DOI 10.18.276/978-83-8419-053-1-11

28th European Conference on Artificial Intelligence (ECAI 2025)

ASSESSING LLM'S THROUGH SCENARIO-BASED SOFT SKILLS TASKS FOR MANAGERS: EXPLORATORY STUDIES

Sara Drożdżyńska[a], Paweł Kucharski[b]

[a] *University of Szczecin, Institute of Management, Szczecin, Poland,*
[b] *University of Szczecin, Doctoral School, Szczecin, Poland*

ABSTRACT

Purpose: *The aim of this study is to analyze the training potential of large language models (LLMs) in developing managerial soft skills. By potential, we refer to the model's demonstrated understanding of the key elements involved in resolving conflict scenarios.*

Need for the study: *Soft skills remain a critical component of managerial effectiveness, yet traditional training methods often lack scalability and accessibility. With the rise of advanced language models capable of human-like interaction, there is an opportunity to explore their applicability in leadership development. This study addresses a gap in current literature by evaluating the practical use of LLMs for supporting the growth of interpersonal and emotional intelligence among managers.*

Methodology: *The study employed an experimental design based on ten conflict scenarios representing typical managerial challenges. Each scenario consisted of a short background narrative, a description of the employee's personality, and a task for the manager. The scenarios were processed by three large language models—GPT-4o, Gemini 2.0 Flash, and Claude 3.5 Haiku. The model responses were evaluated by the authors and supervisory LLM model across seven categories.*

Findings: *All tested models demonstrated a solid understanding of conflict dynamics and offered appropriate managerial responses. The supervisory distinctly favored Claude 3.5 Haiku, which it rated notably higher than the other models.*

Practical Implications: *The study confirms that large language models can potentially support managerial development by offering interactive, on-demand, and risk-free simulations. A solid understanding of the dynamics underlying conflict scenarios may indicate the validity of using them in simulation-based training.*

Keywords: Soft Skills Development; Managerial Training; Large Language Models; Digital Learning Tools

Jel codes: M12; O33; D83

1. INTRODUCTION

The rapid advancement and widespread availability of large language models (LLMs) have significantly transformed the contemporary landscape of work and education. These models are now accessible to

Peer-review under responsibility of the scientific committee of 28th European Conference on Artificial Intelligence (ECAI 2025) and Intelligent Management Workshop.

the general public, require no specialized technical expertise for operation, and can be utilized virtually anywhere – on various devices, at any time, and across a broad spectrum of professional and personal domains. However, this technological expansion has also introduced a degree of uncertainty, particularly with regard to job security and the evolving role of human labor in knowledge-based and soft skills professions. Many individuals across diverse sectors perceive LLMs as a disruptive force, potentially threatening the stability of employment and altering the traditional value of human experience.

Such anxieties, while understandable, are not new. They represent a recurring social pattern in which the introduction of groundbreaking technologies – regardless of their actual capacity – triggers widespread concern about disintermediation, automation, and redundancy. Historically, these concerns often precede periods of adaptation, during which new tools are eventually integrated into existing professional frameworks, frequently augmenting rather than replacing human capabilities.

In light of this context, we chose to critically examine the potential applications of LLMs in the domain of managerial education and development. Specifically, the authors focused on their capacity to support and enhance the acquisition of soft skills, which are essential to effective leadership. Our central research hypothesis was that language models are capable of accurately identifying key elements necessary for conducting a conflict scenario simulation, taking into account the characteristics of the employee involved. An additional hypothesis assumed that supervision of the simulation by a language model is feasible, which would be reflected in a similar pattern of key point evaluations between human raters and the model.

To explore this question, we designed an experimental study aimed at analyzing the training potential of LLMs in the context of managerial development. The aim of our study was to determine whether language models demonstrate a solid understanding of conflict situations and are able to identify key elements relevant to their resolution.

This inquiry is particularly relevant in today's volatile, uncertain, complex, and ambiguous business environment, where organizations face continuous pressure to innovate, adapt, and maintain a competitive edge through strong and agile leadership. Within such a context, the ability to access intelligent, AI-based tools for managerial development offers a promising complement to traditional training programs. LLMs, by virtue of their conversational and generative capabilities, may serve as platforms for leadership simulation, role-play, reflective practice, and experiential learning – methods long recognized as effective in adult education and executive coaching.

Our preliminary findings suggest that large language models can indeed play a meaningful role in the development of soft managerial skills. Rather than posing an existential threat to human professionals, these models may enhance human potential by offering continuous, scalable, and personalized learning experiences. Furthermore, the integration of LLMs into organizational learning ecosystems may contribute to error reduction, more informed decision-making, and greater alignment between managerial behavior and organizational goals. The researchers also observed that the use of LLMs in simulation-based learning enables a new form of interactive, on-demand mentoring—one that could easier the access to high-quality managerial development resources, especially in organizations with limited budgets for traditional training interventions.

2. LITERATURE REVIEW

Managerial competency training has emerged as a cornerstone of leadership development, ensuring that managers at all levels are equipped with the necessary skills to navigate complex organizational environments. Managers play a crucial role in facilitating knowledge transfer and embedding learning within workplace practices. (Gumuseli & Ergin, 2002) In the subject literature, we can find many definitions of competencies, both detailed and general. Regardless of how we examine and analyze them, when we speak of 'competencies', we refer to the ability to effectively utilize one's own potential or the potential of an organization. This potential lies in knowledge, skills, attitudes, employees, effective leadership, motivation, available resources, psychological assets, environmental conditions, etc. The twenty-first century is a time focused on efficiency. Therefore, we expect competencies to lead to the desired outcomes. All administrative levels in nowadays organizations, need talented, well prepared top managers and this need is constantly increasing, and the effectiveness of management

becomes one of the most crucial conditions for organization's success. (Zakarevičius & Župerkienė, 2008)

Several competency frameworks have been developed to structure training programs. One widely referenced model is the Kirkpatrick Framework, which evaluates training effectiveness through four levels: reaction, learning, behavior, and results. The Kirkpatrick model provides a systematic method for assessing the impact of managerial training interventions. Training is an instrument to expand the knowledge base of the employees and allows them to transfer this on their jobs in the form of improved performance (Sahni, 2020).

The general strengths of the Kirkpatrick model in evaluation theory and practice have been extolled by scholars. They recognize the model for its ability to provide the following: simple system or language in dealing with the different outcomes and how information about these outcomes can be obtained; descriptive or evaluative information about the kind of training that are needed, thus allows organizations to anchor the results of what they do in business points of view; and practical approach for the typically complex evaluation process. With these strengths, it cannot be denied that Kirkpatrick model has offered significant contributions to the evaluation theory and practice (Cahapay, 2021).

Knowing the value of training and its impact on an organization, it is worth considering the possibility of creating an equally effective model using artificial intelligence and large language models (LLMs) available on the market. This is not about preparing training materials per se, but rather about exploring the potential for collaboration with modern technologies, grounded in the structure and strengths of the aforementioned training model. The authors of the article see potential in the use of LLMs for employee training, as well as in the practical application of tools that, at this moment, may appear more as a threat than as a clear opportunity for development.

2.1. Defining Managerial Competencies

Managerial competencies encompass a wide array of knowledge, behaviors, and skills required for effective leadership. Competencies can be grouped into strategic orientation, business management, people management, and self-management. These competencies evolve depending on management levels, emphasizing flexibility, decision-making, communication, and emotional intelligence (Kwaśniewski, 2024).

One of the first definitions of this concept was presented by Boyatzis in 1982. He indicated that competencies constitute a combination of general knowledge, motivation, traits, self-image, social roles, and skills that are essential for properly performing a specific job. Another insightful definition was proposed by Levy-Leboyer in 1996, who emphasized that competencies involve the integrated use of abilities, personality traits, and acquired knowledge and skills in order to successfully carry out complex tasks within an organization (Witaszek, 2011).

Additionally, Boyatzis in 2011 highlighted emotional, social, and cognitive intelligence as foundational for behavioral leadership competencies (Boyatzis, 2011). Their research underlines the importance of experiential learning and receiving feedback in fostering these skills. Experiential learning offers several key advantages in managerial development:

- contextual relevance: managers learn in environments that mirror their actual work contexts, increasing the transferability of skills,
- immediate feedback: learning through action enables the rapid identification of strengths and areas for improvement,
- personalization: experiences are subjective, allowing managers to draw lessons that are uniquely meaningful to their roles and personalities,
- emotional engagement: real-life challenges elicit emotions that strengthen memory retention and deepen understanding.

According to Kolb's experiential learning theory (1992), effective learning is a cyclic process involving four stages: 1. Actions. Specific experience from what we call 'the real life'. 2. Thinking. Empiric observation and analysis of activities. 3. Understanding. Logical analysis of the observations, search for respective patterns in observations, memorising what was taught. 4. Checking. Checking how new expertise relates to 'the real world' and how the behaviour has changed (Zakarevičius & Župerkienė, 2008). Managers who participate in this cycle are better equipped to internalize complex competencies such as strategic thinking, conflict resolution, and adaptive leadership. Integrating experiential learning into competency development initiatives fosters a growth mindset, enhances

resilience, and encourages lifelong learning among managers. It also supports the cultivation of critical metacompetencies, such as self-awareness, systems thinking, and emotional regulation, which are increasingly essential in complex organizational environments.

The authors assume that these elements can be achieved with proper use of LLMs as a training tool. Experiential learning plays a pivotal role in the development of managerial competencies. Unlike traditional training methods focused primarily on the transmission of theoretical knowledge, experiential learning emphasizes the active engagement of managers in real-world problem-solving, decision-making, and leadership challenges. It promotes deep, durable learning by allowing individuals to reflect on their experiences, adapt their behaviors, and continuously refine their skills. AI and large language models can work collaboratively with trainees to analyze real-life situations and evaluate the solutions they have applied. This enables individuals to receive feedback and test their own decisions and approaches in a fast, widely accessible manner – without the need to wait for consultation with a supervisor. In high-pressure situations, modern technologies can support managers and their skillsets by providing analysis of alternative solutions without triggering real-world consequences.

Competencies Managerial training programs significantly contribute to the development of individual and organizational capabilities. Managerial support is essential in ensuring that knowledge gained through training is effectively transferred to the job (Gumuseli & Ergin, 2002). Training initiatives enable managers to better handle strategic tasks, lead teams with confidence, and align departmental goals with broader organizational strategies.

In small business contexts, management training correlates positively with improved organizational performance, especially when the training content is tailored to industry-specific challenges (Panagiotakopoulos, 2020).

The impact of managerial training extends beyond skill acquisition. It directly influences employee satisfaction and retention. Investments in leadership development lead to reduced turnover intentions and higher engagement among staff (Malek et al., 2018). Managers who undergo structured training are more likely to foster positive work cultures, provide effective feedback, and support employee development. Of course, this perspective applies to training in its traditional sense. While the benefits are substantial, organizations must also consider the design and delivery of training programs. Context, duration, follow-up, and managerial involvement are key success factors. Lack of strategic alignment or insufficient post-training support can undermine program effectiveness. All these reasons present a perspective to explore the topic of modern tools that can provide managers with the necessary skills training at almost any time. The use of AI and large language models (LLMs) can serve as tools to support managers in moments when they are unsure of their competencies in a given situation. This kind of support is likely to enhance a sense of stability and confidence among employees.

Furthermore, transforming managerial competencies becomes critical in ensuring resilience, innovation, and sustainability. That is a global challenge that can reshape business practices (Kwaśniewski, 2024).

2.2. LLMs in manager's soft skills development

In recent years, we have observed rapid advancements in language models (Zhao et al., 2023). This process was undoubtedly influenced by two key events. From a technical perspective, it was the introduction of the attention mechanism proposed by Vaswani (Vaswani et al. 2017), which enabled the creation of modern language model architectures. From a societal perspective, a significant milestone was OpenAI's release of the ChatGPT chatbot in 2022 (OpenAI, 2022), based on the GPT-3.5 model (OpenAI, 2022), allowing a broad audience to experiment with the technology and seek practical applications. Since then, many competing models have emerged, such as Gemini (Anil et al., 2023), Grok (xAI, 2024), Claude (Anthropic, 2024), DeepSeek (Bi et al., 2024), and LLama (Touvron et al., 2023), which also achieve excellent evaluation results (Rangapur & Rangapur, 2024).

It did not take long before attempts were made to apply language models to management (Nihal et al., 2023), particularly in the area relevant to this publication: the development of managerial soft skills (Sun, 2024).

We can distinguish two main directions in utilizing language models in this context. The first is an advisory approach, where the language model acts as a virtual assistant for management staff. AI-based systems supporting business decision-making are not new (Huang, 2024); however, only recent advances in language models have enabled their use in a 'conversation with an expert' format. In this

arrangement, the model analyzes the situation, offers advice, identifies potential risks, and generates possible solutions. Additionally, these systems can be fed large amounts of internal company data, recognizing subtle patterns and relationships that might otherwise escape human notice. This allows for data-driven decision-making and minimizes the impact of personal biases and political factors on decision-making processes. Nevertheless, it is crucial to emphasize that these systems are intended solely as advisory tools, with ultimate decisions and responsibility resting on the manager, who must also consider immeasurable or hard to measure factors such as team morale or organizational culture.

In addition to complex advisory systems, language model-based chatbots can be used less formally by managers, for instance, in brainstorming sessions. Interestingly, research suggests that advice generated by language models is perceived as highly substantive (Howe et al., 2023). Naturally, this also carries certain risks. Biases present in language models' training datasets can influence managers' management styles. Studies have shown (Geleilate & Humberd, 2024) that managers who used ChatGPT to solve case studies tended to prefer more rigid and controlling solutions. Another study (Kirshner et al., 2025) demonstrated that ChatGPT exhibits typical human cognitive biases in decision-making. This highlights the necessity of providing the model with comprehensive context and considering the softer aspects of decisions. Despite this, recommendations generated by language models should always be critically evaluated and combined with the manager's own perspective.

The second direction involves using language models as simulators, enabling managers to practice skills such as conflict resolution, team management, coaching, or providing feedback. This leverages the ability of language models to simulate the personalities of team members (Wang et al., 2025). Chatbots assume the roles of subordinates, customers, or collaborators interacting with each other. Virtual dialogues allow managers to gain practical experience, which proves more effective than purely theoretical learning (Dai & Ke, 2022). Simulators enable testing various management styles without fearing real-world consequences of incorrect decisions. An additional advantage of this approach is the ability to generate a simulation report for the participating manager, highlighting areas needing improvement.

Market needs, given possibilities, a review of the literature and curiosity encouraged the authors to conduct tests on large language models in order to verify and analyze their capabilities in the area of training, developing, and supporting managerial competencies, recognizing their potential for practical application. Competency training is an indispensable element of workforce development. It strengthens leadership capacity, enhances strategic alignment, and fosters a culture of continuous improvement. As organizations embrace technological innovations like LLMs, there is a growing opportunity to simulate experiential learning digitally. Artificial Intelligence-powered simulations, virtual coaching, and scenario-based training modules can replicate real-world managerial challenges, providing managers with rich experiential learning opportunities in a scalable and flexible manner.

The primary aim of this study is to explore the potential of modern language models in the second of the two previously described roles – simulation agents. By potential, we refer to the model's demonstrated understanding of the key elements involved in resolving conflict scenarios. Given the pilot nature of this research, we will evaluate the model-generated outputs ourselves, and we will not assess the simulated conversations per se, but rather the model's ability to identify the key elements that must be included in such a simulation. Here, the model's self-awareness becomes critical and resembles reasoning processes used in prompting strategies such as chain-of-thought (Wei et al., 2022). Evaluating the capacity of language models to perform this function will allow us to determine whether this is a research direction worth pursuing with the current generation of models, or whether it would be more prudent to wait for future, more advanced iterations.

An additional goal of this study is to investigate how well language models perform in the evaluation of other language models, and how their assessments compare to those made by human. The potential of such an approach – where an LLM oversees the performance of another LLM during simulation – is important to examine, given the emerging interest in self-supervising multi-agent systems. In such a configuration, an evaluator model could generate guidelines and feedback for simulation agents (those emulating employees) based on ongoing analysis or even intervene in real time during a simulation. For this to be feasible, the evaluator model must perceive and interpret the behavior of other models in a manner comparable to human evaluators.

3. METHODOLOGY

To conduct an evaluation of the potential of large language models in managerial soft skills development, we have prepared a dataset containing 10 tasks (conflict situations) designed for managers to solve. Each task consists of three key components:

1. Scenario – a description of a problematic situation,
2. Employee Personality – a description of the role that the language model should play,
3. Manager's Task – instructions outlining the objectives that the participant in the simulation must achieve.

Examples of these tasks are presented in Table 1.

Table 1. Example tasks used to test the usefulness of language models for developing soft skills

Scenario	Employee Personality	Manager's Task
Emily Taylor recently requested a salary increase based on her past achievements. However, her recent interactions with clients have been marred by sarcastic remarks and a noticeable lack of professionalism, which has raised questions about her current performance levels.	Emily is confident and knowledgeable, with a proven track record. Nevertheless, she becomes notably defensive when faced with criticism and her current approach has begun to erode trust among clients and colleagues alike.	Conduct a performance review meeting with Emily. Present specific examples of her recent behavior, discuss the discrepancies between her achievements and current performance, and collaboratively create a development plan aimed at addressing these issues.
Your employee John Doe has been repeatedly late to work and missing scheduled virtual meetings. This inconsistency has disrupted project timelines and frustrated team members, especially during critical phases of projects.	John is highly capable and creative but struggles with organization and time management. While he has shown great potential in his work, personal challenges seem to affect his punctuality. He is usually receptive to feedback but can become defensive when he feels singled out.	Conduct a candid disciplinary conversation with John. Explain the impact his tardiness and lack of communication have on the team, and work together to develop a clear plan to improve his punctuality and participation in meetings.
Michael Brown has recently submitted a vacation request during a critical project period. Compounding this issue, there are reports that he has been neglecting essential safety procedures at work, which is concerning for both team productivity and overall workplace safety.	Michael is generally dedicated and conscientious, but recent personal stress may have led to lapses in attention and an increasing tendency to take shortcuts. This shift in behavior is uncharacteristic of his usual performance.	Discuss the timing of his vacation in the context of the current project demands. Emphasize the importance of following safety protocols, and collaboratively devise a plan to ensure that his absence does not adversely affect team operations.

Source: own elaboration.

To assess the usefulness of language models as simulators, we evaluated the prepared tasks using large language models. The prompt used in our assessment can be found in Appendix A. As part of the task evaluation, we requested responses addressing the following aspects:

1. What key discussion points must be included for the task to be considered successfully completed?
2. What key discussion points must be included for the task to be considered unsuccessfully completed?

3. What are the critical aspects of problem-solving that the participant must achieve to complete the task?

The responses generated by the models were compared and analyzed. For this comparison, we selected the following large language models (LLMs) available as of today: GPT-4o, Gemini 2.0 Flash, and Claude 3.5 Haiku.

3.1 Evaluation of model-generated responses by human

Given the pilot nature of the study, the initial evaluation of model-generated responses was conducted by the authors themselves. To ensure consistency and depth in the analysis, we proposed seven evaluation categories, detailed in Table 2 along with their definitions. In addition to standard aspects such as clarity of expression and relevance of the response, we introduced criteria focused on the model's capacity for personalization, and its sensitivity to contextual dimensions such as empathy, collaboration, and the importance of feedback in interpersonal scenarios. In their evaluations, the authors strictly followed the definitions introduced in Table 2, drawing on both their theoretical and practical knowledge in the field of management.

The evaluation categories were as follows:

Table 2. Example tasks used to test the usefulness of language models for developing soft skills

Category	Description
Success Points	Did the model correctly identify the key elements that should be present for the managerial conversation to be considered successful?
Failure Points	Did the model sensibly highlight the main risk factors that may lead to failure in the given scenario? Did it overlook critical threats?
Empathy	Did the model demonstrate an understanding of the emotional dynamics of the situation? Did it acknowledge the need for empathy toward the employee?
Cooperation	Did the model reflect an awareness of the collaborative nature of the problem? Did it emphasize joint responsibility, active listening, or team engagement?
Feedback	Did the model understand the importance of feedback in the scenario? Did it include concrete and constructive feedback strategies?
Structure	Was the response logically structured, clearly articulated, and easy to follow? Or was it vague, inconsistent, or overly general?
Context	Did the model refer specifically to the scenario context (e.g., employee traits, situation-specific details), or did it rely solely on generic statements?

Source: own elaboration.

All model responses were evaluated in random order using a five-point Likert-type scale (see Table 3):

Table 3. Rating scale used to evaluate model responses in soft skills assessment tasks

Score	Description
1 – very poor	The model failed to recognize key elements, provided random or irrelevant answers, and showed no connection to the scenario.
2 – poor	The model captured some relevant points but omitted or misinterpreted key elements. The response was either vague or inaccurate.

3 – fair	The main aspects were correctly identified, but the response lacked depth or full alignment with the scenario.
4 – good	Most key points were correctly recognized, context was considered, and the response was logically structured.
5 – very good	The response was accurate, deep, and highly adapted to the scenario, incorporating subtle elements appropriately.

Source: own elaboration.

In the context of the evaluation scale for model responses proposed above, we assume that a score of 3 or higher indicates the potential for using a language model in training managerial soft skills. This is based on the premise that if the model is capable of correctly understanding the scenario's assumptions, it is also capable of incorporating them appropriately within a simulated interaction.

The statistical analysis was based on the null hypothesis assuming no significant performance differences among the evaluated LLMs. To compare model performance across the defined categories (see Table 2), we conducted the non-parametric Friedman test, which is appropriate for repeated-measures comparisons across multiple conditions. The significance threshold was set at $\alpha = 0.05$. For categories where significant differences were detected, we performed post-hoc pairwise Wilcoxon tests, corrected for multiple comparisons using the Holm-Bonferroni method. The decision to use non-parametric methods was driven by the limited number of observations (10 scenarios per category) and the potential violation of normality assumptions, which are required for classical ANOVA.

3.2. Evaluation of model-generated responses by LLM

In the next stage of the study, we employed the Gemini 2.5 Pro model as a reviewer to assess the same responses generated by the three LLMs. The model was prompted to assign scores using the same scale and categories as the human evaluators (the exact prompt is included in Appendix B). The scores generated by Gemini 2.5 Pro were then analyzed in the same way as the human ratings - via Friedman tests and Wilcoxon comparisons with Bonferroni correction, maintaining the $\alpha = 0.05$ significance threshold. The choice of the model was due to the fact that it is currently one of the most powerful models available on the market. The model received the category definitions and rating definitions mentioned in Table 2 and Table 3 as part of the prompt.

Finally, we compared the evaluations assigned by the human raters with those produced by the Gemini 2.5 Pro model. To do so, we used the Spearman rank correlation test to assess the alignment between the two sources of ratings. To detect any systematic bias, we additionally conducted a paired Wilcoxon signed-rank test.

4. RESULTS

The evaluation results for each model across all categories are presented in Figure 1, along with standard deviations to illustrate the variability in scoring. Most ratings across all models fell within the 2 – 4 point range. As shown in the Figure 1, Claude 3.5 Haiku achieved the highest scores across all categories, except for Cooperation, where it ranked second. GPT-4o in most categories occupied the second position overall, while Gemini 2.0 Flash ranked lowest in most categories.

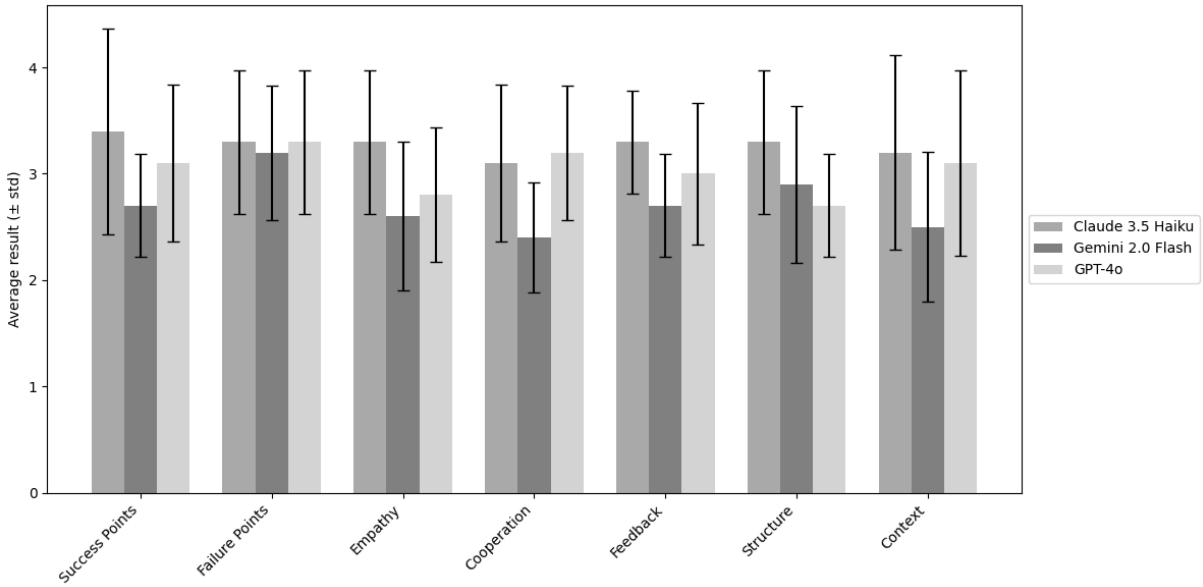

Figure 1. Average results of evaluation in all categories for all three tested models
Source: own elaboration.

The greatest divergence in model scores was observed in the Cooperation category, while the Failure Points category exhibited the smallest variation. Regarding standard deviation, it did not exceed the value of 1 in any category. The Feedback category had the lowest score dispersion, whereas Context exhibited the highest. In terms of average score variability per model, Gemini 2.0 Flash showed the most consistent performance, while Claude 3.5 Haiku had the highest variation.

To statistically assess the significance of differences between the models across all categories, we applied the Friedman test for three models and ten scenarios per category. The test statistics are summarized in Table 4:

Table 4. Results of the Friedman test for significance level at $\alpha = 0.05$. Statistically important results are bolded

Category	χ^2 (df=2)	p
Success Points	4.522	0.104
Failure Points	0.080	0.961
Empathy	7.000	**0.030**
Cooperation	8.960	**0.011**
Feedback	4.526	0.104
Structure	4.323	0.115
Context	4.200	0.122

Source: own elaboration.

Based on these results, we identified two categories (Empathy and Cooperation) where the differences between models were statistically significant. From a managerial practice perspective, it is important not only that language models accurately identify elements of conflict, but also that their understanding and responses can genuinely support managers in making interpersonal decisions. For example, higher scores in the 'Empathy' category may suggest that model is more capable of recognizing the emotional aspects of a situation and adjusting its communication to the needs of the employee – an essential factor for effective conflict resolution and building trust within a team. For categories with statistical significance, we conducted post-hoc pairwise Wilcoxon tests with Bonferroni correction. The results are summarized in Table 5:

Table 5. Post-hoc pairwise comparisons between models in categories with significant group-level differences (Wilcoxon tests with Holm-Bonferroni correction). Statistically important results are bolded

Category	Model comparison	p Bonferroni
Empathy	Claude 3.5 Haiku vs Gemini 2.0 Flash	0.244
	Claude 3.5 Haiku vs GPT 4o	0.573
	Gemini 2.0 Flash vs GPT 4o	0.629
Cooperation	Claude 3.5 Haiku vs Gemini 2.0 Flash	0.319
	Claude 3.5 Haiku vs GPT 4o	1.000
	Gemini 2.0 Flash vs GPT 4o	0.080

Source: own elaboration.

In the next part of research, we employed the Gemini 2.5 Pro model to evaluate the same responses from the three models across seven categories. The results are presented in Figure 2. What can be immediately observed is that, overall, the LLM model assigned higher scores, as indicated by the ratings ranging between 3 and 5. Another noticeable feature is the lower variability in the scores assigned by human raters. In extreme cases - such as the Structure category across all models and the Failure Points category for Gemini 2.0 Flash and GPT-4o – the standard deviation is zero, which means the model assigned identical scores to all scenarios within these models and categories. Similar to the ratings provided by the authors, Claude 3.5 Haiku achieved the highest scores. The other two models obtained nearly identical results in most categories.

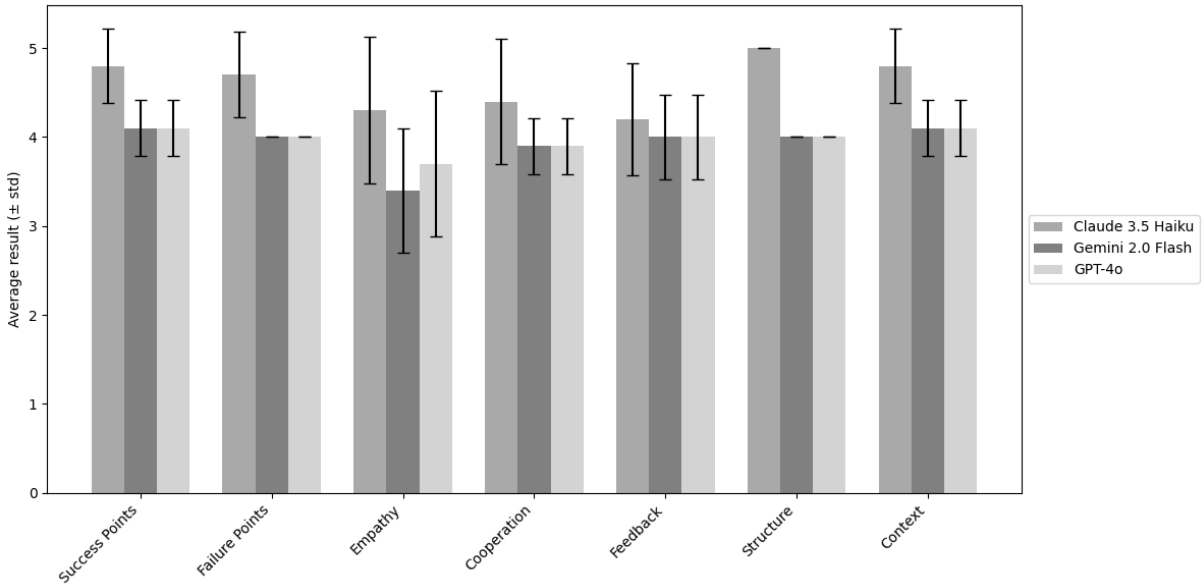

Figure 2. Average results of evaluation by Gemini 2.5 Pro model in all categories for all three tested models
Source: own elaboration

To statistically assess the significance of differences between the models across all categories, we applied again the Friedman test for three models and ten scenarios per category. The test statistics are summarized in Table 6. In this case, we see statistically significant differences for all categories except for Feedback. For Structure, we obtained an extremely low p-value, which results from the zero scatter of estimates for all three models in this category.

Table 6. Results of the Friedman test for rating created by Gemini 2.5 Pro, significance level at α = 0.05. Statistically important results are bolded

Category	χ^2 (df=2)	p
Success Points	14.000	**0.001**
Failure Points	14.000	**0.001**
Empathy	12.087	**0.002**
Cooperation	10.000	**0.007**
Feedback	4.000	0.135
Structure	20.000	**0.000**
Context	14.000	**0.001**

Source: own elaboration.

For all categories except Feedback, we conducted post-hoc pairwise Wilcoxon tests with Bonferroni correction. The results are summarized in Table 7:

Table 7. Post-hoc pairwise comparisons between models in categories with significant group-level differences (Wilcoxon tests with Holm-Bonferroni correction) for rating created by Gemini 2.5 Pro. Statistically important results are bolded

Category	Model comparison	p Bonferroni
Success Points	Claude 3.5 Haiku vs Gemini 2.0 Flash	**0.049**
	Claude 3.5 Haiku vs GPT 4o	**0.049**
Failure Points	Claude 3.5 Haiku vs Gemini 2.0 Flash	**0.049**
	Claude 3.5 Haiku vs GPT 4o	**0.049**
Empathy	Claude 3.5 Haiku vs Gemini 2.0 Flash	**0.049**
	Claude 3.5 Haiku vs GPT 4o	**0.049**
Cooperation	Claude 3.5 Haiku vs Gemini 2.0 Flash	0.051
	Claude 3.5 Haiku vs GPT 4o	0.051
Structure	Claude 3.5 Haiku vs Gemini 2.0 Flash	**0.014**
	Claude 3.5 Haiku vs GPT 4o	**0.014**
Context	Claude 3.5 Haiku vs Gemini 2.0 Flash	**0.049**
	Claude 3.5 Haiku vs GPT 4o	**0.049**

Source: own elaboration.

The obtained results indicate that in all evaluated categories – except for Cooperation – we observed statistically significant differences in favor of the Claude 3.5 Haiku model, which outperformed the other two models in a statistically meaningful way. It is worth noting here that outside the Structure category, statistical significance value is on the border and must be treated with caution. In the automated evaluation conducted by Gemini 2.5 Pro, Claude 3.5 Haiku obtained the highest scores in all categories except for Cooperation. This may indicate its stronger capacity to identify emotionally and contextually relevant elements of conflict scenarios, which is particularly important in training soft managerial skills. The lack of a clear advantage in the Cooperation category suggests that current models may still face limitations in reflecting the interactive and reciprocal nature of collaborative workplace situations.

Finally, we turn to the comparison between human evaluations and those provided by the automated evaluator, in this case represented by the Gemini 2.5 Pro model. The model's ratings were discussed in the preceding section. To examine the consistency in rank ordering, we used Spearman's rank correlation analysis, with the results presented in Table 8.

Table 8. Spearman rank correlation coefficients between human and LLM-based evaluations across categories. Significance level at $\alpha = 0.05$. Statistically important results are bolded

Category	Spearman ρ	p
Success Points	0.283	0.129
Failure Points	0.168	0.375
Empathy	0.538	**0.002**
Cooperation	0.111	0.560
Feedback	0.224	0.235
Structure	0.357	0.053
Context	0.305	0.102

Source: own elaboration.

Statistically significant Spearman coefficients were observed only for Empathy catgory. The coefficient suggests a moderate positive correlation.

In the next step, we examined whether one of the evaluation systems (human vs. LLM-based) consistently rated model responses higher or lower than the other. For this purpose, we used the Wilcoxon signed-rank test, and the results are presented in Table 9.

Table 6. Wilcoxon signed-rank test results comparing the level of ratings between human and LLM-based evaluations

Category	Wilcoxon p
Success Points	**0.0001**
Failure Points	**0.0001**
Empathy	**0.0001**
Cooperation	**0.0001**
Feedback	**0.0001**
Structure	**0.0001**
Context	**0.0001**

Source: own elaboration.

As shown in the table above, statistically significant results were obtained across all categories, indicating that the LLM systematically assigns higher scores to the evaluated model responses than the human raters. The discrepancies between human evaluations and those generated by the Gemini 2.5 Pro model indicate a possible lack of rigor in automated assessment or a different prioritization of communicative aspects. In practice, this means that using LLMs to supervise simulations (e.g., as coaching agents) requires further research into their independence and their capacity for reliable evaluation.

5. DISCUSSION

The results obtained in our study indicate that all evaluated models performed well in identifying the core challenges associated with the development of soft skills among managerial staff. Model ratings, which mostly ranged between 2 and 4 on the applied scale, suggest that LLMs are indeed useful tools for addressing the problem space described in this study. Particularly noteworthy are the results related to the inclusion of empathy, cooperation, and feedback in simulated managerial conversations. All three models appear to recognize the relevance of these interpersonal dimensions in effective leadership.

Despite apparent differences in mean scores – and the seemingly dominant performance of Claude 3.5 Haiku – statistical analysis revealed significant differences only in two categories: Empathy and Cooperation. However, these differences were not confirmed by the post-hoc analysis with Holm correction, which identified only a single statistically significant pairwise difference: Gemini 2.0 Flash vs GPT-4o in the Cooperation category, where Gemini exhibited the weakest performance.

The lack of sufficient evidence to reject the null hypothesis – stating that all models demonstrate comparable levels of understanding the conflict scenarios – may serve as an indication that these LLMs are, in practice, similarly applicable for supporting the development of managerial soft skills.

The situation differs in the case of evaluations conducted by the Gemini 2.5 Pro model. Compared to human raters, the model consistently assigned higher scores to the responses of the other models and exhibited lower variance across scenarios, suggesting it applied a narrower scoring range. Under this evaluation scheme, Gemini 2.0 Flash and GPT-4o achieved very similar results – a finding that was also confirmed by statistical analysis. The case of Claude 3.5 Haiku, however, presents a different picture. Pairwise comparisons with the other two models revealed that Claude outperformed them in five out of seven evaluated categories, indicating a more decisive advantage under the automated review. One particularly noteworthy finding is the consistency of statistically significant distinctions identified by the Gemini 2.5 Pro model when evaluating outputs generated by other LLMs. This suggests a promising direction for systems aiming to leverage agent-based architectures for soft skills training in managerial contexts.

In the final stage of the study presented in this paper, we compared the evaluation of model responses conducted by the authors with that of an automated scoring system implemented via the Gemini 2.5 Pro model. Preliminary analysis based on score distribution plots suggested a potential systematic overestimation of responses by the LLM compared to the human raters which was later confirmed with statistical analysis. Rank correlation analysis showed only moderate correlation for Empathy category.

Our study indicate that all evaluated models performed adequately in identifying interpersonal and procedural elements central to conflict resolution. From a managerial perspective, the ability of Claude 3.5 Haiku to more effectively address categories such as empathy or contextualization may enhance the quality of feedback during employee conversations and signal its greater utility for leadership development tools.

However, it is essential to highlight the potential biases in LLM-generated evaluations. As shown in our results, the automated model (Gemini 2.5 Pro) consistently rated responses higher than human experts. This may stem from shared architectural features or alignment strategies during training that prioritize politeness or generic completion over critical analysis. Such tendencies raise concerns about overestimating the effectiveness of peer LLMs, especially if used for evaluative purposes in multi-agent systems.

Naturally, several improvements should be considered for future iterations. Most notably, the model evaluation was conducted solely by the authors, resulting in a limited sample size and potential evaluator bias. Replicating the study with a larger and more diverse group of raters would undoubtedly strengthen the robustness and generalizability of the results. In particular, involving experienced managers as evaluators could enable deeper insights into the behavioral validity of the model responses and provide more credible assessments of their practical utility.

6. CONCLUSIONS

In this publication, we reviewed the issue of developing soft skills among managers and the role of large This study demonstrated that large language models are capable of recognizing key elements in managerial conflict scenarios and that they differ in the quality and depth of such recognition. Claude 3.5 Haiku achieved the highest scores in multiple categories, particularly in emotional and contextual understanding, suggesting that it may offer greater utility in developing soft skills related to empathy and communication.

At the same time, we identified that LLM-based evaluations tend to be more lenient and less differentiated than human evaluations. This finding underscores the importance of human oversight in simulation-based training and highlights potential risks in fully automating evaluation processes.

While the study supports the use of LLMs as simulation agents, the application of multi-agent systems or fully autonomous supervisors was not tested. As such, any recommendations in that direction remain hypothetical and should be treated cautiously.

Claude 3.5 Haiku showed higher effectiveness in the evaluation of soft managerial skills scenarios, which may suggest its potential applicability in personalized leadership training programs. While human raters did not observe statistically significant differences between the models – indicating that all performed satisfactorily – the automated evaluation suggests that certain models may better capture specific interpersonal competencies. This may point to subtle but meaningful differences in their suitability for particular training contexts.

Future research should explore whether these results hold in real-time simulations, involve diverse managerial participants, and test the longitudinal effectiveness of LLM-based training programs in real organizational settings.

Funding: Co-financed by the Minister of Science under the "Regional Excellence Initiative".

REFERENCES

Anil, R., Borgeaud, S., Alayrac, J. -B., et al. (Gemini Team, Google DeepMind). (2023). *Gemini: A family of highly capable multimodal models*. arXiv. https://doi.org/10.48550/arXiv.2312.11805.

Anthropic. (2024). *Claude 3: Technical overview*. Retrieved from: https://www.anthropic.com/claude-3-technical-overview (04.04.2025).

Bi, X., Chen, D., Chen, G., et al. (2024). *DeepSeek LLM: Scaling open-source language models with longtermism*. arXiv. https://doi.org/10.48550/arXiv.2401.02954

Boyatzis, R. (2011). Managerial and Leadership Competencies: A Behavioral Approach to Emotional, Social and Cognitive Intelligence. Vision: *The Journal of Business Perspective, 15*, 91–100. https://doi.org/10.1177/097226291101500202.

Cahapay, M. (2021). Kirkpatrick Model: Its Limitations as Used in Higher Education Evaluation. *International Journal of Assessment Tools in Education*, *8*(1), 135–144. https://doi.org/10.21449/ijate.856143.

Dai, C. P., & Ke, F. (2022). Educational applications of artificial intelligence in simulation-based learning: A systematic mapping review. *Computers and Education: Artificial Intelligence, 3*, 100087. https://doi.org/10.1016/j.caeai.2022.100087.

Geleilate, J. M. G., & Humberd, B. K. (2024). GenAI tools and decision-making: Beware a new control trap. MIT Sloan Management Review. Retrieved from: https://sloanreview.mit.edu/article/genai-tools-and-decision-making (04.04.2025).

Gumuseli, A. I., & Ergin, B. (2002). The manager's role in enhancing the transfer of training: A Turkish case study. *International Journal of Training and Development, 6*, 80–97. https://doi.org/10.1111/1468-2419.00151.

Howe, P. D. L., Fay, N., Saletta, M., & Hovy, E. (2023). ChatGPT's advice is perceived as better than that of professional advice columnists. *Frontiers in Psychology, 14*, 1281255. https://doi.org/10.3389/fpsyg.2023.1281255

Huang, Y. (2024). *Levels of AI agents: From rules to large language models*. arXiv. https://doi.org/10.48550/arXiv.2405.06643

Kirshner, S. N., Ovchinnikov, A., Andiappan, M., & Jenkin, T. (2025). A manager and an AI walk into a bar: Does ChatGPT make biased decisions like we do? *Manufacturing & Service Operations Management, 27*(1), 1–18. https://doi.org/10.1287/msom.2023.0279.

Kwaśniewski, J. (2024). *Models of managerial competencies* (s. 74–91).

Malek, K., Kline, S., & DiPietro, R. (2018). The impact of manager training on employee turnover intentions. *Journal of Hospitality and Tourism Insights, 1*. https://doi.org/10.1108/JHTI-02-2018-0010.

Nihal, K. S., Pallavi, L., Raj, R., Babu, C. M., & Mishra, B. (2023). Enhancing soft skill development with ChatGPT and VR: An exploratory study. In: *Proceedings of the 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering* (RMKMATE 2023) (IEEE).

OpenAI. (2022, March 15). Introducing GPT-3.5 series: text-davinci-002 and code-davinci-002 models. Archived March 20, 2023.
OpenAI. (2022, November 30). ChatGPT: Optimizing language models for dialogue. OpenAI Blog. Retrieved from: https://openai.com/blog/chatgpt/ (05.04.2025).
OpenAI. (2025, January 31). OpenAI O3 mini system card. Retieved from: https://openai.com/research/o3-mini-system-card (05.04.2025).
Panagiotakopoulos, A. (2020). Exploring the link between management training and organizational performance in the small business context. *Journal of Workplace Learning*, ahead-of-print. https://doi.org/10.1108/JWL-10-2019-0121/.
Rangapur, A., & Rangapur, A. (2024). *The battle of LLMs: A comparative study in conversational QA tasks*. arXiv. https://doi.org/10.48550/arXiv.2405.18344.
Sahni, J. (2020). Managerial training effectiveness: An assessment through Kirkpatrick framework. *TEM Journal*, 1227–1233. https://doi.org/10.18421/TEM93-51.
Sun, J. (2024). Research on the application of large language models in human resource management practices. *International Journal of Emerging Technologies and Advanced Applications, 4*(8), 125–134. https://doi.org/10.62677/IJETAA.2408125.
Sypherd, C., & Belle, V. (2024). *Practical considerations for agentic LLM systems*. arXiv. https://doi.org/10.48550/arXiv.2412.04093.
Touvron, H., Lavril, T., Izacard, G., Martinet, X., et al. (2023). *LLaMA: Open and efficient foundation language models*. arXiv. https://doi.org/10.48550/arXiv.2302.13971.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30* (NeurIPS 2017), 6000–6010. https://doi.org/10.48550/arXiv.1706.03762.
Wang, Y., Zhao, J., Ones, D. S., He, L., & Xu, X. (2025). Evaluating the ability of large language models to emulate personality. *Scientific Reports, 15*(1), 519.
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv. https://doi.org/10.48550/arXiv.2201.11903.
Witaszek, Z. (2011). Rozwój kompetencji menedżerskich przesłanką sukcesu organizacji. *Zeszyty Naukowe Akademii Marynarki Wojennej, R. 52 nr 4*(187), 303–322.
Zakarevičius, P., & Župerkienė, E. (2008). Improving the Development of Managers' Personal and Professional Skills. *Engineering Economics*, *5(60)*, 104–113.
xAI. (2024, March 17). Open release of Grok-1. Retrieved from: https://x.ai/blog/grok-os (04.04.2025).
Zhao, W. X., Zhou, K., Li, J., Tang, T., et al. (2023). *A survey of large language models*. arXiv. https://doi.org/10.48550/arXiv.2303.18223.

APPENDIX A

PROMPT:
You are an expert in training managerial staff in soft skills through conflict situation simulations. After each simulation, you can assess its strengths and weaknesses and prepare a report for the user.
Below, you will find a task scenario enclosed within triple backticks (```), where you will take on the role of a problematic employee. Example:
```
Scenario:
{A scenario outlining a problematic situation}
Employee Personality:
{The personality traits of the employee you need to embody during the simulation}
Task:
{The objective that the participant engaging with you must achieve}
```

Your task is to respond to the following questions for each scenario:

1. What key discussion points must be included for the task to be considered successfully completed?
2. What key discussion points must be included for the task to be considered unsuccessfully completed?
3. What are the critical aspects of problem-solving that the participant must achieve to complete the task?

Be precise and limit your response to a maximum of 10 sentences.

APPENDIX B

PROMPT:
You are an expert in training managerial staff in soft skills through conflict situation simulations. After each simulation, you can assess its strengths and weaknesses and prepare a report for the user.
Below, you will find a task scenario and model response about most important things to include during the simulation. They are enclosed within triple backticks (```). Example:

```
Scenario:
{A scenario outlining a problematic situation}
Employee Personality:
{The personality traits of the employee you need to embody during the simulation}
Task:
{The objective that the participant engaging with you must achieve}
Critical aspects:
{Model response – What are the critical aspects of problem-solving that the participant must achieve to complete the task?}
Success Points:
{Model response – What key discussion points must be included for the task to be considered successfully completed?}
Failure Points:
{Model response – What are the critical aspects of problem-solving that the participant must achieve to complete the task?}
```

Your goal is to evaluate the model's response in the 7 categories defined below:

1. Success Points – Did the model correctly identify the key elements that should be present for the managerial conversation to be considered successful?
2. Failure Points – Did the model sensibly highlight the main risk factors that may lead to failure in the given scenario? Did it overlook critical threats?
3. Empathy – Did the model demonstrate an understanding of the emotional dynamics of the situation? Did it acknowledge the need for empathy toward the employee?
4. Cooperation – Did the model reflect an awareness of the collaborative nature of the problem? Did it emphasize joint responsibility, active listening, or team engagement?
5. Feedback – Did the model understand the importance of feedback in the scenario? Did it include concrete and constructive feedback strategies?
6. Structure – Was the response logically structured, clearly articulated, and easy to follow? Or was it vague, inconsistent, or overly general?
7. Context – Did the model refer specifically to the scenario context (e.g., employee traits, situation-specific details), or did it rely solely on generic statements?

Use the 5-point scale defined below:

1 – very poor – The model failed to recognize key elements, provided random or irrelevant answers, and showed no connection to the scenario.
2 – poor – The model captured some relevant points but omitted or misinterpreted key elements. The response was either vague or inaccurate.
3 – fair – The main aspects were correctly identified, but the response lacked depth or full alignment with the scenario.
4 – good – Most key points were correctly recognized, context was considered, and the response was logically structured.
5 – very good – The response was accurate, deep, and highly adapted to the scenario, incorporating subtle elements appropriately.