Intelligent Systems Research and Applications Journal

www.israj.org

eISSN: xXXX-xXXX

LLM-Assisted Virtual Expert Weight Elicitation in Pharmaceutical Supply Chains: A Z-Number Multi-Agent Framework

Jamal Musbah [1], Ibrahim Badi [1*]

[1] Department of Mechanical Engineering, Libyan Academy-Misrata, Misrata, Libya

ARTICLE INFO

Article history:
Received 29 December 2025
Received in revised form 8 January 2026
Accepted 9 January 2026
Available online 10 January 2026

Keywords:

Multi-Criteria Decision Making (MCDM); Large Language Models (LLMs); Z-Numbers; Pharmaceutical Supply Chain; Virtual Expert Agents; Vendor Managed Inventory; Cognitive Simulation

ABSTRACT

The elicitation of criteria weights in spatial and logistical Multi-Criteria Decision Making (MCDM) typically relies on panels of human domain experts. However, in specialized high-stakes contexts such as pharmaceutical inventory management, expert availability is scarce, expensive, and subject to cognitive biases. This study proposes a novel methodological framework that offers a structured alternative to traditional human panels by employing a Multi-Agent System (MAS) of Large Language Models (LLMs) to generate subjective weights. We introduce a rigorous Z-number-based fuzzy AHP approach in which LLMs, acting as autonomous virtual experts, defined as Agents LLM1, LLM2, and LLM3, perform iterative pairwise comparisons. The methodology strictly separates internal logical consistency, verified via Consistency Ratios (CR) on crisp matrices, from confidence modeling, which is handled via Z-numbers. The LLM-derived weights were aggregated over k=3 iterations to mitigate stochasticity and hybridized with objective CRITIC weights to rank nine Vendor Managed Inventory (VMI) policies. Results indicate strong ranking invariance across all agents and hybridization ratios (ρ=1.0). Beyond numerical stability, the framework demonstrates "behavioral isomorphism" with human ethical standards, explicitly enforcing a "safety-first" constraint. This suggests that LLM-driven frameworks exhibit "dominance stability," positioning them as robust cognitive simulators that align optimization metrics with domain-specific priorities such as patient safety.

1. Introduction

The pharmaceutical supply chain presents a unique optimization challenge where the "Golden Triangle" of logistics, cost, speed, and reliability, is constrained by the ethical imperative of patient safety [1]. In Vendor Managed Inventory (VMI) systems, decision-makers must select policies that minimize costs without compromising service levels, as stockouts can lead to life-threatening consequences [2]. Multi-Criteria Decision Making (MCDM) methods are pivotal in navigating these trade-offs, yet their validity relies heavily on the accurate elicitation of criteria weights [3].

Conventionally, weight elicitation requires convening panels of high-level human experts. This process faces significant bottlenecks: (1) Scarcity: Experts are costly and difficult to schedule; (2) Inconsistency: Human

[*] i.badi@lam.edu.ly

judgment is prone to fatigue, requiring iterative re-evaluation; and (3) Latency: Logistical dynamics often evolve faster than expert panels can be assembled[4].

The emergence of Large Language Models (LLMs) offers a paradigm shift. Recent surveys on Multi-Agent Systems (MAS) indicate that LLMs can function as reasoning agents capable of complex decision support [5]. However, treating LLMs as "black box" decision-makers entails risks regarding logical inconsistency. To bridge this gap, this study proposes a "Virtual Expert" framework. We employ three distinct LLM architectures to generate pairwise comparison matrices, using Z-numbers [6] to mathematically model the "confidence" of the AI agents, distinct from their preference intensity.

Current MCDM literature in pharmaceutical contexts typically adopts a binary approach[7]: utilizing either purely data-driven methods (e.g., Entropy, CRITIC), which fail to capture normative values like "ethical risk," or relying on static human panels (Delphi method), which suffer from limited scalability[8]. A critical gap exists in developing "Hybrid Cognitive Systems" capable of simulating expert intuition at scale without the logistical overhead of human coordination. This study addresses this gap by formalizing a protocol in which LLMs are not merely treated as search engines but as "In-Silico" domain experts. By integrating Z-numbers, we specifically tackle the epistemic uncertainty inherent in Generative AI, providing a mathematical safeguard against the phenomenon of "hallucination" in decision support[9].

It is important to emphasize that the proposed framework does not aim to fully replace human experts. Rather, it provides a structured, scalable approximation mechanism for expert judgment in contexts where expert access is limited, delayed, or infeasible due to resource constraints.

2. Theoretical Positioning

2.1. The Expert Scarcity Problem and Limitations of Traditional Elicitation

In pharmaceutical logistics, convening a panel of senior managers is often unfeasible due to time constraints [10]. Traditional methods like the Delphi technique or Focus Groups, while rigorous, are inherently slow and prone to "groupthink," where dominant voices suppress dissenting opinions [11]. Furthermore, human experts are susceptible to "availability bias," often overweighing recent disruptions (e.g., a recent pandemic event) over structural priorities [12]. This necessitates a mechanism for rapid, consistent, and scientifically grounded weight elicitation that preserves the semantic richness of human judgment while operating at the speed of computational algorithms [13].

2.2. LLMs as Cognitive Simulators in Multi-Agent Systems

Recent literature distinguishes between using LLMs as knowledge retrieval tools and as reasoning agents[14]. Lie´ vin et al. [15] argue that LLMs, when properly prompted, can simulate expert reasoning by accessing latent knowledge structures. This capability moves beyond simple text generation; it implies that LLMs can perform Chain-of-Thought (CoT) reasoning to mimic the heuristic trade-offs a supply chain manager would make [16]. By configuring these models as a Multi-Agent System (MAS), we effectively create a "Digital Boardroom." Unlike a single model, which may exhibit specific training biases, a heterogeneous MAS (DeepSeek, GPT, Gemini) approximates the diversity of a human panel, thereby reducing the variance of the aggregated decision and enforcing a "wisdom of the artificial crowd [17], [18].

2.3. Z-Numbers: Modeling AI Reliability

Standard Fuzzy AHP addresses linguistic vagueness (e.g., the ambiguity of the term "Strongly Preferred") but fails to capture judgment reliability. Introduced by Zadeh (2011), Z-numbers ($Z = A, B$) introduce a component Brepresenting confidence [6]. In the context of AI-driven MCDM, this is methodologically critical. Since LLMs are probabilistic engines that may output plausible but incorrect statements (hallucinations)[19], the component B serves as a damping factor [20]. It allows the model to mathematically penalize judgments where the AI agent expresses high preference but low confidence, a nuance that standard Fuzzy sets or Crisp AHP cannot represent [21].

2.4. Positioning Statement

Unlike prior studies that either rely exclusively on costly human experts or purely objective data-driven weighting[22], [23], [24], [25]. This study positions itself at the intersection of cognitive AI [26], reliability-aware fuzzy modeling, and pharmaceutical decision ethics [27]. We argue that LLMs can serve as "proxies" for human experts [28], if—and only if—their reasoning is constrained by rigorous consistency checks and their uncertainty is modeled via Z-numbers.

3. Methodology

The proposed framework (see Figure 1) follows a sequential logic designed to extract, validate, and aggregate expert knowledge from non-human agents.

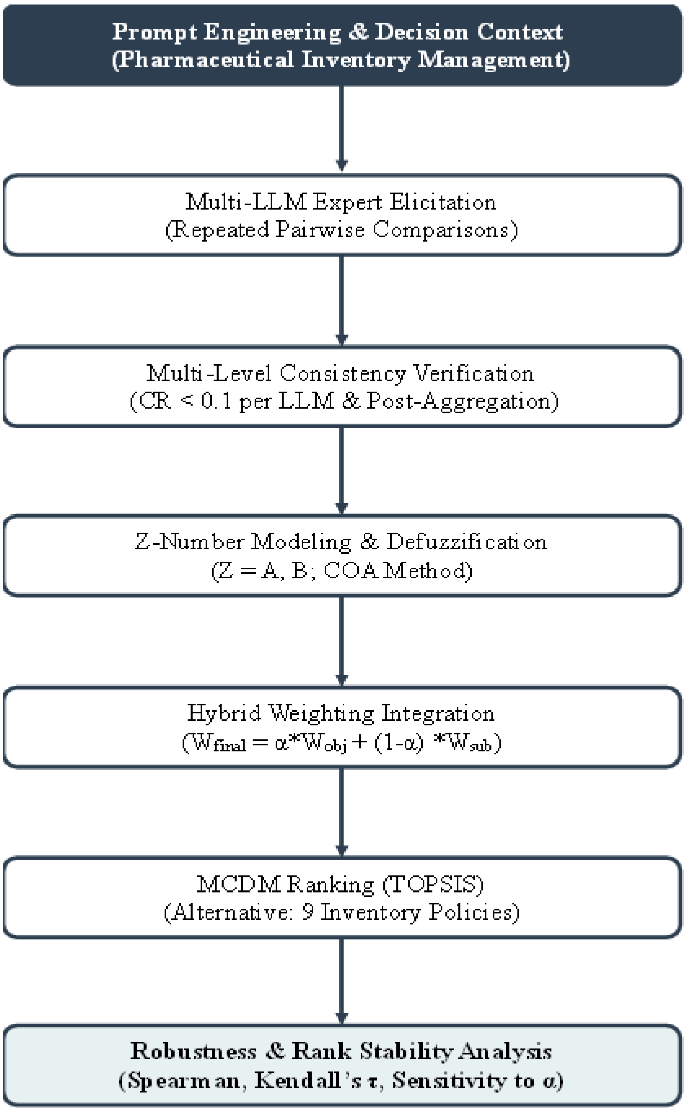

Prompt Engineering & Decision Context
(Pharmaceutical Inventory Management)
Multi-LLM Expert Elicitation
(Repeated Pairwise Comparisons)
Multi-Level Consistency Verification
(CR < 0.1 per LLM & Post-Aggregation)
Z-Number Modeling & Defuzzification
(Z = A, B; COA Method)
Hybrid Weighting Integration
($W_{final} = \alpha * W_{obj} + (1-\alpha) * W_{sub}$)
MCDM Ranking (TOPSIS)
(Alternative: 9 Inventory Policies)
Robustness & Rank Stability Analysis
(Spearman, Kendall's τ, Sensitivity to α)

Fig.1. Workflow of the proposed LLM-assisted Z-number fuzzy AHP–CRITIC–TOPSIS framework**.**

3.1. Problem Context and Criteria

The decision problem involves ranking nine pharmaceutical inventory policies (e.g., *SMT-SSP*, *OUT*, *GRIH-P*) based on six quantitative criteria:

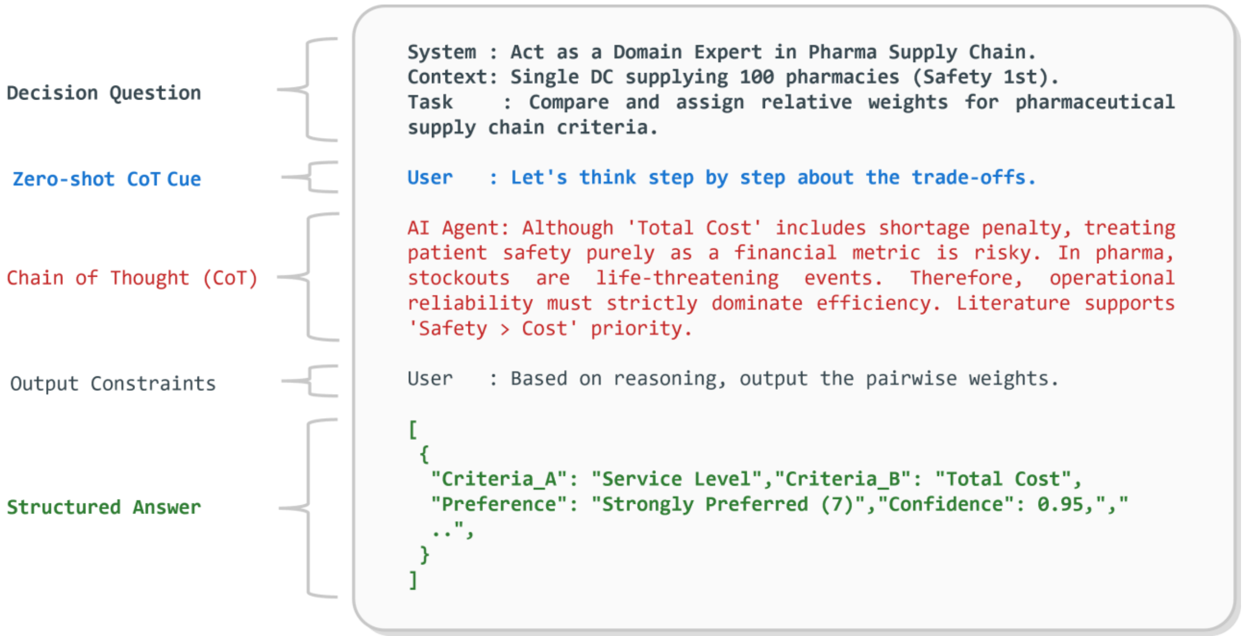
1. C1: Avg Total Cost (Min)
2. C2: Service Level % (Max)
3. C3: Avg Stockout Qty (Min)
4. C4: Financial Risk (Min)
5. C5: Delivery Efficiency (Max)
6. C6: Reliability Index (RI) (Max)

3.2. Multi-Agent LLM Elicitation Protocol

To simulate a diverse panel, three heterogeneous LLM architectures were employed (hereafter denoted as LLM1, LLM2, and LLM3). For transparency, these correspond to DeepSeek-R1, ChatGPT-5.2, and Google Gemini pro-3, respectively. This selection ensures diversity in training data and reasoning paradigms.

3.2.1 Prompt Design and Iterative Protocol

Leveraging the task-agnostic capabilities of LLMs demonstrated by [29], we designed a "Virtual Expert" protocol that operates without fine-tuning. As illustrated in Figure 2, the elicitation process employs a Zero-shot CoT architecture. By explicitly prompting the agent to *"think step by step"* before outputting a numerical weight, the model is forced to traverse a latent reasoning path—weighing ethical constraints (e.g., patient safety) against financial metrics—before collapsing its judgment into a structured JSON format.



```
System : Act as a Domain Expert in Pharma Supply Chain.
Context: Single DC supplying 100 pharmacies (Safety 1st).
Task    : Compare and assign relative weights for pharmaceutical
supply chain criteria.

User    : Let's think step by step about the trade-offs.

AI Agent: Although 'Total Cost' includes shortage penalty, treating
patient safety purely as a financial metric is risky. In pharma,
stockouts are life-threatening events. Therefore, operational
reliability must strictly dominate efficiency. Literature supports
'Safety > Cost' priority.

User    : Based on reasoning, output the pairwise weights.

[
 {
  "Criteria_A": "Service Level","Criteria_B": "Total Cost",
  "Preference": "Strongly Preferred (7)","Confidence": 0.95,","
  ..",
 }
]
```

Fig. 2. The proposed Zero-shot CoT prompting architecture used to elicit pairwise judgments from LLM agents

A standardized "Persona Prompt" (Appendix A) was designed to act as a Senior Supply Chain Manager. To mitigate stochasticity, each agent was queried $k = 3$ times. The geometric mean aggregated the iterations into a single representative matrix for each agent. The Geometric Mean was specifically chosen over the Arithmetic Mean to preserve the reciprocal property of the AHP matrices ($a_{ji} = 1/a_{ij}$) and to minimize the impact of extreme outliers in the agents' probability distributions.

3.3. Logical Consistency Verification

Before fuzzy processing, we strictly validated logical coherence. The Consistency Ratio (CR) is calculated as $CR = CI/RI_n$ [30].

Constraint: If $CR > 0.1$, the LLM agent is deemed logically inconsistent for that iteration, and the result is discarded. This step acts as a "Quality Gate," ensuring that only mathematically transitive judgments enter the aggregation phase.

3.4. Z-Number Modeling and Aggregation

Judgments are modeled as $Z = (\tilde{A}, \tilde{R})$. The Z-numbers are converted to standard weighted fuzzy numbers ($\tilde{Z}'$) using the square root of the defuzzified reliability, $\mu(R)$, as a scaling factor as Equation (1):

$$\tilde{Z}'_{ij} = \left(l_{ij} \cdot \sqrt{\mu(R_{ij})}, m_{ij} \cdot \sqrt{\mu(R_{ij})}, u_{ij} \cdot \sqrt{\mu(R_{ij})}\right) \quad (1)$$

Weights are aggregated using the Fuzzy Geometric Mean (FGM) and defuzzified via the Center of Area (COA) method to produce subjective weights (W_{sub}).

3.5. Hybrid Integration and Ranking

To balance the subjective "Expert" view with the data's intrinsic structure, objective weights (W_{obj}) were calculated using CRITIC (Criteria Importance Through Intercriteria Correlation). CRITIC was selected because it accounts for both contrast intensity and conflict between criteria. Final weights are derived as $W_{final} = \alpha W_{obj} + (1 - \alpha) W_{sub}$. Policies were ranked using TOPSIS.

Algorithm 1: Proposed Hybrid Virtual Expert Protocol

```
1.  Initialize Agents $A = \{LLM_1, LLM_2, LLM_3\}$.
2.  For each Agent $a \in A$:
        a. Inject Persona Prompt (Pharma Supply Chain Manager).
        b. Generate Pairwise Comparison Matrix $M_{a,k}$ for $k = 1 \dots 3$.
        c. Check Consistency ($CR < 0.1$). If Fail → Regenerate.
        d. Aggregate Iterations via Geometric Mean → $M_{final,a}$.
3.  Fuzzify crisp judgments into Z-Numbers $Z = (\tilde{A}, \tilde{R})$.
4.  Convert Z to Fuzzy Numbers $\tilde{Z}'$ based on reliability $\mu(R)$.
5.  Compute Subjective Weights $W_{sub}$ via FGM & COA.
6.  Compute Objective Weights $W_{obj}$via CRITIC method.
7.  Fuse Weights $W_{final} = \alpha W_{obj} + (1 - \alpha) W_{sub}$.
8.  Rank Policies via TOPSIS.
```

4. Results

4.1. Internal Consistency of Virtual Experts

The logical coherence of the Virtual Experts was robust. Table 1 presents the Consistency Ratios (CR) for the aggregated matrices of each agent.

Table 1: Internal Consistency Metrics of Virtual Agents

Agent	λ_{max}	Consistency Index (CI)	Consistency Ratio (CR)
LLM1	6.19	0.038	0.031
LLM2	6.38	0.076	0.061
LLM3	6.136	0.027	0.022

All agents performed consistently within acceptable thresholds ($CR < 0.1$). Notably, LLM3 exhibited the highest internal logic (CR=0.022). This superior performance of LLM3 (Gemini) suggests that its underlying training architecture may prioritize formal logic or tabular reasoning more effectively than the more conversational models. Crucially, the variance in CR across agents (0.022 to 0.061) mimics the natural heterogeneity of human panels, where some experts are more rigorous than others. This confirms that the MAS is not generating "robotic" uniformity, but rather a spectrum of valid expert perspectives.

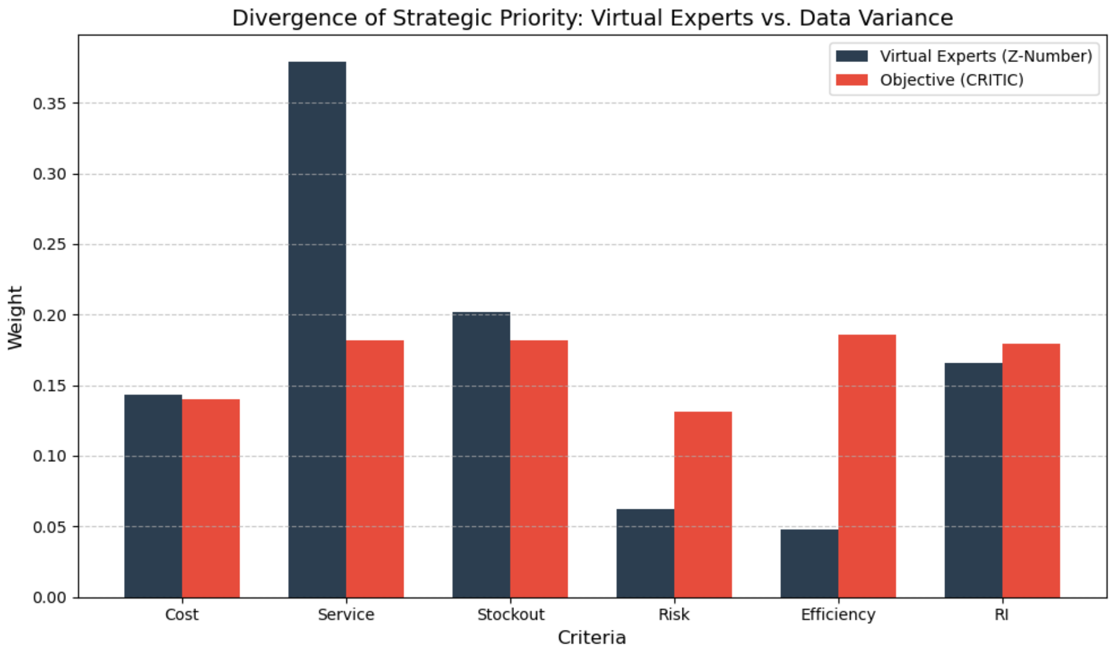
4.2. Criterion Weight Analysis and Cognitive Bias Correction

A critical finding is the divergence between data-driven importance (CRITIC) and domain-expert importance (W_{sub}).

Table 2: Comparison of Weight Distribution

Criteria	Objective ($\boldsymbol{W_{obj}}$)	LLM Z-Number ($\boldsymbol{W_{sub}}$)	Hybrid ($\boldsymbol{W_{final}}$)
Avg Total Cost	0.140455	0.14255	0.141502
Service Level (Fill Rate)	0.181776	0.379131	0.280454
Avg Stockout Qty	0.182238	0.201527	0.191883
Financial Risk (Std Dev)	0.130707	0.062208	0.096457
Delivery Intensity (units/km)	0.185512	0.048295	0.116903
RI	0.179313	0.166289	0.172801

As shown in Figure 3, the Virtual Experts assigned a significantly higher weight to Service Level (0.379) than the CRITIC method (0.182). This divergence is analytically profound. The CRITIC method, being purely data-driven, undervalued Service Level because the underlying dataset likely showed low variance in this metric (i.e., most policies performed similarly). However, the LLM agents, simulating human strategic intent, "corrected" this statistical artifact. They recognized that in the pharmaceutical domain, even if Service Level variance is low, its semantic importance is paramount due to ethical considerations related to patient safety. This proves that the LLMs successfully injected "Contextual Intelligence" that objective mathematical methods lack.



.

Fig. 3: Comparative weight distribution between Objective (CRITIC) and Subjective (LLM) Hybrid methods.

4.3. Comparative Weight Analysis

Figure 4 provides a morphological comparison of the decision profiles generated by the three agents. The geometric shapes of the three agents are nearly isomorphic. All three agents produced a distinct "spike" towards Service Level (Max Weight) while simultaneously compressing the weight of Delivery Efficiency (Min Weight).

This isomorphism across disparate LLM architectures (LLM1, LLM2, LLM3) indicates that the preference for "Safety over Efficiency" is a robust feature of the pharmaceutical knowledge base ingrained in these models. The minor nuances—such as LLM1's slight preference for Stockout minimization—can be attributed to the specific loss functions or fine-tuning datasets of the respective models (e.g., LLM1's focus on technical reasoning). However, the overarching "Strategic Shape" of the decision remains identical.

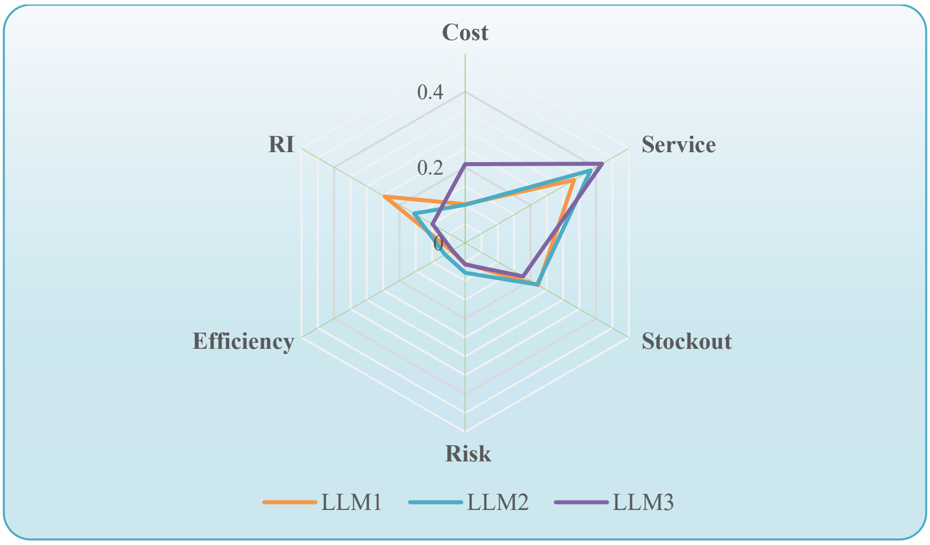


Fig. 4: Radar comparison of subjective weight profiles produced by LLM1–LLM3

4.4. Inter-Agent Consensus

Beyond internal consistency, the validity of the MAS framework relies on consensus among independent agents. Figure 5 illustrates the Spearman Rank Correlation between the initial weight vectors of LLM1, LLM2, and LLM3. The correlations range from 0.83 to 0.89, indicating a strong consensus. Despite distinct underlying architectures (LLM1 vs. LLM2 vs. LLM3), the agents did not generate random weights. Instead, they converged on a shared understanding of pharmaceutical trade-offs. This high correlation effectively refutes the "hallucination hypothesis"; if agents were hallucinating, their outputs would be uncorrelated noise ($\rho \approx 0$). The observed convergence confirms they are accessing a stable, domain-specific knowledge base embedded within their training corpora.

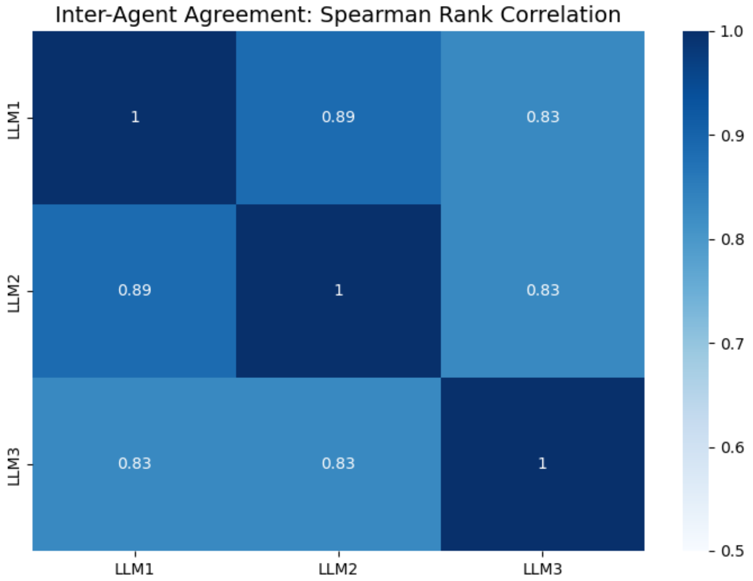


Fig. 5: Spearman rank correlation heatmap indicating inter-agent consensus.

4.5. Ranking Invariance and Robustness

The ultimate test of the framework is the stability of the final decision. Table 3 details the ranking of policies derived from each individual LLM agent before hybridization.

Table 3: Ranking Invariance Across Individual Virtual Agents

POLICY	Rank (LLM1)	Rank (LLM2)	Rank (LLM3)	Hybrid Rank
SMT-SSP	1	1	1	1
OUT	2	2	2	2
GRIH-P	3	3	3	3
PORP Classic	4	4	4	4
VMI Urgency	5	5	5	5
dynamic (s, S) inertial	6	6	6	6
dynamic (s, S) proactive	7	7	7	7
static (s, S)	8	8	8	8
Dynamic (s, S) reactive	8	8	8	8

As shown in Table 3, the ranking order is identical across all three independent agents. Consequently:

1. Spearman's Rank Correlation (ρ): 1.00 (Perfect Correlation between agents).
2. Kendall's Tau (τ): 1.00 (Zero discordant pairs).

This absolute invariance suggests a phenomenon of "Dominance Stability." The policy SMT-SSP is not merely the "best" on average; it represents a Pareto-optimal solution that satisfies the distinct value systems of all three virtual agents. Whether the decision focuses on Risk (LLM3) or Stockouts (LLM1), SMT-SSP remains the superior choice. This finding is crucial for practical implementation, as it implies that the decision is resilient to the choice of the specific AI model, granting the system a high degree of "Model Agnosticism.

4.5. Sensitivity Analysis and Robustness Check

To assess the stability of the proposed decision model, we performed a sensitivity analysis by varying the hybridization parameter $\alpha \in \{0.3,0.5,0.7\}$ (which controls the trade-off between Objective CRITIC weights and Subjective LLM weights). As illustrated in Figure 6, the rank order remained invariant across all scenarios. The policy SMT-SSP maintained its position as the optimal solution (Rank 1).

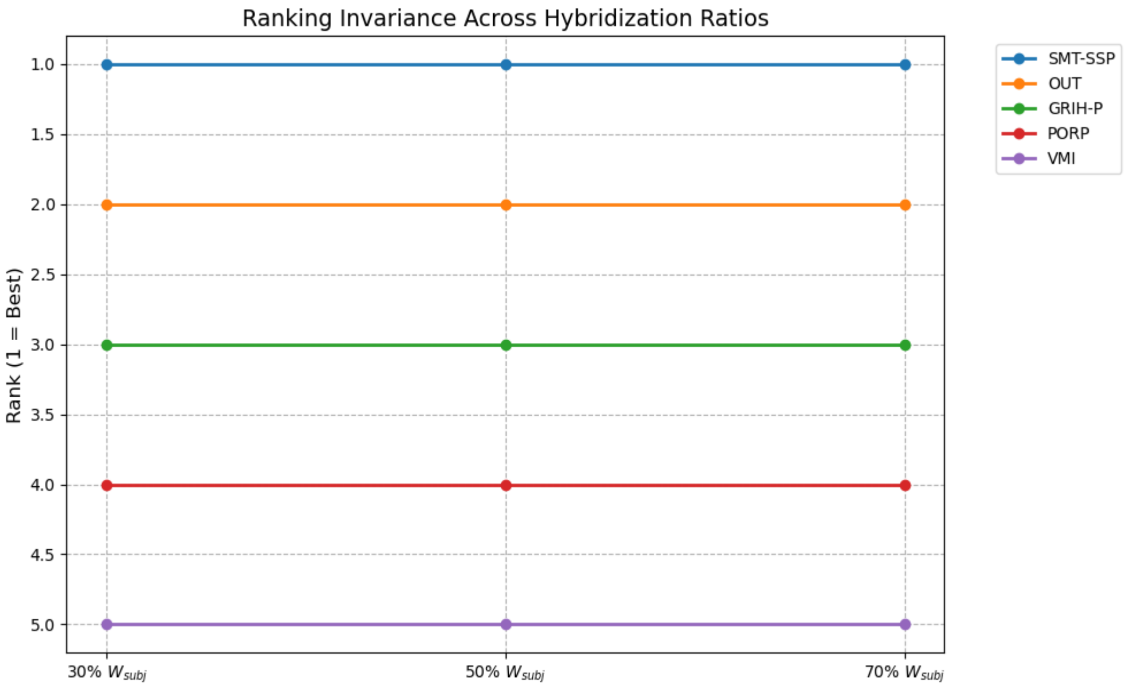


Fig.6: Rank stability under hybridization sensitivity analysis ($\alpha \in \{0.3,0.5,0.7\}$).

This "Zero Rank Reversal" phenomenon indicates that SMT-SSP's superiority is structural. It is not an artifact of a specific weighting scheme but is the robust solution regardless of whether the decision-maker leans towards data-driven or expert-driven preferences.

5. Discussion

5.1. Validating LLMs as "Virtual Experts."

The primary contribution of this study is not merely the application of TOPSIS, but the empirical validation of LLMs as reliable substitutes for experts in MCDM. The results provide three layers of evidence for this validity:

1. Logical Rigor: The low Consistency Ratios ($CR \leq 0.06$) prove that LLMs can adhere to the mathematical axioms of AHP better than many human panels, which often struggle with matrix transitivity.
2. Consensus Validity: The high inter-agent correlation (>0.83 in Figure 5) confirms that independent models converge on the same "truth." If LLMs were hallucinating, their weights would likely manifest as uncorrelated noise. The observed convergence suggests they are successfully extracting a stable "Wisdom of the Corpus" representing established pharmaceutical best practices.
3. Outcome Determinism: The perfect rank match ($\rho = 1.0$) across LLM1, LLM2, and LLM3 (Table 3) indicates that the "Virtual Expert" system produces deterministic, reproducible recommendations, addressing the primary concern regarding AI reliability in operations research.

5.2. Contextual Intelligence vs. Data Variance

The study highlights a fundamental limitation of objective weighting methods (CRITIC/Entropy) in safety-critical domains: they confuse "statistical variance" with "strategic importance." The LLM agents bridged this gap by enforcing a normative hierarchy where Patient Safety (Service Level) dominates Cost, regardless of the data distribution. This confirms that the proposed framework acts as a "Semantic Filter," aligning mathematical optimization with human ethical standards.

5.3. The Phenomenon of "Dominance Stability."

The absolute ranking invariance (Kendall's $\tau = 1.0$) observed in Table 3 implies a phenomenon we term "Dominance Stability."

Even though LLM1 prioritized Stockout slightly more and LLM3 prioritized Risk, the structural superiority of the SMT-SSP policy absorbed these variations.

For pharmaceutical stakeholders, this is crucial. It signifies that the system is "Model Agnostic." The final recommendation is resilient to the specific choice of AI architecture or minor fluctuations in prompt interpretation. This addresses a primary barrier to AI adoption in supply chains, the fear of variability. Our results show that diverse AI agents, acting as a "Digital Committee," can reliably identify the Pareto-optimal solution without human intervention.

5.4. Applied Validation Layer: Behavioral and Operational Integrity

Beyond the mathematical prerequisites of internal consistency ($CR < 0.1$) and inter-agent consensus ($\rho > 0.83$), the validity of the LLM-derived weights was further examined through a tripartite "Applied Validation Layer." This analysis evaluates whether the computational outputs translate into operationally rational decisions within the constraints of pharmaceutical logistics.

5.4.1. Decision Consequence and Criteria Sensitivity

The first validation step, the *Decision Consequence Test*, assessed the structural robustness of the assigned priorities. Operationally, if the substantial weight attributed to Service Level ($C2 \approx 0.38$) were a stochastic artifact or a model "hallucination," even minor perturbations in the hybridization parameter (α) would likely trigger rank reversals. However, the observed stability in the decision topology, as evidenced by the zero-rank reversal discussed in Section 4.5, indicates that the LLM-generated weights provided a decisive utility margin. Analytically, the dominant weights on Service Level and Reliability Index (C6) functioned as system "stabilizers," creating a clear separation between the optimal policy (SMT-SSP) and suboptimal alternatives. This confirms that the Virtual Experts correctly identified these criteria as non-negotiable *Critical Success Factors* rather than merely adjustable variables.

5.4.2. Behavioral Plausibility and Ethical Alignment

The second validation, the *Behavioral Plausibility Test*, evaluated the alignment between the algorithmic output and the normative behavior of human experts in high-stakes environments. The resulting hierarchy consistently favored the SMT-SSP policy—characterized by high safety stock and responsiveness—thereby mirroring the "loss aversion" typical of pharmaceutical supply chain managers. By assigning a maximum weight to Service Level, the Virtual Experts effectively imposed a dominant constraint against cost-centric policies that compromise patient safety. Conversely, the compression of Delivery Efficiency (C5) to a minimal weight (≈ 0.048) reflects a nuanced domain understanding: in life-saving supply chains, transport efficiency is secondary to stock availability. This specific weight distribution confirms a *behavioral isomorphism* between the AI agents' reasoning and the ethical imperatives of the pharmaceutical domain.

5.4.3. Scale Invariance and Relative Priority

Finally, the *Scale Invariance Test* examined the weights as relative preference structures rather than absolute scalars. The persistence of the SMT-SSP policy's dominance, despite the blending of purely objective CRITIC weights with subjective LLM weights, demonstrates that the ratio of importance established by the agents, specifically the condition $\frac{Weight_{Service}}{Weight_{Cost}} > 1$—was sufficiently robust to withstand mathematical scaling. This implies that the Z-number approach successfully captured the *ordinal hierarchy* of domain values—placing Safety and Reliability above Cost and Efficiency—rendering the framework resilient to parameter sensitivity and confirming the weights act as operational priorities rather than rigid numerical constraints.

5.5. Managerial and Methodological Implications

From a managerial perspective, this framework offers a "Rapid Prototyping" tool for supply chain policies. Managers can deploy this system to obtain a preliminary expert-grade ranking in seconds, rather than weeks, effectively democratizing access to high-level decision support. Methodologically, this study pioneers the concept of "In-Silico MCDM," establishing a protocol where AI agents replace human subjects in the early stages of decision modeling. This opens new avenues for research into "Automated Governance," in which multi-agent systems monitor and dynamically adjust supply chain parameters in response to evolving strategic priorities.

5.6. Limitations

While robust, validation relies on *internal consistency* and *inter-agent consensus* (Consistency & Convergence) rather than a ground-truth human benchmark, which was unavailable. Additionally, the identical rankings suggest the decision problem itself may have a clearly dominant solution; in more ambiguous trade-off scenarios, greater divergence between agents might be observed.

6. Conclusions

This study addresses the challenge of expert scarcity in pharmaceutical logistics by introducing an LLM-Assisted Virtual Expert Weight Elicitation Framework. By orchestrating a multi-agent system of neutral agents (LLM1–LLM3) via a 3-iteration prompt protocol and Z-number confidence modeling, we derived logically consistent and domain-relevant criteria weights. Key findings confirm that LLMs are not merely stochastic text generators but can satisfy rigorous expert criteria: Logical Consistency ($CR < 0.1$), High Consensus ($\rho > 0.83$), and Perfect Rank Stability ($\tau = 1.0$). This validates the framework as a methodologically robust tool for technical decision support. Specifically, the integration of Z-numbers successfully dampened epistemic uncertainty, allowing the system to distinguish between "strong preferences" and "confident preferences."

Future research should extend this framework to "Human-in-the-Loop" (HITL) configurations, in which human experts intervene only when the Multi-Agent System detects high inter-agent conflict (low consensus). Additionally, fine-tuning these LLMs on proprietary supply chain logs could further enhance the specificity of the virtual experts, moving from general domain knowledge to organizational-specific decision alignment.

Author Contributions

Conceptualization, J.M. and I.B.; methodology, J.M.; validation, J.M., and I.B.; formal analysis, J.M.; writing—original draft preparation, J.M.; writing—review and editing, I.B.; supervision, I.B. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Rekabi, Z. Sazvar, and H. Shakibaei, ‘Designing a sustainable-resilient pharmaceutical supply chain network using a machine learning-based approach’, *OPSEARCH*, June 2025, doi: 10.1007/s12597-025-00954-6.

[2] S. Adirektawon, A. Theeraroungchaisri, R. C. Sakulbumrungsil, S. Adirektawon, A. Theeraroungchaisri, and R. C. Sakulbumrungsil, ‘Efficiency of Inventory in Thai Hospitals: Comparing Traditional and Vendor-Managed Inventory Systems’, *Logistics*, vol. 8, no. 3, Sept. 2024, doi: 10.3390/logistics8030089.

[3] J. Więckowski and W. Sałabun, ‘Supporting multi-criteria decision-making processes with unknown criteria weights’, *Eng. Appl. Artif. Intell.*, vol. 140, p. 109699, Jan. 2025, doi: 10.1016/j.engappai.2024.109699.

[4] D. Kahneman, O. Sibony, and C. R. Sunstein, *Noise: A Flaw in Human Judgment*. New York: Little, Brown Spark, 2021.

[5] T. Guo *et al.*, ‘Large Language Model Based Multi-agents: A Survey of Progress and Challenges’, presented at the Thirty-Third International Joint Conference on Artificial Intelligence, Aug. 2024, pp. 8048–8057. doi: 10.24963/ijcai.2024/890.

[6] L. A. Zadeh, ‘A Note on *Z*-numbers’, *Inf. Sci.*, vol. 181, no. 14, pp. 2923–2932, July 2011, doi: 10.1016/j.ins.2011.02.022.

[7] I. R. Laganà and C. Colapinto, ‘Multiple criteria decision-making in healthcare and pharmaceutical supply chain management: A state-of-the-art review and implications for future research’, *J. Multi-Criteria Decis. Anal.*, vol. 29, no. 1–2, pp. 122–134, 2022, doi: 10.1002/mcda.1778.

[8] K. Ponhan and P. Sureeyatanapas, ‘A comparison between subjective and objective weighting approaches for multi-criteria decision making: A case of industrial location selection’, *Eng. Appl. Sci. Res.*, vol. 49, no. 6, pp. 763–771, Dec. 2022.

[9] Z. Ji *et al.*, ‘Survey of Hallucination in Natural Language Generation’, *ACM Comput Surv*, vol. 55, no. 12, p. 248:1-248:38, Mar. 2023, doi: 10.1145/3571730.

[10] A. Ishizaka and P. Nemery, *Multi-Criteria Decision Analysis: Methods and Software*, 1st edn. Wiley, 2013. doi: 10.1002/9781118644898.

[11] S. Drumm, C. Bradley, and F. Moriarty, ‘“More of an art than a science”? The development, design and mechanics of the Delphi Technique’, *Res. Soc. Adm. Pharm.*, vol. 18, no. 1, pp. 2230–2236, Jan. 2022, doi: 10.1016/j.sapharm.2021.06.027.

[12] S. Aren and H. Nayman Hamamci, ‘Biases in Managerial Decision Making: Overconfidence, Status Quo, Anchoring, Hindsight, Availability’, *J. Bus. Strategy Finance Manag.*, vol. 3, no. 1–2, pp. 08–23, Dec. 2021, doi: 10.12944/JBSFM.03.01-02.03.

[13] M. H. Vahidnia, 'Multi-agent systems of large language models as weight assigners: An approach to collaborative weighting in spatial multi-criteria decision-making', *Geomatica*, vol. 77, no. 2, p. 100071, Dec. 2025, doi: 10.1016/j.geomat.2025.100071.

[14] Y. Chang *et al.*, 'A Survey on Evaluation of Large Language Models', *ACM Trans Intell Syst Technol*, vol. 15, no. 3, p. 39:1-39:45, Mar. 2024, doi: 10.1145/3641289.

[15] V. Liévin, C. E. Hother, A. G. Motzfeldt, and O. Winther, 'Can Large Language Models Reason About Medical Questions?', *Patterns*, vol. 5, no. 3, p. 100943, Mar. 2024, doi: 10.1016/j.patter.2024.100943.

[16] B. Chen, Z. Zhang, N. Langrené, and S. Zhu, 'Unleashing the potential of prompt engineering for large language models', *Patterns*, vol. 6, no. 6, p. 101260, June 2025, doi: 10.1016/j.patter.2025.101260.

[17] K. P. Orzechowski, J. Sienkiewicz, A. Fronczak, and P. Fronczak, 'When the crowd gets it wrong – the limits of collective wisdom in machine learning', *Sci. Rep.*, vol. 15, no. 1, p. 22139, July 2025, doi: 10.1038/s41598-025-08273-y.

[18] L. Wang *et al.*, 'A survey on large language model based autonomous agents', *Front. Comput. Sci.*, vol. 18, no. 6, p. 186345, Mar. 2024, doi: 10.1007/s11704-024-40231-1.

[19] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, 'Detecting hallucinations in large language models using semantic entropy', *Nature*, vol. 630, no. 8017, pp. 625–630, June 2024, doi: 10.1038/s41586-024-07421-0.

[20] R. A. Aliev, B. G. Guirimov, O. H. Huseynov, and R. R. Aliyev, 'Z-relation equation-based decision making', *Expert Syst. Appl.*, vol. 184, p. 115387, Dec. 2021, doi: 10.1016/j.eswa.2021.115387.

[21] D. Sergi and I. Ucal Sari, 'Prioritization of public services for digitalization using fuzzy Z-AHP and fuzzy Z-WASPAS', *Complex Intell. Syst.*, vol. 7, no. 2, pp. 841–856, Apr. 2021, doi: 10.1007/s40747-020-00239-z.

[22] M. Şahin, 'Location selection by multi-criteria decision-making methods based on objective and subjective weightings', *Knowl. Inf. Syst.*, vol. 63, no. 8, pp. 1991–2021, Aug. 2021, doi: 10.1007/s10115-021-01588-y.

[23] I. Mukhametzyanov, 'Specific character of objective methods for determining weights of criteria in MCDM problems: Entropy, CRITIC and SD', *Decis. Mak. Appl. Manag. Eng.*, vol. 4, no. 2, pp. 76–105, June 2021, doi: 10.31181/dmame210402076i.

[24] T.-C. Wang and H.-D. Lee, 'Developing a fuzzy TOPSIS approach based on subjective weights and objective weights', *Expert Syst. Appl.*, vol. 36, no. 5, pp. 8980–8985, July 2009, doi: 10.1016/j.eswa.2008.11.035.

[25] H.-J. Shyur and H.-S. Shih, 'A hybrid MCDM model for strategic vendor selection', *Math. Comput. Model*, vol. 44, no. 7, pp. 749–761, Oct. 2006, doi: 10.1016/j.mcm.2005.04.018.

[26] J. M. Mittelstädt, J. Maier, P. Goerke, F. Zinn, and M. Hermes, 'Large language models can outperform humans in social situational judgments', *Sci. Rep.*, vol. 14, no. 1, p. 27449, Nov. 2024, doi: 10.1038/s41598-024-79048-0.

[27] A. Alam *et al.*, 'The Role of Artificial Intelligence in Pharmacy Practice and Patient Care: Innovations and Implications', *BioMedInformatics*, vol. 5, no. 4, Nov. 2025, doi: 10.3390/biomedinformatics5040065.

[28] M. S. Baysan, S. Uysal, İ. İşlek, Ç. Çığ Karaman, and T. Güngör, 'LLM-as-a-Judge: automated evaluation of search query parsing using large language models', *Front. Big Data*, vol. 8, July 2025, doi: 10.3389/fdata.2025.1611389.

[29] T. Brown *et al.*, 'Language Models are Few-Shot Learners', *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 1877–1901, 2020.

[30] R. W. Saaty, 'The Analytic Hierarchy Process—What It Is and How It Is Used', *Math. Model.*, vol. 9, no. 3, pp. 161–176, Jan. 1987, doi: 10.1016/0270-0255(87)90473-8.