

# LLMs-Based Weighting for MCDM: Case Study in Logistics Performance Index

Nurcan DENİZ

Eskisehir Osmangazi University, Department of Operations Management and Marketing, Eskisehir, Türkiye

[ndeniz@ogu.edu.tr](mailto:ndeniz@ogu.edu.tr), [nurcanatikdeniz@gmail.com](mailto:nurcanatikdeniz@gmail.com)

Received on: 02 June 2025

Accepted on: 06 January 2026

Published on: 07 April 2026

## Abstract

**Purpose:** The aim of this paper is to conduct an experimental study to determine whether large language models (LLMs)-based weighting is suitable for Multi-Criteria Decision Making (MCDM) or not. In addition, analyzing the prompt design effect is the other purpose of this study.

**Design/methodology/approach:** In this study, experiments were carried out based on prompts to determine the ranking and weights of six criteria (customs, infrastructure, international shipments, logistics competence and quality, timeliness, tracking and Tracing) in the Logistics Performance Index (LPI).

**Findings/results:** The findings revealed that prompt design is the most critical part to integrate LLMs in the MCDM context. Results show that there is not a single compatible LPI ranking, and the answers of ChatGPT differ with different prompt types (zero-shot, few-shot, and Chain of thought) and decision makers (supply chain director and academician).

**Practical implications:** The findings of this study are beneficial for researchers in the MCDM area who want to integrate LLMs in the field. The study will help not only the professionals in the logistics sector but also the beginners.

**Originality/value:** There are some recent studies using LLMs in the MCDM context. To the best of our knowledge, it is the first study used in logistics performance evaluation.

**Keywords:** LLMs, prompt, ChatGPT, LLMs-based weighting, MCDM, logistics, Logistics Performance Index, generative artificial intelligence

## Introduction

Multi-Criteria Decision Making (MCDM) techniques are beneficial in the presence of conflicting criteria to rank the alternatives, to select the best alternative, and to classify the alternatives. There are numerous MCDM techniques. Some of them are used to determine the weights of each criterion, and some of them make it possible to rank the alternatives based on their own algorithms.

Developments in Artificial Intelligence (AI) have transformed research areas. MCDM research area was also influenced by the development in Generative Artificial Intelligence (GenAI). Literature shows the early attempts of pre-trained large language models (LLMs) usage in the MCDM area. It was reported that LLMs were used in domains such as question answering, professional consulting, and code generation in this context (Wang and Wu, 2024). Frederico (2023) also reported that Chat Generative Pre-Trained Transformer (ChatGPT) can help and support the supply chain's decision-makers. ChatGPT is an AI chatbot based on advanced LLMs, and it was developed by OpenAI (Park et al. 2025).

Some of the researchers (Wang and Wu, 2024; Dehghanimohammadabadi and Kabadayi, 2025; Lu et al., 2025; Park et al., 2025; Svoboda and Lande, 2025; Wang et al., 2025) tried to understand whether ChatGPT (as a virtual expert) can replace MCDM experts and compared the results of real and virtual experts.

There are mainly two kinds of weighing in MCDM: subjective and objective. There are some studies in the literature in which ChatGPT was used as a virtual expert in subjective MCDM techniques like Analytic Hierarchy Process (AHP) (Svoboda and Lande, 2025; Dehghanimohammadabadi and Kabadayi, 2025; Lu et al., 2025; Park et al., 2025; Wang et al., 2025). Domain expertise is crucial in AHP, not only in making pairwise comparisons but also in defining hierarchical structure and maintaining logical consistency (Park et al. 2025).

These studies were carried out in diverse areas such as supplier evaluations (Wang and Wu, 2024; Dehghanimohammadabadi and Kabadayi, 2025), supply chain evaluation (Wang et al., 2025), air quality evaluation (Wang et al., 2025), and customer satisfaction evaluation (Wang et al., 2025). Increase in speed, quality, accuracy, and objectivity, efficiency and reliability, decrease in time, cost, and bias (Wang and Wu, 2024; Svoboda and Lande, 2025) are the expected results of LLMs usage in MCDM techniques.

However, the LLMs-based MCDM research area is still immature. According to "diffusion of innovation theory,"

Rogers (1983) identified five adopter categories: Innovators (2.5% of adopters), early adopters (13.5%), the early majority (34%), the late majority (34%), and laggards (16%) (Wamba et al. 2024). It is highly important to carry out studies to make contributions to the literature.

These studies used LLMs as a way of implementing an MCDM technique instead of using software / MS Excel spreadsheets or replacing human experts with virtual experts to make pairwise comparisons. To the best of our knowledge, there is no study that uses LLMs as a new MCDM technique for weighting and ranking. In this study, there is an attempt to see the capabilities of ChatGPT as a virtual MCDM technique to see whether it can produce consistent weights or not. Three research questions (RQs) are listed below:

RQ1: Is LLM-based weighting suitable for MCDM?

RQ2: Which type of prompt (zero-shot, few-shot, Chain of Thought (CoT)) is better to get the result?

RQ3: Are the weights consistent?

Answering these research questions makes it possible to contribute to both the MCDM, performance management, and prompt engineering areas. First of all, this study will be one of the pioneer studies in the LLMs-based MCDM area, and the results of the study will contribute to the weighting techniques in the MCDM area. In addition, this study also contributes to the logistics performance management area in which MCDM techniques are frequently used, as explained in the literature review part in detail. Lastly, prompt design is a crucial part of communication with AI, and the results of this study will contribute to this area by proposing prompt templates. In a nutshell, the contributions of this research can be listed as follows:

- LLMs' usage in the field of weight determination in the MCDM context
- LLMs usage in the field of logistics performance management
- The effect of different prompt strategies

The paper is organised as follows. The following section provides a literature review about MCDM usage in performance measurement with LPI. The next section consists of methodology based on prompt design. The study continues with the results and the discussion. The conclusion part is the last section, which includes limitations and recommendations for future studies.

## Literature Review

LPI is the acronym of “Logistics Performance Index,” which was presented by the World Bank biannually since 2007 to measure the performance of logistics activities based on a worldwide survey of logistics professionals. The seventh edition compares 139 countries based on defined criteria (The World Bank Report, 2023). To calculate overall performance scores, six criteria are used, and these criteria are equally weighted. These six criteria are listed as follows:

- Customs (C1)
- Infrastructure (C2)
- International shipments (C3)
- Logistics competence and quality (C4)
- Timeliness (C5)
- Tracking and tracing (C6)

It helps to gain a competitive advantage (Işık et al., 2020), and it is used to identify improvement areas and investment in logistics infrastructure by policymakers and practitioners (Çıray et al., 2024). It is also used for training logistics professionals (Çıray et al., 2024).

The main criticism of the LPI is the equal weight assumption (Gürler et al., 2024). There is considerable literature proposing some of the MCDM techniques to determine the weights of LPI criteria based on these comments. The results of the Chejarla et al. (2022) literature review study about performance evaluation of logistics based on 124 research articles (2010 to 2019) show that most of the papers (64%) used MCDM methods. Table 1 shows some of the studies that used MCDM techniques to determine the weights of LPI criteria in the literature based on countries and MCDM techniques.

Table 1: Literature Review about MCDM Studies based on LPI Scores

	Ulutaş and Karaköy, 2019	Yıldırım and Adigüzel Mercangoz, 2020	Işık et al., 2020	Mesic, 2022	Çalık et al., 2023	Hadzikadunic et al., 2023	Çıray et al., 2024	Gürler et al., 2024	Yılmaz, 2025	Özekenci, 2025
<b>Countries</b>										
European Union member states	√					√		√		
OECD countries		√			√					√
Top 23 countries									√	
11 Central and Eastern European Countries			√							
Western Balkans				√						
139 Countries							√			
<b>MCDM Techniques (for weighting LPI criteria)</b>										
Fuzzy AHP		√			√					
SWARA	√									
FUCOM						√				
Entropy							√	√		
SV			√							
CRITIC	√			√		√			√	√
MEREC										√
LOPCOW										√
SD										√
<b>MCDM Techniques (for ranking countries)</b>										
MARCOS				√		√		√		
WASPAS						√		√		
EDAS						√		√		

	Ulutaş and Karaköy, 2019	Yıldırım and Adigüzel Mercangöz, 2020	Işık et al., 2020	Mesic, 2022	Çalık et al., 2023	Hadzikadunic et al., 2023	Çiray et al. 2024	Gürler et al., 2024	Yılmaz, 2025	Özekenci, 2025
TOPSIS					√				√	
CRADIS										√
CoCoSo								√		
CODAS					√			√		
COPRAS								√		
GRA								√		
MABAC			√					√		
MOORA								√		
OCRA								√		
VIKOR					√					
ORESTE							√			
ARAS-G		√								
ARAS						√		√		
Integration Techniques (to integrate the results of MCDM techniques)										
AWM										√
Copeland								√		
Borda count					√					
Bonferroni						√				

Table 2 presents some of the LPI criteria weights in the literature. To make it easy to determine and compare the ranking of criteria for readers, Table 3 was derived from Table 2 to show the rankings. Although the weights differ according to year, country, and MCDM technique,

results show that mostly C3 is the criterion placed in the first rank with the highest weight. On the other hand, mostly C4 is the criterion placed in the last rank with the lowest weight.

Table 2: LPI Criteria weights in the literature

Criterion*	Ulutaş and Karaköy, 2019 CRITIC (EU Countries)	Mesic, 2022 CRITIC (Western Balkan)	Çalık et al., 2023 AHP(OECD countries)	Hadzikadunic et al., 2023 CRITIC (EU countries)	Hadzikadunic et al., 2023 FUCOM (EU Countries)	Hadzikadunic et al., 2023 Bonferroni (EU Countries)	Özekenci, 2025 CRITIC (OECD countries)	Özekenci, 2025 AWM(OECD countries)
C1	0.183**	0.119	0.182	0.153	0.132	0.142	0.169	0.170
C2	0.180	0.166	0.213	0.154	0.121	0.137	0.141	0.155
C3	0.225	0.185	0.134	0.219	0.224	0.221	0.218	0.175
C4	0.119	0.156	0.203	0.127	0.149	0.138	0.116	0.153
C5	0.157	0.207	0.164	0.192	0.179	0.185	0.165	0.165
C6	0.135	0.168	0.078	0.155	0.195	0.174	0.192	0.182

\*The ranking of criteria based on the World Bank Report. The readers need to be careful that in each article, the number of criteria is different.

\*\*Some of the values were 0.xxxx, and some of them were 0.xxx. To make a consistency. 0.xxx was adopted.

Table 3: LPI Index Criteria Rankings from highest to lowest

Ulutaş and Karaköy, 2019 CRITIC (EU Countries)	Mesic, 2022 CRITIC (Western Balkan)	Çalik et al., 2023 AHP(OECD countries)	Hadzikadunic et al., 2023 CRITIC (EU countries)	Hadzikadunic et al., 2023 FUCOM (EU Countries)	Hadzikadunic et al., 2023 Bonferroni (EU Countries)	Özekenci, 2025 CRITIC (OECD countries)	Özekenci, 2025 AWM(OECD countries)
C3	C5	C2	C3	C3	C3	C3	C6
C1	C3	C4	C5	C6	C5	C6	C3
C2	C6	C1	C6	C5	C6	C1	C1
C5	C2	C5	C2	C4	C1	C5	C5
C6	C4	C3	C1	C1	C4	C2	C2
C4	C1	C6	C4	C2	C2	C4	C4

## Methods

In this study, the aim is to carry out experiments based on prompts. Firstly, theoretical information about prompt design is provided in this section. Secondly, the designed prompts are explained in detail. Figure 1 shows the workflow diagram of this study.

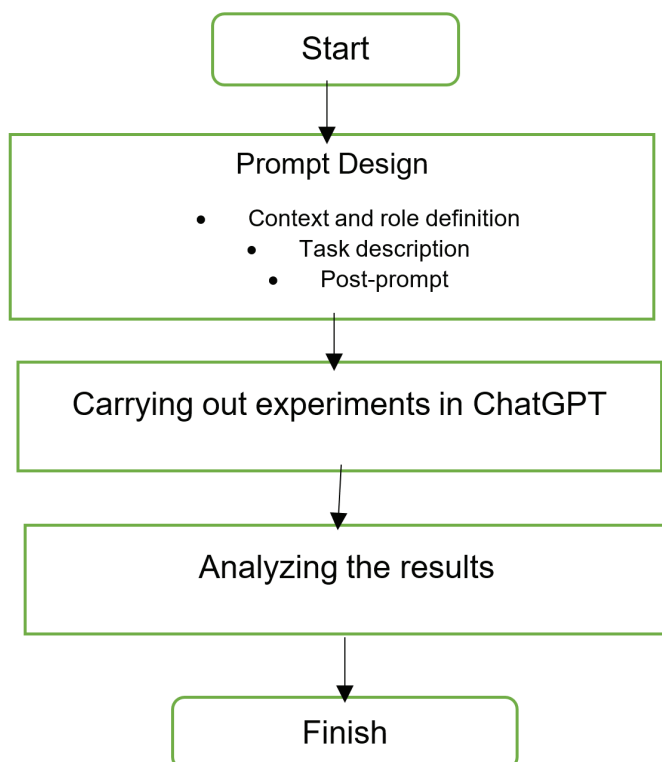


Figure 1. Workflow of the study

## Prompt Design

Bozkurt (2024) defines prompt engineering as “the art of communicating and interacting with generative AI” and introduces some useful principles and strategies. To avoid queries that are too complex or specialized, to break down complex tasks into a series of simpler questions, and to employ a strategy of sequential questioning are some of the principles (Bozkurt, 2024). In addition to assigning a role or persona, providing contextual background, identifying the scope, specifying the format, and identifying the target group or audience are some of the strategies proposed in this paper (Bozkurt, 2024). Due to the hallucination risk, Bozkurt (2024) also advises a strategy to ask for proof.

To obtain optimal results, applying pre-trained models to specific domains directly is not suitable (Wang et al., 2025). Frederico (2023) also highlights that it is highly important to ask plausible questions to provide accurate answers from ChatGPT. Ray (2023) also attracts attention; it is possible to improve the relevance and quality of the AI-generated responses with such prompt engineering techniques. The prompt method is crucial for relatively difficult tasks (Lu et al., 2025).

The study of Wang et al. (2025) shows three prompt approaches. The task description is provided directly without any examples in the first approach, called “zero-shot”. The second approach, called “few-shot”, aims to help the model with a small number of task-specific examples. “Chain of Thought (CoT)” is the third approach in which a step procedure is followed to get the final answer. The accuracy rate increase was reported by Wang et al. (2025) via CoT or few-shot prompting. The LoRA-based fine-tuning technique was also reported for further performance improvement (Wang et al., 2025).

Ray (2023) gathers up the advice overlapping with Bozkurt (2025) for prompts in five steps:

- Start with clear and specific prompts
- Provide context and background information
- Specify desired format and structure
- Apply constraints and limitations
- Iterative prompting

In the study of Wang and Wu (2024), the prompt has three main parts: the prefix prompt, main content, and post prompt. On the other hand, Dehghanimohammadabadi and Kabadayi (2025) followed a three-step approach in prompt development: context specification, persona crafting, and instruction configuration. They also added a refinement process to enhance the AI’s performance

through iterative development with testing and iterative feedback.

Svoboda and Lande (2025) created "AHP Guide" as a special custom ChatGPT. Park et al. (2025) provided that the AHP model hierarchies consist of criteria, sub-criteria, and their weights in the Appendices. These model hierarchies start with the reference model as a ground truth and continue with AHP models created by LLM without and with presenting criteria, AHP models created by LLM without and with presenting criteria and sub-criteria, fine-tuning LLM AHP without and with presenting criteria, fine-tuning LLM AHP without and with presenting criteria and sub-criteria.

In this study, the prompt design contains three main parts: Context and role definition, task description, and post-prompt.

The roles were defined based on the information of real experts in Çalık et al. (2023)'s article and role definition examples in Wang et al. (2025). There are two main roles in this study. The first decision maker (DM1) is a supply chain director, and the second decision maker (DM2) is an academician.

The task described takes into account examples in Wang et al. (2025), Dehghanimohammadabadi and Kabadayi (2025), Svoboda and Lande (2025), and Park et al. (2025). There are three prompt types Wang et al. (2025) used in this study: Zero-shot, few-shot, and CoT. The weights determined by Hadzikadunic et al. (2023) and Özekenci (2025) were used as Example 1 (E1) and Example 2 (E2) in few-shot prompts. The reason for the selection of these examples is that they used more than one MCDM technique and combined them.

Park et al. (2025) pointed out that LLMs can uncover novel evaluation criteria that human experts might overlook. In the post-prompt part, PP1 tries to learn new criteria in light of this information. In addition, to check consistency, a post-prompt (PP2) was added to compare two random criteria. It is advised to apply a refinement process if the results are not satisfactory in the study of Dehghanimohammadabadi and Kabadayi (2025). In this study, there is no need to make a refinement process because the results are satisfactory.

Zero-shot and few-shot prompts contain only one prompt. Each CoT type prompt consists of 7 prompts. There are a total of 18 prompts generated with the prompts parts below. ChatGPT is used as an LLM in this study.

#### **Context and Role definition part:**

*DM1: "You are a supply chain director, and you have ten years of expertise in the logistics sector, possessing the ability to evaluate logistics sector performance."*

*DM2: "You are an academician, and you have ten years of expertise in the logistics sector, possessing the ability to evaluate logistics sector performance."*

#### **Task description part:**

##### **Zero-shot**

*"There are six criteria: Customs, infrastructure, international shipments, logistics competence and quality, timeliness, tracking, and Tracing. You are required to assess the weights of each criterion. The results of each criterion should be between 0 and 1. In addition, the sum of all weights should be 1."*

##### **Few-shot**

*"There are six criteria: Customs, infrastructure, international shipments, logistics competence and quality, timeliness, tracking, and Tracing. You are required to assess the weights of each criterion. The results of each criterion should be between 0 and 1. In addition, the sum of all weights should be 1."*

*Example 1 (E1): Here is an example from the study of Hadzikadunic et al. (2023): weight of customs: 0.142, weight of infrastructure: 0.137, weight of international shipments: 0.221, weight of logistics competence and quality: 0.138, weight of timeliness: 0.185, weight of tracking and tracing: 0.174.*

*Example 2 (E2): Here is an example from the study of Özekenci (2025): weight of customs: 0.170, weight of infrastructure: 0.155, weight of international shipments: 0.175, weight of logistics competence and quality: 0.153, weight of timeliness: 0.165, weight of tracking and tracing: 0.182."*

##### **CoT**

*"There are six criteria: Customs, infrastructure, international shipments, logistics competence and quality, timeliness, tracking, and Tracing. You are required to assess the weights of each criterion."*

##### **Step-by-Step Analysis**

*Step 1: Do you think these criteria are equally important?*

*Step 2: What is the most important criterion?*

*Step 3: What is the least important criterion?*

*Step 4: Please rank the criteria from the (most important one) to the (least important one).*

*Step 5: Please determine weights to show the importance of each criterion. The results of each criterion should be between 0 and 1. In addition, the sum of all weights should be 1.*

**Post-prompt part:**

PP1: “Is there any other criterion you can advise to measure the logistics performance?”

PP2: “Do you think (Cx) is more important than (Cy)?”

**Results**

In this study, there are two decision makers (DM1 and DM2), three prompting types (zero shot, few shot, and CoT), and two sub-types in few shot (E1, E2). As a result, eight different weights were determined using ChatGPT, as shown in Table 4. Table 5 is derived from Table 4 to make it easy to see the ranking.

In the DM1 CoT Step-by-Step analysis, since ChatGPT gave the ranking at the end of the answer Step 3, Step 4 was skipped. In the DM2 CoT Step-by-Step analysis, since ChatGPT gave the ranking at the end of the answer Step 2, Step 4 was skipped. To analyze consistency, the Step 3 prompt was applied, and the “infrastructure” answer was verified.

Table 4: ChatGPT-based weights of LPI criteria

	DM1 Zero shot	DM1 Few shot E1	DM1 Few shot E2	DM1 CoT	DM2 Zero shot	DM2 Few shot E1	DM2 Few shot E2	DM2 CoT
C1	0.15	0.130	0.160	0.120	0.15	0.14	0.16	0.12
C2	0.20	0.180	0.170	0.190	0.20	0.13	0.15	0.08
C3	0.15	0.200	0.165	0.140	0.15	0.22	0.18	0.22
C4	0.20	0.160	0.155	0.180	0.20	0.14	0.15	0.15
C5	0.15	0.170	0.180	0.210	0.20	0.18	0.16	0.25
C6	0.15	0.160	0.170	0.160	0.10	0.19	0.20	0.18

Table 5: Rankings of LPI criteria

DM1 Zero shot	DM1 Few shot E1	DM1 Few shot E2	DM1 CoT	DM2 Zero shot	DM2 Few shot E1	DM2 Few shot E2	DM2 CoT
C2-C4	C3	C5	C5	C2- C4-C5	C3	C6	C5
-	C2	C2-C6	C2	-	C6	C3	C3
C1-C3- C5-C6	C5	-	C4	-	C5	C1-C5	C6
-	C4-C6	C3	C6	C1-C3	C1-C4	-	C4

DM1 Zero shot	DM1 Few shot E1	DM1 Few shot E2	DM1 CoT	DM2 Zero shot	DM2 Few shot E1	DM2 Few shot E2	DM2 CoT
-	-	C1	C3	-	-	C2-C4	C1
-	C1	C4	C1	C6	C2	-	C2

**DM1 results**

Table 4 shows the results for DM1:

- “Infrastructure (C2)” and “logistics competence and quality (C4)” are slightly more weighted in the answer to the “DM1 Zero shot” prompt.
- “International shipments (C3)” received the highest weight in the answer to the “DM1 Few shot E1” prompt.
- “Timeliness (C5)” received the highest weight in the answer of “DM1 Few shot E2” prompt.
- The answers of “DM1 CoT” are step-by-step
- Step 1 (Prompt asking for opinion about equal weight): “Six criteria are not equally important ..... Factors like timeliness, infrastructure, and logistics competence typically deserve higher weights due to their cross-cutting and direct impact on logistics performance. Customs and international shipments are context-dependent; their weight should reflect a country’s or industry’s trade orientation. Tracking and Tracing is gaining importance with digitalization and rising customer expectations, but still ranks just below core physical enablers.”
- Step 2 (prompt asking for most important criterion): The most important criterion is “Timeliness (C5)”
- Step 3 (prompt asking for the least important criterion): The least important criterion is “Customs (C1)” and the ranking is in Table 5.
- Step 4 (prompt asking for ranking): Skipped
- Step 5 (prompt asking for weights): The weights are listed in Table 4.
- PP-1 (prompt asking for additional criteria): Additional criteria proposed by ChatGPT are “supply chain resilience, sustainability/green logistics, cost efficiency, inventory management effectiveness, technology adoption, and last-mile delivery (optional).”

- PP2 (prompt asking for consistency check): The answer for the prompt “Do you think ‘international shipments’ is more important than ‘tracking and tracing’? Was *“No, tracking and tracing is generally more important than international shipments.”*”

## DM2 results

Table 4 also shows the results for DM2:

- “Infrastructure (C2)”, “logistics competence and quality (C4)”, and “Timeliness (C5)” are slightly more weighted in the answer of “DM2 Zero shot” prompt.
- “International shipments (C3)” received the highest weight in the answer to the “DM2 Few shot E1” prompt.
- “Tracking and tracing (C6)” received the highest weight in the answer to the “DM2 Few shot E2” prompt.
- The answers of “DM2 CoT” are step-by-step
- Step 1 (Prompt asking for opinion about equal weight): *“From my 10 years of experience in logistics, I would say these six criteria are not equally important—each plays a distinct role and impacts overall logistics performance differently..... Operational and service-related criteria like International Shipments, Timeliness, and Logistics Competence & Quality tend to carry higher weight because they directly affect supply chain performance and customer satisfaction. Structural and regulatory criteria like Customs and Infrastructure are essential but usually more stable or long-term focused. Tracking and Tracing are growing in importance as visibility becomes a key competitive advantage. If the criteria were equally weighted (each ~0.167), it would oversimplify the complex reality and potentially misrepresent the real drivers of logistics performance.”*
- Step 2 (prompt asking for most important criterion): The most important criterion is “Timeliness (C5)” and the ranking is in Table 5.
- Step 3 (prompt asking for the least important criterion): The least important criterion is “Infrastructure (C2)”
- Step 4 (prompt asking for ranking): Skipped
- Step 5 (prompt asking for weights): The weights are listed in Table 4.
- PP-1 (prompt asking for additional criteria): Additional criteria proposed by ChatGPT are

*“cost efficiency, flexibility and responsiveness, sustainability / environmental impact, safety and security, customer service/satisfaction, inventory management efficiency”.*

- PP2 (prompt asking for consistency check): The answer for the prompt “Do you think ‘international shipments’ is more important than ‘tracking and tracing’? *Whether International Shipments is more important than Tracking and Tracing really depends on the context of the logistics operation, but here’s how I see it from my experience: In traditional trade-focused logistics, International Shipments is often weighted higher. In highly digitized or customer-focused logistics, tracking and Tracing may take precedence.”*”

## Discussion

This section delves into the discussion of the results after the experiments via designed prompts with ChatGPT. The answers of “DM1 Zero shot” and “DM2 Zero shot” prompts were nearly the same, except “Timeliness (C5)” got the highest weight in the answer of “DM2 Zero shot” prompt. Indeed, there is not a huge difference between “supply chain director” and “academician” roles. On the other hand, the answer of “DM1 Zero shot” concludes with *“Let me know if you’d like to adapt this for a specific country or industry context,”* while the answer of DM2 Zero shot concludes with *“Would you like me to adjust weights for a particular region, sector, or type of logistics operation? Or, if you want, I can provide a quantitative method (like AHP or entropy weighting) to calculate these weights systematically.”* This shows the importance of role definition.

The most important criterion is “International shipments (C3)” in both of the answers of “DM1 Few Shot E1” and “DM2 Few Shot E1”. The idea behind the change of the highest weighted criterion is obviously the example given, because in E1, “international shipments (C3)” has the highest weight. This situation changes in the “Few-Shot E2” prompt. The answer to the “DM2 Few Shot E2” prompt is the same as E2, with “Tracking and tracing (C6)”, but “Timeliness (C5)” in “DM1 Few Shot E2”. The ranking is also nearly consistent with “DM2 Few Shot E1” and “E1,” but “DM1 Few Shot E1” is different. On the other hand, the ranking is nearly consistent with “DM2 Few Shot E2” and “E2”. “DM1 Few Shot E2” is different again. This result shows that the answers of the academician role (DM2) are more similar to the example, which is also an output of an academician. The idea behind the similarity of the ranking is probably the role similarity.

The highest weight is “Timeliness (C5)” in both “DM1

CoT” and “DM2 CoT,” but the other weights are not aligned with it. The answer of “DM2 CoT PP2” also needs to be interpreted by an expert.

When the obtained results are compared with the values in Table 2 and Table 3 (literature review), it is seen that DM2, Few Shot E2, and Özekenci (2025) AWM results are similar, except for equal weights in ChatGPT results. This similarity comes from the fact that the E2 data obtained from Özekenci (2025) are AWM results.

On the other hand, the results of DM2 CoT and Mesic (2022) are similar except for equal weights in ChatGPT results.

## Conclusion

LPI is an important indicator in the logistics sector performance evaluation. MCDM usage in the evaluation of logistics performance is quite common. The emergence of LLMs in the MCDM research area is expressed as a “revolution” (Wang et al., 2025) on the one hand and stated just a “significant innovation” (Dehghanimohammadabadi and Kabadayi, 2025) on the other hand. In this context, the competency of ChatGPT as a virtual expert was demonstrated with various research studies (Wang and Wu, 2024) in the literature. It is aimed to analyse the suitability of ChatGPT as an LLM to produce weights used for MCDM techniques. Eight types of prompts were designed, and weights were obtained. The results were compared with the literature. Results show that although there is an opportunity, there are also some concerns that need to be taken into consideration, as explained in the Discussion part. These concerns are consistent with Frederico’s (2023) statement: “*Technology cannot completely replace the practitioners’ ability in the decision-making process in supply chains*”.

## Theoretical and Practical Implications

The findings of this study are beneficial for researchers in the MCDM area who want to integrate LLMs in the field and researchers in logistics performance evaluation in terms of theoretical implications. On the other hand, in terms of practical implications, this study will help not only the professionals in the logistics sector but also the beginners to give insight into the criteria and their importance in logistics performance.

## Limitations

As in all studies, there are some limitations in this study. The limitations are organized in five main parts and bullets in each part below:

### Limitations of LLM

- The first limitation is using only ChatGPT as an LLM. There are some challenges and ethical issues in the literature about ChatGPT that can be considered as limitations of the study (Ray, 2023): Reliability and accuracy, overreliance on AI, quality control, dataset bias, generalization, explainability, energy consumption, real-time responsiveness, safety and privacy concerns, cultural and linguistic bias, adapting to domain-specific knowledge, adversarial attacks, misinformation, autonomy, human-like interactions are some of them. It is highly important to listen to the warnings of Park et al. (2025) regarding limitations (draft-level quality of code generation, handling data complexity, and computational performance) of LLM-based MCDM models.
- In addition, literature put forth the risk of hallucination in LLMs (Park et al., 2025). Not using a pretrained model and a fine-tuning strategy in this study increases this risk.

### Limitations of the reference model

- Using the LPI index to measure logistics performance.
- In addition, only the six criteria are predefined according to the LPI index. Post-prompt results in this study show that new criteria can be added to the reference model.

### Limitations of the MCDM side

- Using only two decision makers is a limitation of this study.
- In the current CoT type prompt, it is not known whether ChatGPT used an MCDM technique like the Best-Worst method (Rezaei, 2015) or not.
- In addition, as can be seen in the Discussion part, comparison with other MCDM techniques in the literature can not be made exactly due to the equal weights obtained with ChatGPT.

### Limitations of the experiment design

- It was determined that three factors (decision maker number with two levels, prompt type with three levels, and few-shot type with two levels) were used in the experiment design.

### Limitations of the prompt design

- As it is stated by Ray (2023), to meet specific needs using ChatGPT, one needs to gain experience in writing prompts. Due to it is the first attempt of the author(s), lack of experience is a limitation of this study.

### *Future Research*

Based on the aforementioned limitations, future research suggestions are organized in five main parts and bullets in each part below:

#### Future research about LLM

- There are other LLMs, rather than ChatGPT, that can be used and compared in future studies. Comparing the results of Wang et al. (2025) with those of other open-source LLMs or other commercial models can be a research avenue. In addition, domain-specific AI language models can be developed and used to ensure more relevant, accurate, and in-depth information (Ray, 2023).
- To reduce the risk of hallucination, it is advised to pretrain or fine-tune the model (Park et al., 2025).
- To use the potential of ChatGPT effectively, it is also advised to collaborate with human experts (Ray, 2023).

#### Future research about the reference model

- There are other indices, other than the LPI index, used to measure logistics sector performance, such as the Agility Emerging Markets Logistics Index (AEMLI) (Kara et al., 2025). Criteria in this index can be weighted with LLMs in future studies and can be compared with LPI index results.
- In addition, LLMs can be used to develop a new index to measure logistics performance with its own criteria.

#### Future research about the MCDM side

- The number of decision makers can be increased in future studies.
- It is possible to improve the explainability of ChatGPT by using additional post-prompts to learn the algorithm behind it. In this algorithm, the

normalization procedure and also calculations/statistical methods need to be taken into account.

- We used a ranking-based approach in this study. Using a pairwise comparison-based approach will be a research avenue.
- Longitudinal studies (Çıray et al., 2024) and a standardised approach for using LLMs in AHP modelling (Park et al., 2025) are other examples in the literature.
- There is also a need to entitle this emerging research area. AIDM (AI-Driven Decision Making) (Dehghanimohammadabadi and Kabadayi, 2025) is too broad, and ChatGPT-based MCDM (Wang and Wu, 2024) is too narrow. Due to this, "LLMs-based MCDM" is offered. This suggestion is also debatable.

#### Future research about the experiment design

- Experiment design can be redefined using different factors and levels in each factor. In role determination, different expertise levels (15 years/ 20 years, etc.) or genders (male/female) can also be added.

#### Future research about the prompt design

- Prompts can include sector, country, etc.
- The framing effect can also be checked with prompt designs
- Prompts can be revised not to produce equal weights.
- The proposed prompts can be applied in different contexts (e.g., healthcare, etc.) in future studies.
- The results can be compared using some statistical techniques.
- On the other hand, energy and water consumption are increasing depending on the number of prompts. Because of this concern, 18 prompts were used. Last but not least, it is very important to report the number of prompts in this kind of study as a performance criterion.
- While literature reports the accuracy rate increases with few-shot prompts, it is foreseen that this prompt style can cause biases and subjectivity problems in the MCDM weighting area. Due to the lack of information on the number of digits in prompts, nearly half of the results contain two digits after zero. To improve accuracy, adding the output style can be beneficial.

### Declaration of Generative AI in Writing

The author has used the ChatGPT free version while preparing this article to obtain weights of LPI criteria, to ask alternative criteria, and to check consistency on 30.05.2025.

## References

- Bozkurt, A. (2024). Tell Me Your Prompts and I Will Make Them True: The Alchemy of Prompt Engineering and Generative AI. *Open praxis*, 16(2), pp.111–118. doi:<https://doi.org/10.55982/open-praxis.16.2.661>.
- Çalık, A., Erdebilli, B. and Özdemir, Y.S. (2022). Novel Integrated Hybrid Multi-Criteria Decision-Making Approach for Logistics Performance Index. *Transportation Research Record: Journal of the Transportation Research Board*, p.036119812211133. doi:<https://doi.org/10.1177/03611981221113314>.
- Chejarla, K.C., Vaidya, O.S. and Kumar, S. (2021). MCDM applications in logistics performance evaluation: A literature review. *Journal of Multi-Criteria Decision Analysis*. doi:<https://doi.org/10.1002/mcda.1774>.
- Çıray, D., Özdemir, Ü. and Mete, S. (2024). An Evaluation of the logistics Performance Index Using the ENTROPY-based ORESTE Method. *Journal of Transportation and Logistics*, 0(0). doi:<https://doi.org/10.26650/jtl.2024.1437070>.
- Dehghanimohammadabadi, M. and Kabadayı, N. (2024). The Ai-Driven Decision-Making (Aidm) Framework: Integrating Ahp and Chatgpt-4 for Supplier Selection. [online] doi:<https://doi.org/10.2139/ssrn.4997750>.
- Frederico, G.F. (2023). ChatGPT in Supply Chains: Initial Evidence of Applications and Potential Research Agenda. *Logistics*, 7(2), p.26. doi:<https://doi.org/10.3390/logistics7020026>.
- Gürler, H.E., Özçalıcı, M. and Pamucar, D. (2023). Determining criteria weights with genetic algorithms for multi-criteria decision making methods: The case of logistics performance index rankings of European Union countries. *Socio-Economic Planning Sciences*, 91, pp.101758–101758. doi:<https://doi.org/10.1016/j.seps.2023.101758>.
- Hadžikadunić, A., Stević, Ž., Badi, I. and Roso, V. (2023). Evaluating the Logistics Performance Index of European Union Countries: An Integrated Multi-Criteria Decision-Making Approach Utilizing the Bonferroni Operator. *International Journal of Knowledge and Innovation Studies*, 1(1), pp.44–59. doi:<https://doi.org/10.56578/ijkis010104>.
- Işık, Ö., Aydın, Y. and Koşarolu, Ş. (2020). The assessment of the logistics performance index of cee countries with the new combination of sv and mabac methods. *LogForum*, 16(4), pp.549–559. doi:<https://doi.org/10.17270/j.log.2020.504>.
- Kara, K., Bentlyn, Z. and Yalçın, G.C. (2022). Determining the logistics market performance of developing countries by Entropy and MABAC methods. *Logforum*, 18(4), pp.421–434. doi:<https://doi.org/10.17270/j.log.2022.752>.
- Kmiecik, M. (2023). ChatGPT in third-party logistics – The game-changer or a step into the unknown? *Journal of Open Innovation: Technology, Market, and Complexity*, [online] 9(4), p.100174. doi:<https://doi.org/10.1016/j.joitmc.2023.100174>.
- Lu, X., Li, J., Takeuchi, K. and Kashima, H. (2024). AHP-Powered LLM Reasoning for Multi-Criteria Evaluation of Open-Ended Responses. *arXiv (Cornell University)*. doi:<https://doi.org/10.48550/arxiv.2410.01246>.
- Mešić, A., Miškić, S., Stević, Ž. and Mastilo, Z. (2022). Hybrid MCDM Solutions for Evaluation of the Logistics Performance Index of the Western Balkan Countries. *ECONOMICS*, 10(1), pp.13–34. doi:<https://doi.org/10.2478/eoik-2022-0004>.
- Özekenci, E.K. (2025). EVALUATION OF THE LOGISTICS PERFORMANCE INDEX OF OECD COUNTRIES BASED ON HYBRID MCDM METHODS. *Marmara Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 47(1), pp.47–76. doi:<https://doi.org/10.14780/muiibd.1469898>.
- Park, H., Oh, H., Gao, F. and Kwon, O. (2025). Enhancing Analytic Hierarchy Process Modelling Under Uncertainty With Fine-Tuning LLM. *Expert Systems*, 42(6). doi:<https://doi.org/10.1111/exsy.70051>.

- Ray, P.P. (2023). ChatGPT: a Comprehensive Review on background, applications, Key challenges, bias, ethics, Limitations and Future Scope. *Internet of Things and Cyber-Physical Systems*, [online] 3(1), pp.121–154. doi:<https://doi.org/10.1016/j.iotcps.2023.04.003>.
- Rezaei, J. (2015). Best-worst multi-criteria decision-making method. *Omega*, 53, pp.49–57. doi:<https://doi.org/10.1016/j.omega.2014.11.009>.
- Svoboda, I. and Lande, D. (2024). *Enhancing Multi-Criteria Decision Analysis with AI: Integrating Analytic Hierarchy Process and GPT-4 for Automated Decision Support*. [online] arXiv.org. Available at: <https://arxiv.org/abs/2402.07404>.
- The World Bank (2023). *Connecting to Compete 2023 Trade Logistics in the Global Economy The Logistics Performance Index and Its Indicators*. [online] Washington. Available at: [https://lpi.worldbank.org/sites/default/files/2023-04/LPI\\_2023\\_report\\_with\\_layout.pdf](https://lpi.worldbank.org/sites/default/files/2023-04/LPI_2023_report_with_layout.pdf).
- Ulutaş, A. and Karaköy, Ç. (2019). An analysis of the logistics performance index of EU countries with an integrated MCDM model. *Economics and Business Review*, 5(4), pp.49–69. doi:<https://doi.org/10.18559/ebr.2019.4.3>.
- Wamba, S.F., Guthrie, C., Queiroz, M.M. and Minner, S. (2023). ChatGPT and generative artificial intelligence: an exploratory study of key benefits and challenges in operations and supply chain management. *International Journal of Production Research*, 62(16), pp.1–21. doi:<https://doi.org/10.1080/00207543.2023.2294116>.
- Wang, H., Zhang, F. and Mu, C. (2025). *One for All: A General Framework of LLMs-based Multi-Criteria Decision Making on Human Expert Level*. [online] arXiv.org. Available at: <https://arxiv.org/abs/2502.15778>.
- Wang, X. and Wu, X. (2024). Can ChatGPT Serve as a Multi-Criteria Decision Maker? A Novel Approach to Supplier Evaluation. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.10281–10285. doi:<https://doi.org/10.1109/icassp48485.2024.10447204>.
- Yildirim, B.F. and Adiguzel Mercangoz, B. (2019). Evaluating the logistics performance of OECD countries by using fuzzy AHP and ARAS-G. *Eurasian Economic Review*, 10(1), pp.27–45. doi:<https://doi.org/10.1007/s40822-019-00131-3>.
- Yılmaz, B. (2025). Determining The Digitalization Levels of Leading Countries in Logistics Performance Index: An Application with CRITIC-TOPSIS Approach. *Verimlilik Dergisi*, 59(2), pp.431–450. doi:<https://doi.org/10.51551/verimlilik.1541480>.