and Bard replies. The developers and organizations behind GenAI solutions will take different perspectives on potential supply chain issues. OpenAI has a history of iteratively improving models such as ChatGPT to balance utility and safety. The creators formulated Bards with stricter specifications to prevent potential misuse.

4.4. Readability analysis

We contrasted ChatGPT and Bard using the readability ratings shown in Table 6. The Flesch–Kincaid Grade Level gives ChatGPT's content a grade level of 11.2, meaning that readers who have finished approximately 11.2 years of formal education – roughly at the border between high school and college – are best suited for it. On the other hand, Bard's content aims at people who have completed approximately 11.3 years of education to make it more accessible. ChatGPT and Bard scored 34.6 and 39.9 on the Flesch Readability Ease scale, which ranges from 0 to 100 (higher values indicate easier readability). These results suggest that although both programs generate information that can be difficult (in the 'difficult' range), Bard's outputs are slightly more straightforward to read than ChatGPT's outputs are. Both AI systems produce content appropriate for college reading levels, with Bard having a minor accessibility advantage.

4.5. Comparative analysis

We compared the outcomes of ChatGPT with those of Bard in several scenarios. As shown in Table 7, the analysis revealed an unexpected trend where ChatGPT or Bard answered some questions. Upon analyzing the questions provided, it is evident that they revolve around the procurement strategy, warehousing strategies, reverse logistics, and specific technical aspects of the supply chain. The developer of ChatGPT, OpenAI, has a well-established procedure for continuously adjusting and fine-tuning its models to strike a compromise between use and security. On the other hand, Bards might adhere to stricter rules to prevent potential misuse, especially in fields plagued with complexity and sensitivity. This difference in methodology reflects broader difficulties and factors to be considered when creating generative AI systems that are accountable and informative.

5. Discussion

By evaluating the accuracy of the responses displayed in [Tables 8 to 13](#), it is possible to perform a comparison analysis of Bard's and ChatGPT's effectiveness in offering strategic supply chain management solutions. In our study, we consider both quantitative and qualitative data, including metrics such as the percentage of correct and incorrect answers and the thoroughness shown in the problem-solving process.

As seen in [Table 8](#), both Bard and ChatGPT provided the correct answer, 'C. Decoupling,' explaining that decoupling allows each process to operate at its maximum efficiency by using buffers to separate stages, preventing slower operations from constraining the overall process. However, Bard provided a more detailed example, while ChatGPT offered a concise explanation of the concept.

The results for a specific query about manufacturing techniques and their effects on process efficiency are illustrated and presented in [Table 9](#). [Table 9](#) answers a specific query concerning manufacturing techniques and emphasizes that decoupling is the best technique for ensuring that each process runs as efficiently as possible without being restricted by slower operations. Additionally, it compares decoupling to alternative solutions and explains each one. The concern raised in [Table 9](#) concerns procedures that enable every process in a production system to run as efficiently as possible without being slowed down

Table 7. Comparative analysis.

Scenarios	Table Number
Both ChatGPT and Bard gave correct answers.	Table 8
Both ChatGPT and Bard gave wrong answers but agreed on the wrong answer	Table 9
Both ChatGPT and Bard gave wrong answers but did not agree on the answer	Table 10
ChatGPT's answer is correct, and Bard's answer is wrong.	Table 11
ChatGPT's answer is wrong, and Bard's answer is correct	Table 12

Table 8. Both ChatGPT and Bard provided correct answers.

Q. No	Question	Bard Explanation	ChatGPT Explanation
[Q3]	Which of the following practices allows each process to operate at maximum efficiency rather than having the speed of the entire process constrained by the slowest operation? A. Theory of constraints (TOC) B. Supply/demand balancing C. Decoupling D. Pull production	The answer is C. Decoupling. Decoupling is a practice that allows each process to operate at maximum efficiency rather than having the speed of the entire process constrained by the slowest operation. Buffers are introduced between operations, allowing faster techniques to carry on while the slower ones catch up. For example, a manufacturing plant may have a bottleneck in the painting process. This means the painting process is the slowest, slowing down the entire production line. Decoupling can solve this problem by introducing a buffer between the painting and the following procedures. This buffer will allow faster methods to continue operating while the painting process catches up. Decoupling is a powerful tool for improving efficiency in a production process. It allows each process to operate at its maximum speed, which can lead to significant productivity gains. The other options are not as effective at allowing each process to operate at maximum efficiency. Theory of constraints (TOC) is a management philosophy that focuses on identifying and eliminating restrictions in a system. While TOC can effectively improve efficiency, decoupling is not as specific. Supply/demand balancing is a practice that aims to match supply with demand. This can help prevent bottlenecks but does not address the underlying issue of a slow process. Pull production is a lean manufacturing practice that uses kanban cards to control production. Pull production can effectively reduce waste but does not explicitly address the slow process issue. Push production is a traditional manufacturing practice of producing goods anticipating demand. Push production can lead to bottlenecks, as the production line cannot adapt to changes in demand. In conclusion, decoupling is the best practice for allowing each process to operate at maximum efficiency rather than having the speed of the entire process constrained by the slowest operation.	The practice that allows each process to operate at maximum efficiency by decoupling the processes so that the slowest operation does not constrain the speed of the entire process is: C. Decoupling Decoupling involves using inventory buffers to separate or 'decouple' different stages in a production process. This allows each stage to function at its optimal speed, rather than a subsequent or preceding stage slowing it down.

Table 9. Both ChatGPT and Bard gave wrong answers but agreed on the wrong answer.

Q. No	Question	Initial answer	Are you sure
Q [19] ChatGPT	Which of the following criteria is the primary factor when considering storage locations for high-volume products? A. Seasonality B. Quality C. Velocity D. Perishability	When considering storage locations for high-volume products, the primary factor is how quickly and frequently these products move in and out of storage. This is known as the 'velocity' of the product. High-velocity goods will be kept in areas that are simple to get to reduce handling time and improve warehouse operations. Therefore, the correct choice is: C. Velocity.	'Yes, I am sure. In warehousing and distribution, 'velocity' refers to the speed or frequency at which inventory moves through a facility or the supply chain. For high-volume products, which naturally have a high turnover rate, the primary consideration in determining storage locations would be their velocity. This ensures that the most frequently picked or moved items are stored in the most accessible areas, facilitating efficient warehouse operations. Therefore, for high-volume products, the most relevant criterion for storage location is valid: C. Velocity.
Q [19] Bard	Which of the following criteria is the primary factor when considering storage locations for high-volume products? A. Seasonality B. Quality C. Velocity D. Perishability	The correct answer is C. Velocity. Velocity is the rate at which products move through a supply chain. It is a measure of how quickly products are sold or consumed. For high-volume products, velocity is the primary factor when considering storage locations. This is due to the need for marketplaces to store high-volume goods nearby. This will help to reduce transportation costs and improve customer service. The other options are not as important as velocity when considering storage locations for high-volume products. Seasonality: Seasonality is the variation in demand for products over time. It is not as crucial for high-volume products as for low-volume products. This is because high-volume products are typically in high demand throughout the year. Quality: Quality is the degree to which a product meets its specifications. It is essential for all products, but it is not as crucial for high-volume products as it is for low-volume products. This is because high-volume products are typically less expensive and are not as critical to the customer experience. Perishability: Perishability is the degree to which a product can spoil or go bad. It is essential for some products, such as food and flowers, but it is not as crucial for high-volume products as for low-volume ones. This is because high-volume products are typically not perishable. In conclusion, velocity is the primary factor when considering storage locations for high-volume products. This will help to reduce transportation costs and improve customer service.	I am sure that velocity is a primary factor when considering storage locations for high-volume products. However, I think seasonality and perishability can also be essential factors. Seasonality is the variation in demand for products over time. For seasonal high-volume products, such as Christmas decorations or swimsuits, storing them in locations close to the markets during the peak season is essential. This will help reduce transportation costs and ensure that products are available when needed. Perishability is the degree to which a product can spoil or go bad. For high-volume products that are perishable, such as food or flowers, storing them in locations with the right temperature and humidity controls is essential. This will help extend the products' shelf life and ensure they are safe to eat or use. In conclusion, velocity, seasonality, and perishability are critical factors when choosing storage locations for high-volume products.

by the slowest process. Option C: Decoupling is the proper response to this query, according to both Bard and ChatGPT. Bard and ChatGPT explain why 'Decoupling' is the correct response. According to their definition, decoupling adds inventory buffers between various production process phases. These buffers aid in separating the stages so that a slower stage does not bind each stage and can work at its maximum speed. The rationales also explain why the other approaches are less successful in enabling each operation to run as efficiently as possible. For example, they noted that while supply/demand balancing and the theory of constraints can increase efficiency, they do not address the problem of slow processes such as decoupling. The effects of pull and push production on production efficiency are also discussed. According to the study findings, decoupling is the best method for ensuring that each process may run

as quickly as possible without being slowed down by another process. This underlines that decoupling is a potent technique for increasing production efficiency.

In Table 10, ChatGPT and Bard give different initial responses to the query about what should be considered when choosing a storage location for high-volume items. However, after some thought, they both had the same response. The initial response from ChatGPT is 'C. Velocity.' This explains that velocity – the speed and frequency with which products flow into and out of storage – is critical for high-volume products. Bard first chooses 'C. Velocity' as the correct response. It offers an explanation that is consistent with ChatGPT's opening statement. The correct response, however, is 'perishability.' ChatGPT reiterates its original decision, claiming that 'velocity' is the most crucial consideration when choosing a storage location for high-volume products when asked if it is sure of its response. This underlines the need for velocity for effective warehouse operations. Interestingly, Bard switches its mind in the 'Are You Sure' area after initially choosing 'C. Velocity.' While velocity is a critical component, 'seasonality' and 'perishability' are essential considerations for high-volume items. After more thought, ChatGPT and Bard concur that 'C. Velocity' is the incorrect primary criterion when choosing a storage location for high-volume items. Although Bard first added more detail on the significance of seasonality and perishability, it finally concurs with ChatGPT's assessment that velocity is the essential feature.

In Table 11, ChatGPT and Bard give different initial responses to the query about the activity most likely to increase overall supply chain risk. However, after some thought, they both had the same response. Initially, 'B. Expanding operations to multiple locations' is chosen by ChatGPT as the action that is most likely to increase overall supply chain risk. That is untrue. In Bard's opinion, the correct response is 'D. Reducing the supplier base for commodity-type components,' which is his initial response. The correct response, however, is C: 'Consolidating manufacturing locations.' When questioned about the certainty of its response, ChatGPT reconsiders the possibilities and modifies its position. It now chooses 'D. Reducing the supplier base for commodity-type components' as the action most likely to increase overall supply chain risk since it concurs with Bard's original response. When asked if it is inevitable, Bard reiterates its previous decision that the correct response is 'D. Reducing the supplier base for commodity-type components.' Bard explains in great detail why this approach raises supply chain risk. After considering it, ChatGPT and Bard concur that 'D. Reducing the supplier base for commodity-type components' is the move most likely to increase overall supply chain risk. They initially had varied replies, but after additional consideration, they agreed.

Table 12 shows a mismatch between ChatGPT and Bard's initial responses, and Bard initially provides the incorrect response. Even after being questioned if it is sure of its response, Bard still responds incorrectly. When the customer who supplies point-of-sale data remains the sole decision maker about order quantities, ChatGPT initially chooses 'D. Quick response' as the correct response for the type of inventory strategy. The reply from ChatGPT is accurate. The first response given by Bard, 'C. Vendor-managed replenishment (VMR),' is untrue. Bard incorrectly links VMR to the idea that the client will always be the only one making decisions. ChatGPT explains why 'D. Quick response' is the correct response in keeping with its original response. The quick response technique, which relies on point-of-sale data from the client yet gives them control over order quantity, is accurately described. Sadly, even being questioned if it is certain of its response, Bard still gives the wrong response. Bard incorrectly asserts that the answer is 'C. Vendor-managed replenishment (VMR).'

Table 13 contains a mismatch between ChatGPT's and Bard's initial responses, and ChatGPT initially provides an incorrect response. However, after more thought, ChatGPT modified its initial response. As the principal use of medium-term demand management projections, Bard initially chooses 'C. Aggregate demand for production planning' as the right response. Bard gave the right response. The first time it is used, ChatGPT chooses 'D. Develop the master production schedule' as the right response. According to Bard, this response is untrue. When asked if it is confident in its answers, Bard sticks with its original decision of 'C. Aggregate demand for production planning' and explains why. After the second thought, ChatGPT changed its response from 'D. Develop the master production schedule' to 'C. Aggregate demand for production planning.' It justifies the modification and matches its response with Bard's opening statement. After considering it, ChatGPT and Bard concur that the main use of medium-term demand management projections is 'C. Aggregate demand for production planning'.

Table 10. Both ChatGPT and Bard gave wrong answers but did not agree on the answer.

Q. No	Question	Initial Answer	Are you sure?
Q [46] ChatGPT	Which of the following actions will most likely increase total supply chain risk? A. Standardizing components used in a product family B. Expanding operations to multiple locations C. Consolidating manufacturing locations D. Reducing the supplier base for commodity-type components	Out of the options provided, the action most likely to increase total supply chain risk is: B. Expanding operations to multiple locations. Expanding operations to multiple locations introduces complexities such as coordinating among various sites, ensuring consistency in processes and products across different regions, managing logistics over larger geographic areas, and dealing with different regional risks (natural disasters, political instabilities, economic conditions, etc.). These complexities can introduce more points of potential failure or disruption in the supply chain, thus increasing overall risk. While other options can also increase risk (e.g., by increasing dependence on fewer suppliers), expanding operations to multiple locations presents a broader set of potential challenges and risks to manage.	Let us reevaluate the options: A. Standardizing components used in a product family: This reduces risk by simplifying inventory management, manufacturing processes, and procurement. Using standardized components across various products can often decrease the likelihood of stockouts, reduce manufacturing complexities, and increase suppliers' purchasing power. B. Expanding operations to multiple locations: This can introduce complexities such as coordinating among various sites and managing regional risks. However, it can also serve as a risk mitigation strategy since disruptions at one location may not necessarily affect operations at another. C. Consolidating manufacturing locations: Consolidating can increase risk because if there is a disruption at a single manufacturing location, it can halt the entire production. Having all operations centralized makes the supply chain vulnerable to location-specific risks. D. Reducing the supplier base for commodity-type components: This can significantly increase risk. Relying on fewer suppliers means that if one supplier has an issue (e.g., production halt, quality problems, or bankruptcy), it can significantly impact the company's supply chain. It reduces the flexibility and resilience of the supply chain against disruptions. Upon reconsideration, option D, 'Reducing the supplier base for commodity-type components,' is most likely to increase total supply chain risk, as it creates greater dependency on fewer suppliers, making the supply chain more vulnerable to any disruptions related to those specific suppliers.
Q [46] Bard	Which of the following actions will most likely increase total supply chain risk? A. Standardizing components used in a product family B. Expanding operations to multiple locations C. Consolidating manufacturing locations D. Reducing the supplier base for commodity-type components	The answer is D. Reducing the supplier base for commodity-type components. Supply chain risk is the likelihood of a disruption that could impact the availability, quality, or cost of goods or services. Reducing the supplier base for commodity-type components can increase total supply chain risk for several reasons. Increased concentration of risk: If a company reduces the number of suppliers for a particular component, it increases the concentration of risk. This is because the company is now more dependent on fewer suppliers. If one of these suppliers experiences a disruption, it could significantly impact the company's supply chain. Reduced flexibility: Reducing the supplier base can also reduce the company's flexibility. The company may have fewer options if it needs to find a new supplier for a particular component. Increased bargaining power of suppliers: Suppliers with fewer customers have more bargaining power. This means that they may be able to charge higher prices or demand better terms from the company. The other options are not as likely to increase total supply chain risk. Standardizing components used in a product family can reduce supply chain risk by reducing the number of components the company needs to manage. Expanding operations to multiple locations can increase supply chain risk by increasing the complexity of the supply chain. However, it can also reduce risk by providing the company with more options in case of a disruption. Consolidating manufacturing locations can reduce supply chain risk by reducing the number of locations the company needs to manage. Therefore, the action that is most likely to increase total supply chain risk is reducing the supplier base for commodity-type components.	Yes, I am sure. I have explained my reasoning in detail above. Reducing the supplier base for commodity-type components can increase total supply chain risk for several reasons. Increased concentration of risk: If a company reduces the number of suppliers for a particular component, it increases the concentration of risk. This is because the company is now more dependent on fewer suppliers. If one of these suppliers experiences a disruption, it could significantly impact the company's supply chain. Reduced flexibility: Reducing the supplier base can also reduce the company's flexibility. The company may have fewer options if it needs to find a new supplier for a particular component. Increased bargaining power of suppliers: Suppliers with fewer customers have more bargaining power. This means that they may be able to charge higher prices or demand better terms from the company. The other options are not as likely to increase total supply chain risk. Standardizing components used in a product family can reduce supply chain risk by reducing the number of components the company needs to manage. However, it can also increase risk if the standardized components are unavailable from a reliable supplier. Expanding operations to multiple locations can increase supply chain risk by increasing the complexity of the supply chain. However, it can also reduce risk by providing the company with more options in case of a disruption. Consolidating manufacturing locations can reduce supply chain risk by reducing the number of locations the company needs to manage. However, it can also increase risk if the consolidated location is not well-managed or experiences a disruption. Therefore, the action that is most likely to increase total supply chain risk is reducing the supplier base for commodity-type components.

Table 11. ChatGPT's answer is correct, and Bard's answer is wrong.

Q. No	Question	Initial answer	Are you sure?
Q [113] ChatGPT	The customer who provides point-of-sale data remains the sole decision-maker regarding order quantities; when what type of inventory strategy is used? A. Continuous replenishment B. Advanced continuous replenishment C. Vendor-managed replenishment D. Quick response	When using the Quick Response inventory strategy: D. Quick response The customer provides point-of-sale data to suppliers to enable them to react rapidly to customer demand. However, the customer remains the decision-maker regarding order quantities. This strategy aims to minimize lead times and ensure that products are available to customers when needed without holding excessive inventory. It is a collaboration between the retailer and supplier, but the retailer (customer) retains decision-making authority over order quantities.	
Q [113] Bard	The customer who provides point-of-sale data remains the sole decision-maker regarding order quantities; when what type of inventory strategy is used? A. Continuous replenishment B. Advanced continuous replenishment C. Vendor-managed replenishment D. Quick response	The answer is C. Vendor-managed replenishment (VMR). In VMR, the supplier manages the inventory levels of their products at a customer's location. The customer provides point-of-sale data to the supplier, but the supplier is the sole decision-maker regarding order quantities. In contrast, in continuous replenishment (CR), the customer manages their inventory levels. The customer uses point-of-sale data to determine when to order more inventory. Advanced continuous replenishment (ACR) is a more sophisticated version of CR. In ACR, the customer uses demand forecasting and inventory optimization software to determine when to order more. Quick response (QR) is a strategy that reduces the time between customer orders and product delivery. QR does not involve any specific inventory management strategy. Here are some additional details about VMR: Benefits: VMR can help customers to reduce inventory levels, improve customer service, and increase profitability. Drawbacks: VMR can be expensive to implement and manage. It can also be difficult to get suppliers to agree to VMR. Overall, VMR is an inventory strategy that can benefit customers who want to reduce inventory levels, improve customer service, and increase profitability.	I apologize for the previous answer. The answer to the question is A. Continuous replenishment. In continuous replenishment (CR), the customer manages their inventory levels. The customer uses point-of-sale data to determine when to order more inventory. The customer is the sole decision-maker regarding order quantities. In vendor-managed inventory (VMI), the supplier manages the inventory levels of their products at a customer's location. The supplier uses point-of-sale data to determine when to order more inventory. The supplier is the sole decision-maker regarding order quantities. Advanced continuous replenishment (ACR) is a more sophisticated version of CR. In ACR, the customer uses demand forecasting and inventory optimization software to determine when to order more. However, the customer is still the sole decision-maker regarding order quantities. Quick response (QR) is a strategy that reduces the time between customer orders and product delivery. QR does not involve any specific inventory management strategy. Here are some additional details about CR: Benefits: CR can help customers to reduce inventory levels, improve customer service, and increase profitability. Drawbacks: CR can be more difficult to manage than other inventory management strategies. It can also lead to stockouts if the customer does not have accurate point-of-sale data. Overall, CR is an inventory strategy that can benefit customers who want to reduce inventory levels, improve customer service, and increase profitability. I hope this helps! Let me know if you have other questions or requests.

Initially, there was a disparity. However, the response of ChatGPT changed in light of additional analysis.

A comparison of the justifications offered by ChatGPT and Bard reveals that Bard provides a more thorough and in-depth explanation of the factors influencing the choice or rejection of each alternative (Table 13). With this approach, more information is considered before making decisions, leading to a more thorough analysis of each problem. Strong statistical support and a compelling justification were

Table 12. ChatGPT's answer is wrong, and Bard's answer is correct.

Q. No	Question	Initial answers	Are you sure?
Q [80] Bard	<p>Medium-term demand management projections are used primarily to:</p> <p>A. Complete strategic business planning of facilities.</p> <p>B. Complete forecasts at the item level.</p> <p>C. aggregate demand for production planning.</p> <p>D. Develop the master production schedule.</p>	<p>The answer is C. aggregate demand for production planning.</p> <p>Medium-term demand management projections are used to aggregate demand for production planning. This is because medium-term demand management projections cover a period that is too long for item-level forecasts to be accurate. By aggregating demand for production planning, companies can ensure they have the proper inventory to meet demand.</p> <p>Option A, complete strategic business planning of facilities, is not the primary purpose of medium-term demand management projections. Strategic business planning is a long-term process not typically based on medium-term demand management projections.</p> <p>Option B, complete forecasts at the item level, is not the primary purpose of medium-term demand management projections. Item-level forecasts are typically used for short-term planning, such as inventory management and order fulfillment.</p> <p>Option D, developing the master production schedule, is not the primary purpose of medium-term demand management projections. The master production schedule is a detailed plan that specifies the quantity and timing of production for each product. It is typically based on short-term demand forecasts.</p> <p>In conclusion, medium-term demand management projections primarily aggregate demand for production planning. This is because medium-term demand management projections cover a period that is too long for item-level forecasts to be accurate. By aggregating demand for production planning, companies can ensure they have the proper inventory to meet demand.</p>	
Q [80] ChatGPT	<p>Medium-term demand management projections are used primarily to:</p> <p>A. Complete strategic business planning of facilities.</p> <p>B. Complete forecasts at the item level.</p> <p>C. aggregate demand for production planning.</p> <p>D. Develop the master production schedule.</p>	<p>Medium-term demand management projections typically span from a few months up to a year or two. They are not as long-term as strategic business plans, which might consider multiple years or decades, and they are not as short-term as immediate production schedules or detailed item-level forecasts.</p> <p>Given the options provided, the most suitable use of medium-term demand management projections is:</p> <p>D. Develop the master production schedule.</p> <p>The master production schedule (MPS) provides a plan for producing individual end items or products. It is more detailed than long-term strategic plans but not as granular as short-term daily or weekly schedules. The MPS typically falls within the medium-term range and uses medium-term demand forecasts to balance and align production plans with anticipated demand.</p> <p>Therefore, the correct answer is D. Develop the master production schedule.</p>	<p>Let us reevaluate the options:</p> <p>A. Complete strategic business planning of facilities typically pertains to long-term projections. Strategic business planning would involve decisions like building new facilities, closing old ones, or making significant capital investments. Such decisions usually require a much longer horizon than the medium-term.</p> <p>B. Complete forecasts at the item level can be short-term or medium-term. Nevertheless, item-level forecasts are usually more detailed and short-term to drive immediate production or purchasing decisions.</p> <p>C. Aggregate demand for production planning: This is a potential use for medium-term demand management projections. Aggregating demand over a medium-term horizon provides a macrolevel view to help balance resources and capacities and make decisions about broader production strategies.</p> <p>D. Develop the master production schedule (MPS): The MPS translates the strategic business plan into a tactical production plan. It typically covers the medium-term horizon, often in monthly buckets, to plan production for specific items in response to forecasted demand.</p> <p>Upon reconsideration, both C and D are relevant for medium-term demand management projections. However, when distinguishing between the aggregate level of production planning and the more specific MPS, the master production schedule might lean more toward the shorter side of the medium term.</p> <p>Given the emphasis on aggregation in option C, which fits the medium-term scope, 'aggregate demand for production planning' could be a more fitting answer than the MPS.</p> <p>Therefore, revising my earlier response, the correct answer is likely C. aggregate demand for production planning.</p>

Table 13. A comparison of justifications offered by ChatGPT and Bard.

Aspect	Bard Response	ChatGPT Response
Supply Chain Expertise	Bard provides precise and thorough explanations, demonstrating a solid grasp of supply chain management concepts.	While ChatGPT demonstrates a fair understanding of supply chain concepts, it occasionally gives less thorough or precise responses.
Initial Accuracy	Bard frequently gives the right responses and justifications right away, demonstrating consistent expertise in the supply chain industry.	Initial responses from ChatGPT could be inaccurate or mistaken, necessitating more analysis and correction.
Responsiveness to question	Bard typically gives concise, direct comments that provide understandable and pertinent information.	The lengthy nature of ChatGPT's comments may make it more difficult to get the precise details needed to answer the query.
Flexibility in approaches	Bard frequently takes a systematic and thorough approach to supply chain scenarios, considering several elements before providing solutions.	The method taken by ChatGPT is more adaptable and may provide quick, practical answers without carefully considering every aspect.
Handling of complex scenarios	With its comprehensive and strategic solutions, Bard excels at managing challenging supply chain situations.	When faced with complex or multifaceted problems, ChatGPT may struggle and offer a more straightforward or less comprehensive solution.
Depth of knowledge	Bard frequently demonstrates extensive topic expertise in his comments, offering subtle insights into supply chain management fundamentals.	The breadth of ChatGPT's expertise can vary, and it might not always offer in-depth justifications or insights.
Verifiability of information	Bard's solutions are typically well-researched and based on accepted supply chain ideas and procedures.	Critical judgments may need to be confirmed externally if a ChatGPT response lacks particular citations or references.

provided for ChatGPT's top picks. However, the precise standards for excluding particular comments were not immediately clear.

ChatGPT and Bard strengths and weaknesses

ChatGPT is adaptable in its responses and may offer solutions for various questions and domains. It effectively adjusts to various questions and situations. ChatGPT handles queries rapidly and produces responses. It is suitable for applications that demand quick information retrieval because it is responsive and accessible around the clock. ChatGPT is a useful tool for jobs involving data analysis and information synthesis since it is excellent at processing and summarizing enormous amounts of information. ChatGPT is adaptable for various use cases and applications since it can handle several queries and tasks. Initial ChatGPT responses could be inaccurate or mistaken, and those responses frequently need to be verified and corrected. ChatGPT may have trouble comprehending context and providing nuanced responses, occasionally resulting in misunderstandings. Ethical questions are raised by ChatGPT's potential to reinforce biases unintentionally found in its training data. Understanding how ChatGPT arrives at a specific answer may be difficult because of the possible lack of openness and interpretability of ChatGPT responses.

Bard exhibits deep knowledge of the specialized field in which it is trained. In its specialist field, Bard consistently and accurately provides information in its responses. Typically, Bard's succinct responses address their inquiries immediately and offer pertinent and detailed information. Bards continually emphasize moral issues, particularly in circumstances requiring moral quandaries, encouraging ethical and responsible decisions. Bards' responses might not be as adaptable or diverse as those of ChatGPTs, and they might not be able to handle inquiries or situations outside of their purview. Despite being the best in its field, Bard's responses might not always provide logical justifications or explanations, which makes its decision-making procedures less transparent.

Upon closer examination, one unexpected finding was Bard's slight edge in accessibility, as indicated by its higher Flesch readability ease score, despite ChatGPT outperforming Bard in accuracy and relevance. This suggests that while Bard's responses may be easier to understand, they may not always be as precise or contextually relevant as ChatGPT's. ChatGPT's superior performance in terms of accuracy and relevance likely stems from its more advanced training on a broader dataset and its iterative improvement model, which prioritizes utility and safety. Additionally, we observed that ChatGPT was better at handling complex, multipart questions, potentially because of its architecture's stronger focus

on problem-solving and decision-making scenarios. These factors contribute to the model's overall advantage in professional applications such as supply chain management.

Although thorough, the study uses 150 multiple-choice questions (MCQs) from the APICS CSCP exam, which focuses on a particular area of supply chain management. The dataset may not fully capture the real-world practical, hands-on experience that professionals may encounter because it largely focuses on the theoretical parts of supply chain management. There might be a bias in this dataset toward queries with a more academic focus as opposed to scenarios involving real-world problem solving. As a result, the study may not adequately represent how well AI models perform in real-world settings where situational awareness and on-the-ground decision-making are crucial.

According to the study, ChatGPT and Bard can both be quite helpful as study tools, especially for students who are preparing for tests such as the APICS CSCP. They can be used to present different viewpoints on supply chain management principles, explain things, and mimic test situations. By encouraging students to consider other perspectives and emphasizing the need to critically evaluate and cross-reference AI-generated content with traditional textbooks and professional guidelines, educators can use these AI technologies to improve learning outcomes for their students. Both approaches can provide more educational materials because of ChatGPT's flexibility and speed as well as Bard's ability to provide clear, concise explanations. To achieve particular learning objectives, this entails developing comprehensive lesson plans, practice questions, and instructive examples. By using these models, educators may quickly create personalized content that targets particular areas in which students struggle and offers more specialized resources and help. AI models can be utilized for continuing education and professional development for supply chain management experts. They can offer succinct responses and useful ideas that might support brainstorming sessions, particularly in situations where a variety of viewpoints might be advantageous. Businesses can include AI tools in their training programs to provide workers with up-to-date knowledge about industry norms and practices. This is a dynamic approach to learning about new tactics and technology.

6. Conclusions and implications

Deep learning uses generative AI tools such as ChatGPT and Bard to process and comprehend human language. Owing to their thorough training in textual data, these models can understand grammar and context and produce language that resembles human responses. GAI tools are essential for improving human-computer interactions for chatbots, sentiment analysis, and content creation. They can consistently address new obstacles owing to their aptitude for learning and flexibility. While Bard by Google focuses on more dynamic, human-like discussions, ChatGPT, developed by OpenAI, is intended for interactive text-based conversations. Both methods have been used across various industries, from content generation to niche markets. Research is still being conducted to determine their advantages and disadvantages in various situations. The performance of two well-known large language models, ChatGPT and Bard, in answering supply chain management-related queries was thoroughly compared in this study. A complex understanding of their capabilities was developed by thoroughly examining their accuracy, relevance, and clarity and the effect of confirmation inquiries. The results showed that both models have great potential for understanding and reacting to supply chain management topics.

The results of this investigation have numerous consequences. First, the analysis revealed that ChatGPT performed better than Bard did in terms of accuracy and relevance, with statistically significant differences. Furthermore, the effectiveness of confirmation questions in boosting response accuracy points to a viable direction for increasing the dependability of AI-generated answers via iterative feedback mechanisms.

Examining cosine similarity values also highlighted that both models showed various degrees of alignment in their responses, suggesting the possibility of agreement on particular supply chain issues. However, the uniqueness of some queries that one model successfully answers but not the other highlights the importance of diversity in AI-driven solutions for intricate and multidimensional areas such as supply chain management. With both models targeting readers with college-level reading ability, the readability analysis highlighted the accessibility of created content. Both systems generated writing that was usually in line with higher educational levels, showing their application in academic and professional

situations, albeit Bard showed a modest readability advantage. The comparative examination of question coverage indicated that the tactics used by developers to adjust and enforce rules substantially impact the knowledge bases and sensibility of these AI systems. Bard's stricter requirements, which prevent possible misuse, contrast with OpenAI's iterative process and commitment to balancing utility and safety.

The bullwhip effect practices to increase process efficiency, the synchronization of supply and demand activities, the impact of globalization on supply chains, involving suppliers in design, measuring reliability, the SCOR model, the goal of GRI, undesirable attributes, revenue-focused CRM activities, quality-based supplier selection, and other concepts were all accurately addressed by ChatGPT and Bard. Additionally, they were aware of issues such as customs clearance for international sourcing, reducing the bullwhip effect, forecasting methods, reducing cash-to-cash cycle time, and the role of technologies such as bar codes and RFID. Both organizations, however, needed help with some ideas. The importance of price differentiation, the value of interpersonal communication and negotiation with suppliers, the significance of manufacturing facilities' location for cost competitiveness in exports, handling a key supplier's violation of the UN Global Compact, and balancing costs in lot size decisions were among them. They also included assessing the delivery performance of forecast-built items. Additional explanations for these ideas could be needed to resolve potential misunderstandings or comprehension gaps.

The study offers a nuanced understanding of the potential and limitations of GAI tools in the realm of supply chain management. These systems not only enhance decision-making and operational efficiency through advanced data analysis and text processing but also offer capabilities for improved stakeholder communication and customer engagement. Despite these advancements, the study calls attention to the imperative of responsible AI development, emphasizing the need for ongoing model adaptation to ensure ethical and safe application, particularly as these technologies become integral to complex domains such as supply chain management.

This study has certain limitations. The analysis is based on a dataset with only 150 questions. As a result, the size of the dataset may constrain the applicability of the findings. It is also critical to consider the possibility of bias introduced by raters and coders when the Likert scale is used, which could impact the accuracy of the evaluation.

A larger dataset comprising a wider variety of queries and scenarios may offer a better understanding of the models' capabilities. Open-ended questions, simulations, and real-world case studies could be used to gauge how well an AI can solve challenging problems and provide original answers. A more thorough assessment of AI tools in diverse settings would be made possible by this extension, offering insights into the tools' suitability for use in a range of sectors and businesses beyond supply chain management.

Examining how iterative feedback loops affect the models' learning and accuracy of responses may provide insightful information. This study investigated how model performance changes over time when models are continuously trained with real-time user feedback. Comprehending the function of feedback in fine-tuning artificial intelligence responses might result in the creation of AI tools that are more flexible and responsive, thus increasing their usefulness in ever-changing and dynamic domains.

By examining how various training datasets affect model performance, biases and knowledge gaps may be found. Future research might assess how data diversity influences accuracy and relevance by comparing models trained on a variety of datasets, such as non-English sources and domain-specific content. The results of this study may direct the creation of more inclusive and thorough AI models that can comprehend and address a greater variety of viewpoints and situations.

Author contributions

Conceptualization, R.R.; Methodology, R.R.; Software, A.S.; Validation, A.S.; Formal Analysis, R.R.; Data Curation, R.R.; Writing – Original Draft Preparation, R.R., P.N., and M.S.; Writing – Review & Editing, R.R., P.N., A.S., M.S., A.G.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT for English editing and grammar checks. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

About the authors

DR. Raghu Raman, Ph.D., is the Dean of the Amrita School of Business at Amrita Vishwa Vidyapeetham. He holds a Ph.D. degree in Management from Amrita Vishwa Vidyapeetham, India, and an MBA from Haas School of Business, UC Berkeley, USA. His main research interests are Adaptive & Virtual Learning Environments aligned with UN SDGs, Diffusion of ICT Innovations in Socio-technical Systems.

Dr. Aswathy Sreenivasan holds a PhD in management and is currently at Kristu Jayanti College, Bengaluru, India. She focusses on operational excellence in startups using R and Colab for data analysis, data-driven decision-making and continuous process optimization.

Dr. Suresh M holds a PhD in management and is a full professor at Amrita School of Business. His research interests are Social Entrepreneurship; Rural Entrepreneurship; Green Startups; Industry 5.0; Healthcare 5.0; Construction 5.0.

Dr. Angappa Gunasekaran is a Professor of Supply Chain Management at the School of Business Administration at Pennsylvania State University, Harrisburg. He is currently interested in researching benchmarking, agile manufacturing, performance measures and metrics in new economy, logistics, and supply chain management.

Dr. Prema Nedungadi is the Associate of the Dean of Computing at Amritapuri, Amrita Vishwa Vidyapeetham with a PhD in Computer Science. Her research interest is in interdisciplinary projects that combine AI, Reading Difficulties, Human-Computer Interfaces, and Computational Linguistics to build scalable systems.

Data availability statement

The entire dataset used for analysis is available as supplementary material of the manuscript.

References

- Abd-Alrazaq, A., AlSaad, R., Alhuwail, D., Ahmed, A., Healy, P. M., Latifi, S., Aziz, S., Damseh, R., Alabed Alrazak, S., & Sheikh, J. (2023). Large language models in medical education: Opportunities, challenges, and future directions. *JMIR Medical Education*, 9(1), e48291. <https://doi.org/10.2196/48291>
- Agarwal, M., Sharma, P., & Goswami, A. (2023). Analyzing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. *Cureus*, 15(6), e40977. <https://doi.org/10.7759/cureus.40977>
- Ahmed, A. K., Senthilkumar, C. B., & Nallusamy, S. (2018). Study on environmental impact through analysis of big data for sustainable and green supply chain management. *International Journal of Mechanical and Production Engineering Research and Development*, 8(1), 1245–1254.
- Ali, R., Tang, O. Y., Connolly, I. D., Fridley, J. S., Shin, J. H., Sullivan, P. L. Z., ... Asaad, W. F. (2023). Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *Neurosurgery*, 93(5), 1090–1098. <https://doi.org/10.1101/2023.04.06.23288265>
- Ansari, Z. N., & Kant, R. (2017). A state-of-art literature review reflecting 15 years of focus on sustainable supply chain management. *Journal of Cleaner Production*, 142, 2524–2543. <https://doi.org/10.1016/j.jclepro.2016.11.023>
- Bonner, E., Lege, R., & Frazier, E. (2023). Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching. *Teaching English With Technology*, 23(1), 23–41. <https://doi.org/10.56297/BKAM1691/WIEO1749>
- Caramancion, K. M. (2023). News verifiers showdown: A comparative performance evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in news fact-checking. *arXiv preprint arXiv:2306.17176*. <https://doi.org/10.48550/arxiv.2306.17176>
- Centobelli, P., Cerchione, R., Del Vecchio, P., Oropallo, E., & Secondo, G. (2022). Blockchain technology for bridging trust, traceability and transparency in circular supply chain. *Information & Management*, 59(7), 103508. <https://doi.org/10.1016/j.im.2021.103508>
- Chou, D. C., Tan, X., & Yen, D. C. (2004). Web technology and supply chain management. *Information Management & Computer Security*, 12(4), 338–349. <https://doi.org/10.1108/09685220410553550>

- De Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3), 515–535. <https://doi.org/10.1353/lan.2006.0130>
- Doshi, R., Amin, K., Khosla, P., Bajaj, S., Chheang, S., & Forman, H. P. (2023). Utilizing large language models to simplify radiology reports: A comparative analysis of ChatGPT3. 5, ChatGPT4. 0, Google Bard, and Microsoft Bin. *medRxiv*, 2023. 06. <https://doi.org/10.1101/2023.06.04.23290786>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). Opinion Paper: 'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642. <https://doi.org/10.1016/j.ijinfo-mgt.2023.102642>
- Eleyan, D., Othman, A., & Eleyan, A. (2020). Enhancing software comments readability using flesch reading ease score. *Information*, 11(9), 430. <https://doi.org/10.3390/info11090430>
- Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2023). Generative AI. Available at SSRN: <https://doi.org/10.2139/ssrn.4443189>
- Floridi, L. (2023). AI as an agency without intelligence: on ChatGPT, large language models, and other generative models. *Philosophy & Technology*, 36(1), 15. <https://doi.org/10.1007/s13347-023-00621-y>
- Gunasekaran, A., & Ngai, E. W. (2004). Virtual supply-chain management. *Production Planning & Control*, 15(6), 584–595. <https://doi.org/10.1080/09537280412331283955>
- Haver, H. L., Lin, C. T., Sirajuddin, A., Yi, P. H., & Jeudy, J. (2023). Use of ChatGPT, GPT-4, and Bard to improve readability of ChatGPT's answers to common questions on lung cancer and lung cancer screening. *American Journal of Roentgenology*, 221(5), 701–704. <https://doi.org/10.2214/AJR.23.29622>
- Holmes, J., Liu, Z., Zhang, L., Ding, Y., Sio, T. T., McGee, L. A., ... Liu, W. (2023). Evaluating large language models on a highly specialized topic, radiation oncology physics. *arXiv preprint arXiv:2304.01938* <https://doi.org/10.48550/arxiv.2304.01938>
- Huang, A. (2008). *Similarity measures for text document clustering* [Paper presentation]. Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand. April). In (Vol. 4, pp. 9–56).
- Javaid, M., Haleem, A., & Singh, R. P. (2023). ChatGPT for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(1), 100105. <https://doi.org/10.1016/j.tbench.2023.100105>
- Khademi, A. (2023). Can ChatGPT and Bard generate aligned assessment items? A reliability analysis against human performance. *arXiv preprint arXiv:2304.05372* <https://doi.org/10.48550/arxiv.2304.05372>
- Khan, W., Daud, A., Khan, K., Muhammad, S., & Haq, R. (2023). Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends. *Natural Language Processing Journal*, 4, 100026. 100026. <https://doi.org/10.1016/j.nlp.2023.100026>
- Luitse, D., & Denkena, W. (2021). The great transformer: Examining the role of large language models in the political economy of AI. *Big Data & Society*, 8(2), 205395172110477. <https://doi.org/10.1177/20539517211047734>
- Mohamadi, S., Mujtaba, G., Le, N., Doretto, G., & Adjeroh, D. A. (2023). ChatGPT in the age of generative AI and large language models: A concise survey. *arXiv preprint arXiv:2307.04251*. <https://doi.org/10.48550/arxiv.2307.04251>
- Noda, R., Izaki, Y., Kitano, F., Komatsu, J., Ichikawa, D., & Shibagaki, Y. (2023). Performance of ChatGPT and Bard in self-assessment questions for nephrology board renewal. *medRxiv*, 2023-06. <https://doi.org/10.1101/2023.06.06.23291070>
- Patnaik, S. S., & Hoffmann, U. (2023). Comparison of ChatGPT vs. Bard to anesthesia-related queries. *medRxiv*, 2023-06. <https://doi.org/10.1101/2023.06.29.23292057>
- Qurashi, A. W., Holmes, V., & Johnson, A. P. (2020, August). *Document processing: Methods for semantic text similarity analysis* [Paper presentation]. 2020 International Conference on INnovations in Intelligent Systems and Applications (INISTA), In (pp. 1–6). IEEE. <https://doi.org/10.1109/INISTA49547.2020.9194665>
- Rahsepar, A. A., Tavakoli, N., Kim, G. H. J., Hassani, C., Abtin, F., & Bedayat, A. (2023). How AI responds to common lung cancer questions: ChatGPT vs Google Bard. *Radiology*, 307(5), e230922. <https://doi.org/10.1148/radiol.230922>
- Raman, R., Calyam, P., & Achuthan, K. (2024). ChatGPT or Bard: Who is a better Certified Ethical Hacker? *Computers & Security*, 140, 103804. <https://doi.org/10.1016/j.cose.2024.103804>
- Raman, R., Lathabai, H. H., Mandal, S., Das, P., Kaur, T., & Nedungadi, P. (2024). ChatGPT: Literate or intelligent about UN sustainable development goals? *PloS One*, 19(4), e0297521. <https://doi.org/10.1371/journal.pone.0297521>
- Raman, R., Mandal, S., Das, P., Kaur, T., Sanjanasri, J. P., & Nedungadi, P. (2024). Exploring university students' adoption of ChatGPT using the diffusion of innovation theory and sentiment analysis with gender dimension. *Human Behavior and Emerging Technologies*, 2024(1), 3085910. <https://doi.org/10.1155/2024/3085910>
- Raman, R., Nair, V., Dinesh, S., & Acharyulu, R. (2024). Comparative analysis of ChatGPT and Bard in digital governance: Accuracy, adaptability, and readability insights. *Digital Government: Research and Practice*. <https://doi.org/10.1145/3674846>
- Raman, R., Venugopalan, M., & Kamal, A. (2024). Evaluating human resources management literacy: A performance analysis of ChatGPT and bard. *Heliyon*, 10(5), e27026. <https://doi.org/10.1016/j.heliyon.2024.e27026>

- Roscoe, R. D., Balyan, R., McNamara, D. S., Banawan, M., & Schillinger, D. (2023). Automated strategy feedback can improve the readability of physicians' electronic communications to simulated patients. *International Journal of Human-Computer Studies*, 176, 103059. <https://doi.org/10.1016/j.ijhcs.2023.103059>
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225–237. <https://doi.org/10.3758/PBR.16.2.225>
- Seth, I., Bulloch, G., & Rozen, W. M. (2023). Applications of artificial intelligence and large language models to plastic surgery research. *Aesthetic Surgery Journal*, 43(10), NP809–NP810. <https://doi.org/10.1093/asj/sjad210>
- Shen, Y., Heacock, L., Elias, J., Hentel, K. D., Reig, B., Shih, G., & Moy, L. (2023). ChatGPT and other large language models are double-edged swords. *Radiology*, 307(2), e230163. <https://doi.org/10.1148/radiol.230163>
- Williamson, J. M. L., & Martin, A. G. (2010). Analysis of patient information leaflets provided by a district general hospital by the Flesch and Flesch–Kincaid method. *International Journal of Clinical Practice*, 64(13), 1824–1831. <https://doi.org/10.1111/j.1742-1241.2010.02408.x>
- Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information Sciences*, 307, 39–52. <https://doi.org/10.1016/j.ins.2015.02.024>
- Xu, F., Lin, Q., Han, J., Zhao, T., Liu, J., & Cambria, E. (2023). Are large language models truly good logical reasoners? A comprehensive evaluation from deductive, inductive and abductive views. *arXiv preprint arXiv:2306.09841* <https://doi.org/10.48550/arxiv.2306.09841>
- Zhao, B., Jin, W., Del Ser, J., & Yang, G. (2023). ChatAgri: Exploring potentials of ChatGPT on cross-linguistic agricultural text classification. *arXiv preprint arXiv:2305.15024*. <https://doi.org/10.48550/arXiv.2305.15024>