Article

Not peer-reviewed version

A Multi-Criteria Decision Framework for Enterprise LLM Routing

Marcin Nowak *

Posted Date: 14 April 2026

doi: 10.20944/preprints202604.0905.v1

Keywords: large language models; artificial intelligence in the enterprise; prompt routing; multicriteria decision support; AHP; SAW

Article

A Multi-Criteria Decision Framework for Enterprise LLM Routing

Marcin Nowak

Poznan University of Technology, Faculty of Engineering Management; marcin.nowak@put.poznan.pl

Abstract

The increasing use of large language models (LLMs) in enterprises creates a need for the effective selection between lower-cost models and more advanced ones. The aim of the article is to propose a multicriteria decision-making framework for prompt routing to LLMs in an enterprise environment, taking into account organizational preferences regarding cost, response quality, business risk, response time, standardization, and creativity. The study adopts a design-and-evaluation approach. In the design phase, a mechanism was developed in which prompts are assessed according to managerial routing criteria, weighted using the AHP method, and then directed to either a lower-cost or a more powerful model using the SAW method. In the evaluation phase, the solution was tested on a dataset of 100 business prompts and compared with two benchmark strategies: always cheap and always strong. The article's contribution includes framing LLM routing as a managerial decision-support problem, operationalizing managerial routing criteria, and proposing evaluation metrics such as sufficiency rate, average cost per prompt, cost per sufficient response, and incremental cost of sufficiency gain. The results indicate that the proposed solution improves the cost–quality trade-off, while maintaining an acceptable level of response sufficiency and limiting the cost of query handling.

Keywords: large language models; artificial intelligence in the enterprise; prompt routing; multicriteria decision support; AHP; SAW

1. Introduction

The dynamic development of large language models (LLMs) is transforming the way organizations carry out knowledge-based work, communication, information analysis, and decision-support processes. In enterprise environments, these models are increasingly used not only for content generation, but also for document summarization, the preparation of managerial reports, customer service support, the drafting of business communication, the analysis of descriptive data, and the performance of tasks requiring synthesis and reasoning. Consequently, the practical problem is no longer limited to the question of whether organizations should use LLMs, but rather concerns how they can use them effectively, responsibly, and economically across different types of business tasks. In particular, enterprises face the need to determine when a less expensive and faster model is sufficient, and when the use of a more advanced, more costly model is justified because it may offer higher-quality responses.

The importance of this problem is growing alongside the expansion of the range of available models and the increasing differentiation of their parameters in terms of cost, response time, reasoning capability, and the quality of generated outputs. In business practice, users rarely operate under conditions in which only a single model is available. Much more often, they have access to a portfolio of models among which choices must be made depending on the nature of the query. This creates a need for prompt-routing mechanisms, understood as rules for directing queries to different models in a manner consistent with organizational priorities. These priorities are not exclusively technical in nature. They encompass trade-offs among response quality, processing cost, speed of

 doi:10.20944/preprints202604.0905.v1

operation, the level of error risk, the need for standardization, and the business consequences of an inappropriate response.

Existing approaches to the problem of LLM model selection and routing, however, remain dominated by a technical perspective. The primary focus is most often placed on quality benchmarks, computational efficiency, latency reduction, inference cost optimization, or improvements in system parameters. Although these aspects are important, they do not fully reflect the realities of managing the use of artificial intelligence in organizations. Enterprises do not assess model outputs solely through the prism of average technical quality, but also from the standpoint of their operational usefulness, acceptable level of risk, compliance with organizational requirements, and the economic justification of deployment. From a managerial perspective, the key question therefore becomes not so much which model is objectively the best, but rather which model is more appropriate for a specific task under a given profile of organizational preferences.

At this point, a research gap becomes apparent. Despite the growing interest in the use of artificial intelligence in organizations, relatively little attention has still been devoted to LLM routing as a problem of multicriteria decision support. In particular, there is a lack of analytical frameworks that would translate enterprise preferences into formal model-selection rules and integrate business criteria with decision-making methods. The number of studies proposing routing evaluation metrics from a managerial perspective also remains limited, that is, metrics that make it possible to analyze not only the quality of responses itself, but also the relationship between cost and the business sufficiency of the obtained results. Under organizational conditions, the fundamental issue is not whether a model generates the best possible response in absolute terms, but whether the response is sufficient for practical application without incurring unjustified costs.

The aim of this article is to propose and empirically evaluate a multicriteria decision-making framework for prompt routing to LLMs in an enterprise environment. In the proposed approach, model selection depends not only on the nature of the task itself, but also on the organization's preferences regarding cost, quality, risk, response time, standardization, and creativity. This approach assumes that a prompt can be assessed according to a set of managerially relevant criteria and then directed either to a less expensive model or to a more advanced one using a formal decision-making procedure.

The study adopts a design-and-evaluation approach. In the design phase, a routing mechanism was developed based on prompt assessment, criterion weights established by organizational experts, and the Simple Additive Weighting (SAW) method as the core of the multicriteria procedure. In the evaluation phase, the proposed solution was tested on a dataset of 100 business prompts and compared with two benchmark strategies: the fixed selection of a less expensive model and the fixed selection of a stronger model. The evaluation of the results was based on expert verification of the sufficiency of responses for organizational use and on a set of managerial routing metrics relating to the trade-off between cost and effectiveness.

The article's contribution to the literature has several dimensions. First, it conceptualizes LLM routing as a managerial and decision-making problem rather than an exclusively technical one. Second, it systematizes and operationalizes a set of managerial routing criteria that may be used in the assessment of prompts in an enterprise setting. Third, it demonstrates the applicability of multicriteria decision support methods to the allocation of queries among LLMs under organizational conditions. Fourth, it proposes a set of metrics enabling the evaluation of routing effectiveness from the perspective of cost and the business sufficiency of responses, thus in a manner more closely aligned with management practice than classical comparisons of model quality. In this way, the article contributes to the growing stream of research on the use of artificial intelligence in organizations by introducing the perspectives of governance, economic efficiency, and the alignment of AI tools with enterprise needs.

The remainder of the article is structured as follows. The next section presents a review of the literature on the use of LLMs in organizations, model routing, and multicriteria decision support. This is followed by a presentation of the research methodology and the construction of the proposed

 doi:10.20944/preprints202604.0905.v1

decision model. The subsequent section discusses the results of the empirical study, while the conclusion offers a discussion of the findings, managerial implications, study limitations, and directions for future research.

2. Literature Review

Large language models (LLMs) are currently conceptualized in the literature as the next phase in the development of AI systems capable of performing a wide range of linguistic, analytical, and creative tasks, with their significance extending beyond classical NLP applications and increasingly encompassing organizational and managerial contexts [1,2]. Studies on the business value of AI emphasize that the impact of these technologies on firm performance does not arise solely from their technical parameters, but from their capacity to reconfigure processes, support decision-making, and enhance organizational efficiency [3–5]. Recent management research also shows that AI-based decision-support tools can be deployed in specific business functions such as predicting voluntary employee turnover [6]. More recent literature reviews also stress that generative AI is becoming an important component of business model innovation, value creation, and organizational transformation, while at the same time giving rise to challenges related to implementation, governance, and the assessment of usefulness in business practice [7–9]. At the level of concrete managerial applications, AI-based systems are also being developed for revenue-oriented decisions such as dynamic pricing in e-commerce [10]. Similarly, machine learning has been proposed as a quantitative tool for organizational diagnosis, including the prediction of dominant organizational culture types [11].

The significance of this issue is reinforced by empirical studies showing that generative AI can improve the productivity of knowledge workers, although these effects vary markedly across tasks and user groups. Brynjolfsson, Li, and Raymond [12] demonstrated that the use of generative AI in customer service increases productivity by approximately 15% on average, whereas Noy and Zhang [13] identified significant gains in both productivity and quality in writing tasks. In turn, Dell'Acqua et al. [14] describe the phenomenon of the "jagged technological frontier," indicating that AI may improve performance in some tasks while worsening it in others, even when those tasks appear superficially similar. It is precisely this unevenness of effects that is particularly important from the enterprise perspective: it suggests that not every prompt and not every task should be handled by the same model or by the same AI usage strategy [12].

In parallel, a stream of research has been developing that emphasizes the importance of governance, accountability, and the organizational embeddedness of AI systems. Papagiannidis, Mikalef, and Conboy [15] argue that responsible AI governance requires not only a set of general principles, but also procedural, relational, and structural practices that make it possible to operationalize oversight over the technology. Schneider, Kuss, Abraham, and Meske [16], in turn, note that generative AI has its own distinct risk profile and specific implications for enterprise governance that cannot be fully explained by traditional AI governance frameworks. In a similar vein, Leoni et al. [17] show that the adoption of AI in knowledge management processes affects decision-making processes, while Yan, Husted, and Fath [18] point to the importance of AI and GenAI for organizational learning and knowledge creation. This literature leads to the conclusion that the use of LLMs in firms should be analyzed not only as a technological issue, but also as a problem of organizational fit, risk management, and decision effectiveness.

Vidgof, Bachhofner, and Mendling [19] were among the first to systematically describe the opportunities and challenges of using LLMs in BPM, indicating that the potential of these models spans many stages of the process lifecycle, while simultaneously requiring new rules of application. Subsequent studies have presented more specific applications: Bernardi et al. [20] proposed the BPLLM framework for process-aware decision support; Kourani et al. [21] developed an approach for generating process models from textual descriptions; Apaydin and Zisgen [22] investigated the use of local language models for process modeling; and Kourani et al. [23] introduced a benchmark and self-improvement analysis of models in process-modeling tasks. Kampik et al. [24], in turn,

 doi:10.20944/preprints202604.0905.v1

formulated a broader vision of "large process models," in which LLMs are intended to support contextual modeling and the improvement of business processes. The common denominator of these studies is clear: the effectiveness of LLMs in enterprise applications is task-dependent, and model selection should be linked to the type of task, the required level of reliability, and the organizational context of use.

The literature closest to the problem addressed in this article concerns the routing and cascading of language models. Dohan et al. [25] proposed the general concept of "language model cascades," showing that combining models can be treated as a formal probabilistic problem. Chen, Zaharia, and Zou [26], in their work on FrugalGPT, demonstrated that an appropriate composition of models can significantly improve the cost–quality trade-off. Sakota, Peyrard, and West [27] proposed cost-efficient model selection using a meta-model; Hu et al. [28] developed RouterBench as a benchmark for multi-LLM routing systems; and Ong et al. [29] showed that routing trained on preference data can reduce costs without significant loss in response quality. Similarly, Mohammadshahi, Shaikh, and Yazdani [30], Dekoninck, Baader, and Vechev [31], Shirkavand et al. [32], and Yue et al. [33] develop approaches to dynamic model selection, cascading, and cost-aware routing. This literature is highly valuable, but its dominant perspective remains technical or economic optimization, focused primarily on response quality, inference cost, and sometimes latency. Much less developed, however, is an approach in which routing would reflect the managerial preferences of the enterprise, such as the acceptable level of error risk, the requirement for response standardization, or the business consequences of an inappropriate output.

At this point, multicriteria decision support methods become particularly useful [34]. The classic works of Saaty [35,36] and Hwang and Yoon [37] laid the foundations for the AHP and SAW methods, which make it possible to translate decision-makers' preferences into a formal rule for the evaluation and selection of alternatives. Later studies confirm that SAW remains a transparent, interpretable, and convenient method in situations requiring the aggregation of multiple criteria, whereas AHP is a useful tool for determining criterion weights on the basis of expert judgments [38,39]. Importantly, MCDM methods have already been applied to the selection of AI-based systems, including chatbots for customer service, demonstrating that decisions concerning the choice of AI tools can be effectively formalized as multicriteria problems [40]. Related management research also shows that formal decision models can integrate measurable and qualitative factors through expert knowledge and explicit weighting procedures under uncertainty [41]. However, the available literature still lacks convincing attempts to apply AHP and SAW to prompt routing between LLMs in enterprise environments while taking into account criteria of managerial significance rather than exclusively technical ones.

In summary, the literature is currently developing along three related but still weakly integrated streams: research on the use of LLMs and GenAI in organizations, research on model routing and cascading, and research on governance and multicriteria decision support. The first stream demonstrates the growing importance of LLMs for productivity, knowledge creation, and business transformation; the second provides techniques for improving the cost–quality trade-off; and the third offers tools for formalizing organizational preferences. However, what remains insufficiently developed is an approach that would integrate these three perspectives and conceptualize LLM routing as a managerial problem, in which the decision regarding model selection depends on cost, quality, risk, response time, and the requirement for standardization, while the effectiveness of the solution is assessed through the lens of the business adequacy of the response. It is precisely this gap that the present article seeks to address.

3. Materials and Methods

The study is design-and-evaluation in nature. Its objective is to develop and empirically validate a decision-making mechanism that supports LLM selection in an enterprise environment, under the assumption that the organization seeks simultaneously to control model usage costs, maintain the

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2026 doi:10.20944/preprints202604.0905.v1

expected quality of responses, and align model choice with its own managerial and operational preferences. The proposed mechanism employs:

1. a set of business-relevant decision criteria,
2. criterion weights established by organizational experts,
3. prompt evaluation by a classification model,
4. the SAW multicriteria decision-support method,
5. routing of the query to either a less expensive or a more capable model.

From a managerial perspective, the study does not address the question, "Which model is objectively the best?" but rather, "Which model is more appropriate for a given organization and for a given type of task under a specific profile of business preferences?"

The developed model may be presented as a procedure consisting of seven stages.

Stage 1. Identification of managerial decision criteria for LLM model selection

For prompt evaluation, the following set of decision criteria is proposed:

- C1—required substantive accuracy (describes how high the correctness and precision of the response must be for the outcome to be business-useful; the greater the required accuracy, the greater the need to use a more advanced model),
- C2—risk of the business consequences of error (describes the potential effects of generating an incomplete, misleading, or incorrect response; an error in a draft marketing post has different significance than an error in a compliance analysis, HR policy, communication with a strategic client, or the interpretation of a document),
- C3—required depth of reasoning (describes whether the task requires simple information processing or multi-step reasoning, synthesis, and logical analysis; the greater the depth of reasoning required, the greater the likelihood that the organization will prefer a more powerful model),
- C4—sensitivity to processing cost (describes how important cost savings are from the company's perspective for a given type of query; not every task requires maximizing quality at any cost. In many large-scale processes, unit cost is the priority),
- C5—task sensitivity to response time (describes how important it is to obtain the result quickly; in operational, contact-intensive, or high-volume tasks, speed may be just as important as quality),
- C6—required standardization and compliance of the response (describes whether the response must strictly conform to the adopted style, structure, company policy, or communication standard; in organizations, a large part of AI's value derives not from creativity, but from repeatability, consistency, and the scalability of communication),
- C7—required creativity/openness of generation (describes the extent to which the task requires a creative, non-standard, or exploratory approach).

The indicated set of criteria combined four managerial logics: cost efficiency, risk control, quality of the decision-making process, and operational effectiveness.

Stage 2. Determination of criterion weights by organizational experts

To translate organizational preferences into a formal decision model, the Analytic Hierarchy Process (AHP) method was applied. This method makes it possible to determine criterion weights on the basis of pairwise comparisons made by experts. In the first step, each expert compares every criterion with every other criterion by answering the question of which of them is more important from the organization's perspective and to what extent. The comparisons were conducted using the classical nine-point Saaty scale, in which: 1 denotes equal importance of both criteria, 3 a moderate preference of one criterion over the other, 5 a strong preference, 7 a very strong preference, and 9 an extreme preference; the values 2, 4, 6, and 8 represent intermediate judgments. For the k-th expert, a pairwise comparison matrix is constructed:

$$\mathbf{A}^{(k)} = \left[a_{ij}^{(k)}\right]_{n \times n} \tag{1}$$

 doi:10.20944/preprints202604.0905.v1

where $a_{ij}^{(k)}$ denotes the relative importance of criterion C_i with respect to criterion C_j, subject to the following conditions:

$$a_{ii}^{(k)} = 1, \qquad a_{ij}^{(k)} > 0, \qquad a_{ji}^{(k)} = \frac{1}{a_{ij}^{(k)}}. \tag{2}$$

Because the assessments were provided by several experts, it was necessary to aggregate them into a single group matrix. For this purpose, the geometric mean of the experts' judgments was applied, which is the standard solution in the group version of AHP. For each matrix element, the following was adopted:

$$a_{ij} = (\prod_{k=1}^{m} a_{ij}^{(k)})^{\frac{1}{m}} \tag{3}$$

where m denotes the number of experts. In the present study, $m = 3$ was adopted. As a result, a group comparison matrix was obtained:

$$\mathbf{A} = [a_{ij}]_{n\times n} \tag{4}$$

Based on matrix $\mathbf{A}$, the vector of criterion weights was determined. In computational practice, the method of the normalized geometric mean of rows was applied. First, for each criterion, the geometric mean of the judgments in the row was calculated:

$$g_i = (\prod_{j=1}^{n} a_{ij})^{\frac{1}{n}} \tag{5}$$

and the obtained values were then normalized, yielding the weight of each criterion:

$$w_i = \frac{g_i}{\sum_{i=1}^{n} g_i}, i = 1,2, \dots, n \tag{6}$$

The weight vector may therefore be written as:

$$\mathbf{w} = (w_1, w_2, \dots, w_n) \tag{7}$$

subject to the normalization condition:

$$\sum_{i=1}^{n} w_i = 1 \tag{8}$$

The obtained weights reflect the relative importance of the individual criteria from the organization's perspective. The higher the value of w_i, the greater the influence of the given criterion on the subsequent decision regarding the selection of the LLM. An important element of the AHP method is also the assessment of the consistency of expert judgments. For this purpose, the maximum eigenvalue of the comparison matrix was calculated:

$$\lambda_{\max} = \frac{1}{n}\sum_{i=1}^{n} \frac{(\mathbf{Aw})_i}{w_i} \tag{9}$$

On this basis, the consistency index was determined:

$$CI = \frac{\lambda_{\max} - n}{n - 1} \tag{10}$$

followed by the consistency ratio:

$$CR = \frac{CI}{RI} \tag{11}$$

where RI denotes the Random Index, which depends on the number of criteria analyzed. For $n = 7$, the value $RI = 1.32$ is typically adopted. In the literature, pairwise comparisons are considered sufficiently consistent when $CR < 0.10$. If the value of CR exceeds the threshold of 0.10, this indicates that the expert judgments are characterized by excessive inconsistency and should be re-examined. In the present study, the resulting weight vector was subsequently used in the next stage

 doi:10.20944/preprints202604.0905.v1

of the procedure, namely in the SAW method employed for routing prompts between the less expensive model and the more capable model.

Stage 3. Development of the research prompt dataset

At this stage, a diversified database of prompts should be constructed to reflect real-world enterprise applications. The database may comprise approximately 100 prompts. This set should be heterogeneous, that is, it should reflect different classes of business problems, such as: information analysis and synthesis, summarization and reporting, business communication and content drafting, decision-making and analytical tasks, tasks involving high business risk, procedural and standardized tasks, as well as creative and conceptual tasks. This stage results in the creation of a research benchmark representing the space of typical organizational tasks.

Stage 4. Evaluation of prompts by the classification model

At this stage, scores are assigned to each prompt within the specified decision criteria. In this case, the less expensive language model serves as the evaluating decision agent. The agent analyzes each prompt and assigns values to it for each criterion according to the adopted scale. For the sake of transparency and reproducibility, the full system prompt and the response scheme used by the evaluating agent are presented in Appendix A. The following scale is proposed:

C1—Accuracy

To what extent does the required accuracy of the response exceed the typical capabilities of the less expensive model?

1—the less expensive model will probably be fully sufficient
2—the less expensive model should be sufficient with only minor risk
3—borderline task, with no clear advantage
4—higher accuracy clearly favors the more expensive model
5—the required accuracy definitely justifies the more expensive model

C2—Business_Risk

To what extent does the potential error justify escalation to the more expensive model?

1—a possible error has little business significance
2—an error would be undesirable, but not critical
3—an error would have moderate significance
4—an error could have serious consequences
5—an error definitely justifies cautious escalation

C3—Reasoning_Depth

To what extent does the task require reasoning beyond the typical capabilities of the less expensive model?

1—a simple, routine, or template-based task
2—minor analysis or organization of information
3—moderate reasoning
4—complex, multi-step reasoning
5—deep reasoning definitely justifies the more expensive model

C4—Cost_Sensitivity

How important is it to complete the task at the lowest possible cost?

1—cost has little significance
2—cost has minor significance
3—cost has moderate significance
4—cost is important and favors the less expensive model
5—cost is very important and strongly favors the less expensive model

C5—Time_Sensitivity

How important is it to obtain a rapid response?

1—response time has little significance
2—time has minor significance
3—time has moderate significance

 doi:10.20944/preprints202604.0905.v1

4—a rapid response is important
5—response speed strongly favors the less expensive model
C6—Standardization
To what extent is the task template-based, predictable, and grounded in a standard response structure?
1—the response requires a non-standard approach
2—the response is rather non-standard
3—the response is partially standardized
4—the response is highly template-based
5—the response is strongly standardized and predictable
C7—Creativity
To what extent does the task require a creative, conceptual, or non-standard approach?
1—creativity is not needed
2—a small degree of creativity may help
3—moderate creativity
4—high creativity is useful
5—the task clearly requires a more creative model

In this way, for each prompt, a score vector is obtained: $P_i = (p_{i1}, p_{i2}, \dots, p_{i7})$.

Stage 5. Routing of prompts to one of the analyzed LLM models using the SAW method

To assign each prompt to either the less expensive model or the more capable model, the Simple Additive Weighting (SAW) method was applied. This method consists in calculating the total weighted score of each decision alternative and then selecting the alternative with the higher final value. In the present study, the alternatives are two LLMs: the less expensive model and the more capable model. For each prompt P_i, the score vector obtained in Stage 4 is known:

$$\mathbf{p}_i = (p_{i1}, p_{i2}, \dots, p_{i7}) \tag{12}$$

as well as the criterion weight vector determined using the AHP method in Stage 2:

$$\mathbf{w} = (w_1, w_2, \dots, w_7), \qquad \sum_{j=1}^{7} w_j = 1. \tag{13}$$

It was assumed that high values of criteria C1, C2, C3, and C7 favor the selection of the more capable model, whereas high values of criteria C4, C5, and C6 favor the selection of the less expensive model. To maintain a uniform evaluation logic, two synthetic scores were calculated for each prompt: one for the less expensive model and one for the more capable model. The score for the more capable model was determined as:

$$S_i^{(strong)} = w_1 p_{i1} + w_2 p_{i2} + w_3 p_{i3} + w_7 p_{i7} + w_4(6 - p_{i4}) + w_5(6 - p_{i5}) + w_6(6 - p_{i6}), \tag{14}$$

whereas the score for the less expensive model was determined as:

$$S_i^{(cheap)} = w_4 p_{i4} + w_5 p_{i5} + w_6 p_{i6} + w_1(6 - p_{i1}) + w_2(6 - p_{i2}) + w_3(6 - p_{i3}) + w_7(6 - p_{i7}). \tag{15}$$

Such a construction means that criteria favoring a given model increase its score, whereas criteria supporting the alternative model are reversed according to the transformation $(6 - p_{ij})$. As a result, both alternatives can be evaluated within the same aggregation logic. The final decision rule is as follows:

$$d_i = \begin{cases} \text{cheap}, & \text{if } S_i^{(cheap)} \geq S_i^{(strong)}, \\ \text{strong}, & \text{if } S_i^{(strong)} > S_i^{(cheap)}. \end{cases} \tag{16}$$

Additionally, the difference between the scores of the two models may be calculated:

$$\Delta_i = S_i^{(strong)} - S_i^{(cheap)}. \tag{17}$$

If $\Delta_i > 0$, the prompt is routed to the more capable model, whereas if $\Delta_i < 0$, it is routed to the less expensive model. The greater the absolute value of this difference, the stronger the justification for the routing decision. The applied rule thus makes it possible to link the choice of the LLM to a formally specified profile of organizational preferences.

Stage 6. Generation of responses by the set of analyzed reference models

At this stage of the model, responses are generated for all prompts. For each prompt, two responses are generated—one by the less expensive model and one by the more expensive (more capable) model. From these responses, three response vectors are constructed for the compared models, namely:

- a vector of responses originating exclusively from the less expensive model (always cheap strategy),
- a vector of responses originating exclusively from the more expensive model (always strong strategy),
- a vector of responses resulting from prompt routing using the multicriteria model (SAW-based routing).

Costs in the analysis were estimated on the basis of average unit costs derived from the API pricing schedule for the models used, rather than on the basis of actual token consumption in each individual call. Specifically, the cost estimates were based on fixed average per-prompt cost assumptions for each model, intended to provide a comparative approximation of strategy-level expenditure rather than an exact reconstruction of API billing for individual requests.

Stage 7. Evaluation of the always cheap, always strong, and SAW-based routing models

At this stage, it is assessed whether the proposed prompt-routing method offers an advantage over simpler operating strategies. In other words, the purpose of this stage is to determine whether the router is able, with sufficient frequency, to direct simpler/less expensive tasks to the less expensive model and more difficult/sensitive tasks to the more expensive one, so that the company does not overpay while also not sacrificing quality where it matters. The evaluation of model responses should be qualitative and utility-oriented, focusing on whether the generated output meets organizational expectations for a given type of task. Experts from the enterprise are engaged in the evaluation of responses (as users, they are best positioned to assess the responses and compare them with their expectations). At this point, the experts do not replicate the activity of the prompt-evaluating agent and do not assign attributes such as "task complexity" or "prompt error risk" to the responses. Their task is to evaluate the final product, that is, the model's response itself. For each response, the experts answer a single question: Is the response acceptable for use in organizational conditions without significant substantive revisions? Two answers are possible—yes or no. Each response was evaluated independently by three experts. The final classification of a response as sufficient or non-sufficient was established according to the majority-vote rule. This means that a response was considered sufficient if at least two out of three experts assessed it as acceptable for use in organizational conditions without significant substantive revisions. The introduction of this variable makes it possible to avoid the erroneous assumption that the more expensive model should always be selected simply because it generates, on average, higher-quality responses. From the enterprise perspective, what is crucial is not which model is "the best," but which model is sufficient and economically justified for a given type of task.

For each of the three models, the following indicators are proposed:

1. Sufficiency Rate—this is the basic effectiveness indicator of the compared models

$$SR = \frac{\text{number of sufficient responses}}{\text{number of all responses}} \tag{18}$$

2. Average Cost per Prompt (total strategy cost)—this is the indicator of the average cost of handling a single query

$$ACP = \frac{\text{total strategy cost}}{\text{number of prompts}} \tag{19}$$

3. Cost per Sufficient Response—an indicator referring to how much the organization pays on average to obtain one sufficient response. This is a key indicator in the context of comparing the analyzed models

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2026 doi:10.20944/preprints202604.0905.v1

$$CSR = \frac{\text{total strategy cost}}{\text{number of sufficient responses}} \tag{20}$$

A probable scenario to be evaluated in the empirical part of the article may be that the less expensive model, which by assumption will be characterized by low costs, may also have a substantial share of insufficient responses. The more expensive model, by assumption, will have a substantial share of sufficient responses, but at high cost. The routing model, in turn, should achieve the best value of the sufficient-response indicator relative to cost, which would make it the model with the highest efficiency.

4. Incremental Cost of Sufficiency Gain—the incremental cost of improving sufficiency (e.g., relative to cheap-only)

$$ICSG = \frac{Cost_{routing} - Cost_{cheap}}{SR^{pp}_{routing} - SR^{pp}_{cheap}} \tag{21}$$

This indicator shows the additional cost that must be incurred in order to achieve an increase of 1 percentage point in the share of sufficient responses.

4. Results

Research using the developed model was conducted in a selected Polish enterprise during the period from 2 March 2026 to 6 March 2026. In the present empirical evaluation, the categories of cheap model and strong model were operationalized as GPT 4o-mini and GPT 5.2, respectively. However, the adopted terminology is of a general nature, since the proposed method may also be applied to other pairs of models differing in cost and capabilities. The study was carried out in the following seven successive stages:

Stage 1. Identification of managerial decision criteria for LLM model selection

For prompt evaluation, the set of criteria indicated in the model description was used, comprising:

- C1—required substantive accuracy,
- C2—risk of the business consequences of error,
- C3—required depth of reasoning,
- C4—sensitivity to processing cost,
- C5—task sensitivity to response time,
- C6—required standardization and compliance of the response,
- C7—required creativity/openness of generation.

The indicated set of criteria combined four managerial logics: cost efficiency, risk control, quality of the decision-making process, and operational effectiveness.

Stage 2. Determination of criterion weights by organizational experts

Three experts from the company in which the study was conducted participated in determining the weights of the decision criteria: an operations manager, the CEO, and an AI implementation manager. Table 1 presents the criterion weights determined in accordance with the AHP method (following the procedure presented in the methodological section).

Table 1. Weights of the decision criteria determined using the AHP method.

Criterion	Weight	Value [%]
C1 Accuracy	0.2591	25.9078
C4 Cost_Sensitivity	0.2086	20.8578
C5 Time_Sensitivity	0.1894	18.9366
C2 Business_Risk	0.1337	13.3739
C6 Standardization	0.0922	9.2207
C3 Reasoning_Depth	0.0874	8.7351
C7 Creativity	0.0297	2.9680

 doi:10.20944/preprints202604.0905.v1

For the aggregated group comparison matrix, the consistency ratio obtained was $CR = 0.0208$ (with individual expert values ranging from 0.0217 to 0.0269), which indicates that the judgments were highly consistent and clearly satisfied the AHP acceptability criterion of $CR < 0.10$.

Stage 3. Development of the research prompt dataset

At this stage, a database consisting of 100 prompts was developed. Three example prompts are presented in Table 2.

Table 2. Example prompts from the prepared database.

Prompt No.	Prompt Content
1	Read the following excerpt from a quarterly report and prepare a 5-point summary for the Chief Operating Officer. Focus exclusively on the key conclusions regarding sales, costs, and profitability. Do not exceed 150 words. "In the second quarter of 2025, the company's revenues amounted to PLN 18.4 million and were 6.8% higher than in the corresponding period of the previous year. Sales growth was recorded primarily in the subscription services segment, whose share in the revenue structure increased from 34% to 41%. At the same time, operating costs increased by 11.9%, mainly as a result of higher wages, energy costs, and expenditures on customer acquisition campaigns. Gross margin declined from 29.7% to 26.1%, while net profitability fell from 11.4% to 8.9%. The Management Board indicated that the current level of marketing costs is temporary in nature and should translate into sales growth over the next two quarters."
2	Write a professional email message to a strategic client explaining a two-week delay in the implementation of the system. The message should be polite, specific, and reassuring. It should include: the reason for the delay, a corrective action plan, and a proposal for a brief status meeting. Situation: the implementation of a reporting platform for a client from the logistics sector was scheduled to be completed by 12 May, but during testing a problem was detected with the integration of data from the warehouse management system. The technical team has prepared a fix, and the new plan assumes completion of the implementation by 26 May. The client is important to the company and expects regular communication.
3	Propose 5 ideas for the main slogan of a campaign promoting a new advisory service for small and medium-sized enterprises. The slogans should be modern, professional, and emphasize time savings as well as better business decisions. Then, for each slogan, add one short sentence explaining its meaning. Service description: the company offers subscription-based online business advisory services, including sales data analysis, operational consultations, and recommendations for process improvements in small and medium-sized enterprises.

Stage 4. Evaluation of prompts by the classification model

At this stage of the model, the decision criteria (weighted in the second stage of the procedure) are evaluated by the LLM agent, namely the less expensive model (in this case, the cheap model), for each of the 100 analyzed prompts. Table 3 presents the evaluation matrix for a sample of prompts assessed by the evaluating agent.

Table 3. Prompt evaluation matrix produced by the agent (cheap model).

prompt	C1	C2	C3	C4	C5	C6	C7
1	5	4	3	2	4	3	1
2	5	5	4	2	4	5	2
3	4	4	4	5	3	3	2
…	…						

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2026 doi:10.20944/preprints202604.0905.v1

99	5	5	5	3	4	4	2
100	4	3	2	1	4	3	1

Stage 5. Routing of prompts to one of the analyzed LLM models using the SAW method

Based on the previously determined weights of the decision criteria and the evaluations of the individual prompts by the LLM agent, the SAW decision method is used to determine to which of the LLM models (the less expensive or the more expensive one) each prompt should be routed. Example results of this assignment, together with the values that contributed to it, are presented in Table 4.

Table 4. Summary of example prompt-routing results.

Prompt id	Cheap score	Strong score	Score gap	Selected model
1	2.389	3.611	1.222	strong
2	2.322	3.678	1.356	strong
3	2.966	3.034	0.068	strong
4	3.883	2.117	-1.766	cheap
5	3.572	2.428	-1.145	cheap

Across the entire research dataset, the routing mechanism directed 40 prompts to the cheap model and 60 prompts to the strong model. This structure shows that the resulting allocation was not the effect of simply imitating one of the baseline strategies, but rather the result of a selective allocation of prompts between models differing in cost and capabilities.

Stage 6. Generation of responses by the set of analyzed reference models

At this stage of the model, responses were generated for all prompts. For each prompt, two responses were generated—one by the less expensive model and one by the more expensive model. From these responses, three response vectors were constructed for the compared models, namely:

- a vector of responses originating exclusively from the less expensive model (always cheap strategy),
- a vector of responses originating exclusively from the more expensive model (always strong strategy),
- a vector of responses resulting from prompt routing using the multicriteria model (SAW-based routing).

Stage 7. Evaluation of the always cheap, always strong, and SAW-based routing models

At this stage, the responses generated by the three models—always cheap, always strong, and SAW-based routing—were compared for each of the 100 prompts.

In accordance with the adopted methodology, the sufficiency of the responses generated by the models was evaluated. Response sufficiency was likewise assessed by a team of three experts from the studied organization. An example of the evaluation of the sufficiency of model responses for selected prompts is presented in Table 5.

Table 5. Evaluation of model responses for selected prompts.

Prompt id	Model cheap-only	Model strong-only	Model routing (selected)
1	sufficient	sufficient	sufficient (strong)
2	non-sufficient	sufficient	sufficient (strong)
3	sufficient	sufficient	sufficient (strong)
4	sufficient	sufficient	sufficient (cheap)
5	non-sufficient	non-sufficient	non-sufficient (cheap)

Table 6 presents the aggregate results for the indicators SR (Sufficiency Rate), ACP (Average Cost per Prompt), CSR (Cost per Sufficient Response), and ICSG (Incremental Cost of Sufficiency Gain).

 doi:10.20944/preprints202604.0905.v1

Table 6. Aggregate results for the SR, ACP, CSR, and ICSG indicators for the analyzed models.

Model	Number of sufficient responses	Number of sufficient responses (%) (SR)	Total cost (USD)
Cheap only	50	50%	0.8000
Strong only	85	85%	2.0000
Routing	85	85%	1.5200

Model	Average cost per prompt (ACP)	Cost per sufficient response (CSR)	Incremental Cost of Sufficiency Gain (vs cheap) (ICSG)
Cheap only	0.0080	0.0160	-
Strong only	0.0200	0.0235	0.0343
Routing	0.0152	0.0179	0.0206

The results indicate that the SAW-based routing strategy is not the absolutely best solution in terms of a single cost indicator, since the lowest cost is achieved by the cheap-only strategy. Its advantage becomes apparent, however, in a two-criterion perspective combining the level of response sufficiency with the cost of obtaining those responses. Routing achieved the same level of response sufficiency as the strong-only strategy (SR=85%), at a lower total cost (USD 1.5200 versus USD 2.0000), which means a cost reduction of 24.0% while maintaining the same effectiveness. At the same time, routing significantly outperformed the cheap-only strategy in terms of response quality (an increase in SR from 50% to 85%), offering a more favorable trade-off between cost and quality. This means that routing is the preferred solution in situations where the organization does not seek to minimize cost at any price, but rather to achieve high effectiveness at a rational level of expenditure. An additional argument in favor of routing is the lowest ICSG value among the strategies that improve effectiveness relative to the cheap-only variant. This means that each additional unit of improvement in response sufficiency was achieved at a lower cost under routing than under the strong-only strategy.

5. Conclusions

The conducted study indicates that the routing of LLMs in the enterprise environment may be effectively conceptualized not only as a problem of technical optimization, but also as a problem of multicriteria decision support embedded in real organizational preferences. The theoretical contribution of the article lies in integrating three perspectives that have thus far been only weakly connected, namely research on the use of LLMs in organizations, research on model routing, and the stream of multicriteria decision support, within a single coherent analytical model. The article additionally contributes a proposal for the operationalization of managerial routing criteria and a set of indicators for its evaluation, thereby extending the existing, predominantly technical approaches to the efficiency of multi-LLM systems. From a practical perspective, the proposed solution may constitute a useful tool supporting enterprise AI governance, as it enables a more transparent, auditable, and economically justified choice between a less expensive model and a more capable one, taking into account the trade-off between cost, response quality, error risk, response time, and the requirement for standardization. The limitations of the study, however, stem from its pilot character: the analysis was conducted in a single organization, on a set of 100 prompts, and with a comparison of two models, which limits the possibility of fully generalizing the conclusions. In addition, the assessment of response sufficiency was based on expert judgment and was therefore partly subjective and strongly dependent on the organizational context. In future research, it would be advisable to extend the analysis to a larger number of enterprises, industries, models, and use scenarios, as well as to incorporate dynamic calibration of criterion weights, training of the router on historical data,

 doi:10.20944/preprints202604.0905.v1

analysis of the stability of routing decisions, and comparisons with other MCDM methods and machine-learning-based approaches.

Institutional Review Board Statement: According to institutional and national regulations, ethical review and approval were not required for this study. The study involved the evaluation of business prompts and AI-generated responses by adult participants acting in an expert capacity. Participation was voluntary and anonymous, no personally identifiable or sensitive personal data were collected, and the study did not expose participants to physical, psychological, social, or legal risk.

Data Availability Statement: The prompt dataset, routing evaluations, and aggregated assessment results are available from the corresponding author upon reasonable request. The data are not publicly available due to organizational confidentiality considerations.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. System Prompt for Prompt Evaluation

system_prompt = “““
You are a routing agent in an enterprise AI system.
Your task is to assess whether a given prompt requires escalation to a more advanced model, or can be handled by a cheaper model.
Evaluate from the perspective of a routing decision. Important principles:

1. Do not inflate scores just because the task sounds professional.
2. Typical operational, communicative, and template-based tasks should NOT automatically receive high C1/C2/C3 scores.
3. High standardization (C6) → cheaper model.
4. High cost sensitivity (C4) and time sensitivity (C5) → cheaper model.
5. Assess the NEED for escalation, not the general “importance” of the task.

Scales (1–5):
C1_Accuracy—required accuracy vs. capabilities of the cheaper model
1=mini is fully sufficient | 5=strongly justifies a more expensive model
C2_Business_Risk—potential cost of error → escalation?
1=error has minor impact | 5=strongly justifies escalation
C3_Reasoning_Depth—depth of reasoning vs. cheaper model capabilities
1=simple/template-based | 5=deep, multi-step
C4_Cost_Sensitivity—importance of low processing cost
1=cost is not important | 5=cost strongly favors cheaper model
C5_Time_Sensitivity—importance of response speed
1=time is not important | 5=speed strongly favors cheaper model
C6_Standardization—degree of template-based, predictable output
1=non-standard | 5=highly standardized (→ cheaper model)
C7_Creativity—need for creative/non-standard approach
1=creativity unnecessary | 5=clearly requires a more expensive model
Guidelines:

- Simple summaries, emails, lists, messages → low C1/C2/C3, high C4/C5/C6
- Legal, strategic, high-risk tasks → higher C1/C2/C3
- Do not assign multiple criteria a score of 5 without a clear justification.

Return ONLY valid JSON. No markdown. No ```.
Format:
{
“C1_Accuracy”: {“score”: 1, “justification”: “...”},
“C2_Business_Risk”: {“score”: 1, “justification”: “...”},
“C3_Reasoning_Depth”: {“score”: 1, “justification”: “...”},
“C4_Cost_Sensitivity”:{“score”: 1, “justification”: “...”},
“C5_Time_Sensitivity”:{“score”: 1, “justification”: “...”},
“C6_Standardization”: {“score”: 1, “justification”: “...”},
“C7_Creativity”: {“score”: 1, “justification”: “...”},
“overall_comment”: “...”

 doi:10.20944/preprints202604.0905.v1


```
}
"""
human_prompt = "Evaluate the following prompt:\n\n{prompt_text}"
```

References

1. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223* **2023**, *1*, 1–124.
2. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A Comprehensive Overview of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2025**, *16*, 1–72.
3. Wamba-Taguimdje, S.-L.; Fosso Wamba, S.; Kala Kamdjoug, J.R.; Tchatchouang Wanko, C.E. Influence of Artificial Intelligence (AI) on Firm Performance: The Business Value of AI-Based Transformation Projects. *Business process management journal* **2020**, *26*, 1893–1924.
4. Sestino, A.; De Mauro, A. Leveraging Artificial Intelligence in Business: Implications, Applications and Methods. *Technol. Anal. Strateg. Manag.* **2022**, *34*, 16–29.
5. Le Dinh, T.; Vu, M.-C.; Tran, G.T.C. Artificial Intelligence in SMEs: Enhancing Business Functions through Technologies and Applications. *Information* **2025**, *16*, 415.
6. Nowak, M. Prediction of Voluntary Employee Turnover Using Machine Learning. *Zeszyty Naukowe. Organizacja i Zarzadzanie/Politechnika Śląska* **2024**, *201*.
7. Kanbach, D.K.; Heiduk, L.; Blueher, G.; Schreiter, M.; Lahmann, A. The GenAI Is out of the Bottle: Generative Artificial Intelligence from a Business Model Innovation Perspective. *Review of Managerial Science* **2024**, *18*, 1189–1220.
8. Romeo, E.; Lacko, J. Adoption and Integration of AI in Organizations: A Systematic Review of Challenges and Drivers towards Future Directions of Research. *Kybernetes* **2026**, *55*, 1286–1307.
9. Sánchez, M.A. Exploring Value Creation of Generative Artificial Intelligence in Organizations: A Systematic Review. *Strategic Business Research* **2025**, 100015.
10. Nowak, M.; Pawłowska-Nowak, M. Dynamic Pricing Method in the E-Commerce Industry Using Machine Learning. *Applied Sciences* **2024**, *14*, 11668.
11. Nowak, M.; Pawłowska-Nowak, M. Prediction of the Type of Organizational Culture Using Machine Learning Approach. *Przeglad Organizacji* **2023**, 264–272.
12. Brynjolfsson, E.; Li, D.; Raymond, L. Generative AI at Work. *Q. J. Econ.* **2025**, *140*, 889–942.
13. Noy, S.; Zhang, W. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *Science (1979).* **2023**, *381*, 187–192.
14. Dell'Acqua, F.; McFowland III, E.; Mollick, E.; Lifshitz, H.; Kellogg, K.C.; Rajendran, S.; Krayer, L.; Candelon, F.; Lakhani, K.R. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. *Organization Science* **2026**.
15. Papagiannidis, E.; Mikalef, P.; Conboy, K. Responsible Artificial Intelligence Governance: A Review and Research Framework. *The Journal of Strategic Information Systems* **2025**, *34*, 101885.
16. Schneider, J.; Kuss, P.; Abraham, R.; Meske, C. Governance of Generative Artificial Intelligence for Companies. *arXiv preprint arXiv:2403.08802* **2024**.
17. Leoni, L.; Gueli, G.; Ardolino, M.; Panizzon, M.; Gupta, S. AI-Empowered KM Processes for Decision-Making: Empirical Evidence from Worldwide Organisations. *Journal of Knowledge Management* **2024**, *28*, 320–347.
18. Yan, J.; Husted, K.; Fath, B. Transforming Organizational Knowledge Creation through Artificial Intelligence: A Systematic Review of the Emergent Literature. *VINE Journal of Information and Knowledge Management Systems* **2026**, *56*, 522–540.
19. Vidgof, M.; Bachhofner, S.; Mendling, J. Large Language Models for Business Process Management: Opportunities and Challenges. In Proceedings of the International conference on business process management; 2023; pp. 107–123.
20. Bernardi, M.L.; Casciani, A.; Cimitile, M.; Marrella, A. Conversing with Business Process-Aware Large Language Models: The BPLLM Framework. *J. Intell. Inf. Syst.* **2024**, *62*, 1607–1629.

 doi:10.20944/preprints202604.0905.v1

21. Kourani, H.; Berti, A.; Schuster, D.; van der Aalst, W.M.P. Process Modeling with Large Language Models. In Proceedings of the International Conference on Business Process Modeling, Development and Support; 2024; pp. 229–244.
22. Apaydin, K.; Zisgen, Y. Local Large Language Models for Business Process Modeling. In Proceedings of the International Conference on Process Mining; 2024; pp. 605–609.
23. Kourani, H.; Berti, A.; Schuster, D.; van der Aalst, W.M.P. Evaluating Large Language Models on Business Process Modeling: Framework, Benchmark, and Self-Improvement Analysis: H. Kourani et Al. *Softw. Syst. Model.* **2025**, 1–36.
24. Kampik, T.; Warmuth, C.; Rebmann, A.; Agam, R.; Egger, L.N.P.; Gerber, A.; Hoffart, J.; Kolk, J.; Herzig, P.; Decker, G.; et al. Large Process Models: A Vision for Business Process Management in the Age of Generative AI. *KI-Künstliche Intelligenz* **2025**, *39*, 81–95.
25. Dohan, D.; Xu, W.; Lewkowycz, A.; Austin, J.; Bieber, D.; Lopes, R.G.; Wu, Y.; Michalewski, H.; Saurous, R.A.; Sohl-Dickstein, J.; et al. Language Model Cascades. *arXiv preprint arXiv:2207.10342* **2022**.
26. Chen, L.; Zaharia, M.; Zou, J. Frugalgpt: How to Use Large Language Models While Reducing Cost and Improving Performance. *arXiv preprint arXiv:2305.05176* **2023**.
27. Šakota, M.; Peyrard, M.; West, R. Fly-Swat or Cannon? Cost-Effective Language Model Choice via Meta-Modeling. In Proceedings of the Proceedings of the 17th ACM International Conference on Web Search and Data Mining; 2024; pp. 606–615.
28. Hu, Q.J.; Bieker, J.; Li, X.; Jiang, N.; Keigwin, B.; Ranganath, G.; Keutzer, K.; Upadhyay, S.K. Routerbench: A Benchmark for Multi-Llm Routing System. *arXiv preprint arXiv:2403.12031* **2024**.
29. Ong, I.; Almahairi, A.; Wu, V.; Chiang, W.-L.; Wu, T.; Gonzalez, J.E.; Kadous, M.W.; Stoica, I. Routellm: Learning to Route Llms with Preference Data. *arXiv preprint arXiv:2406.18665* **2024**.
30. Mohammadshahi, A.; Shaikh, A.R.; Yazdani, M. Routoo: Learning to Route to Large Language Models Effectively. *arXiv preprint arXiv:2401.13979* **2024**.
31. Dekoninck, J.; Baader, M.; Vechev, M. A Unified Approach to Routing and Cascading for Llms. *arXiv preprint arXiv:2410.10347* **2024**.
32. Shirkavand, R.; Gao, S.; Yu, P.; Huang, H. Cost-Aware Contrastive Routing for Llms. *arXiv preprint arXiv:2508.12491* **2025**.
33. Yue, M.; Zhao, J.; Zhang, M.; Du, L.; Yao, Z. Large Language Model Cascades with Mixture of Thoughts Representations for Cost-Efficient Reasoning. *arXiv preprint arXiv:2310.03094* **2023**.
34. Kazimieras Zavadskas, E.; Antucheviciene, J.; Chatterjee, P. Multiple-Criteria Decision-Making (MCDM) Techniques for Business Processes Information Management. *Information* 2018, *10*, 4.
35. Saaty, T.L. A Scaling Method for Priorities in Hierarchical Structures. *J. Math. Psychol.* **1977**, *15*, 234–281.
36. Saaty, T.L. Decision Making with the Analytic Hierarchy Process. **2002**.
37. Hwang, C.-L.; Yoon, K. Multiple Attribute Decision Making: Methods and Applications a State-of-the-Art Survey; Springer Science & Business Media, 2012;
38. Kaliszewski, I.; Podkopaev, D. Simple Additive Weighting—A Metamodel for Multiple Criteria Decision Analysis Methods. *Expert Syst. Appl.* **2016**, *54*, 155–161.
39. Ciardiello, F.; Genovese, A. A Comparison between TOPSIS and SAW Methods. *Ann. Oper. Res.* **2023**, *325*, 967–994.
40. Chakrabortty, R.K.; Abdel-Basset, M.; Ali, A.M. A Multi-Criteria Decision Analysis Model for Selecting an Optimum Customer Service Chatbot under Uncertainty. *Decision Analytics Journal* **2023**, *6*, 100168.
41. Nowak, M.; Mierzwiak, R.; Butlewski, M. Occupational Risk Assessment with Grey System Theory. *Cent. Eur. J. Oper. Res.* **2020**, *28*, 717–732.