

Article

Not peer-reviewed version

WarehouseGame Training Framework: A Gamified Logistics Training System Integrating ChatGPT, DeepSeek and Grok for Adaptive Learning

[Juan José Romero Marras](#)*, [Luis De la Torre Cubillo](#)*, [Dictino Chaos García](#)*

Posted Date: 8 May 2025

doi: 10.20944/preprints202505.0520.v1

Keywords: gamification; logistics training; large language models; adaptive learning; Unity 3D; ChatGPT; DeepSeek; Grok



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

WarehouseGame Training Framework: A Gamified Logistics Training System Integrating ChatGPT, DeepSeek and Grok for Adaptive Learning

Juan José Romero Marras ^{*,†}, Luis De la Torre Cubillo [†] and Dictino Chaos García [†]

National Distance Education University (UNED)

* Correspondence: jromero292@alumno.uned.es

† These authors contributed equally to this work.

Abstract: Modern warehouses play a fundamental role in today’s logistics, serving as strategic hubs for the reception, storage, and distribution of goods. However, training warehouse operators presents a significant challenge due to the complexity of logistics processes and the need for efficient and engaging learning methods. Training in logistics operations requires practical experience and the ability to adapt to real-world scenarios, which can result in high training costs. In this context, gamification and artificial intelligence emerge as innovative solutions to enhance training by increasing operator motivation, reducing learning time, and optimizing costs through personalized approaches. But is it possible to effectively apply these techniques to logistics training?. This study introduces the WarehouseGame Training Framework, a gamified training tool developed in collaboration with Mecalux Software Solutions and implemented in Unity 3D. The solution integrates large language models (LLMs) such as ChatGPT, DeepSeek, and Grok to enhance adaptive learning. These models dynamically adjust challenge difficulty, provide contextual assistance, and evaluate user performance in logistics training scenarios. Through this framework, the performance of these AI models is analyzed and compared, assessing their ability to improve the learning experience and determine which one best adapts to this type of training.

Dataset: <https://doi.org/10.5281/zenodo.15002938>

Dataset License: Public

Keywords: gamification, logistics training, large language models, adaptive learning, Unity 3D, ChatGPT, DeepSeek, Grok

1. Introduction

A Statista report ([?]) suggests that by 2025, the global e-commerce sector is expected to hit around \$6. 3 trillion in sales, showcasing a compound annual growth rate (CAGR) of 11. 8% since 2021. This increase in demand creates substantial hurdles for the logistics supply chain. Continuous improvement and adjustments have become crucial in aspects such as storage, transportation, and distribution. Warehouses serve a key function in contemporary logistics by acting as centers for the effective receiving, storing, and distributing of products. The incorporation of digitalization and automation in warehouses has greatly boosted their productivity. Technologies like automated guided vehicles (AGVs), digital twins, and virtual reality have optimized decision-making, lowered errors, and enhanced efficiency ([? ? ? ?]). Nonetheless, despite these technological innovations, the human element remains vital in many warehouse logistic operations. Warehouses are crucial for various logistical responsibilities, even within highly automated settings. Repetitive activities, such as order picking and inventory management, can lead to lower motivation. The Self-Determination Theory (SDT) emphasizes that intrinsic motivation flourishes when people feel competent, autonomous, and

socially connected ([?]). This is where training tools can significantly aid operators during their learning journeys. In recent times, a variety of technologies have emerged to improve learning, including serious games, gamification, and artificial intelligence. Serious games are digital solutions specifically designed for educational and training purposes, intertwining training goals with engaging gameplay, leading to increased adoption in educational environments due to their capability to enhance learning across multiple subjects ([? ? ?]). Serious games actively involve users through engaging and challenging interactions, prompting many organizations to implement them for employee training in areas like leadership, sales, or vital operations. Another significant technology is gamification, which applies game-like elements and mechanics in settings outside of gaming, such as business and education ([?]). When gamification is incorporated into training tools, it boosts user engagement, motivation, and learning results. Lastly, artificial intelligence, especially the use of Large Language Models (LLMs), is a key change agent in education and gamification. LLMs are sophisticated neural networks built on extensive textual datasets, designed to comprehend, generate, and adapt language in context. By using architectures like transformers and attention mechanisms, these models can understand queries, produce coherent replies, and flexibly adjust information based on user interactions. Within this framework, LLM-based tools like ChatGPT 4, DeepSeek, and Grok ([? ? ?]) can enhance personalization in gamification and serious games. These models can fluidly adjust challenges and content within the platform, tailoring them to fit the skills and requirements of players ([? ?]). This research aims to assess and contrast the performance of the leading AI models available today to find out which one best fits gamification environments. Important factors for evaluation include the ability to customize challenges, accuracy in providing relevant training content, and adaptability to various user experience levels. By tailoring experiences and facilitating practical learning, the integration of LLMs.

2. Background

The incorporation of gamification, serious games, and large language models (LLMs) into the logistics sector is revolutionizing both training and everyday tasks. These innovations not only improve the efficiency and motivation of operators but also prepare companies to tackle future supply chain issues with increased flexibility and effectiveness. Several investigations ([? ?]) have looked into how serious games merge hands-on learning with interactive simulations, resulting in better outcomes for logistics education. Serious games give users the chance to learn from their mistakes without any real-life risks. For example, a [?] highlights the favorable effects of serious games on independent learning. In addition, gamification enhances logistics operations and workforce training by introducing challenges, rewards, and simulations, which stimulate operators and boost their performance. This leads to less downtime and dwindling mistakes ([? ? ?]). From an educational standpoint, gamified platforms that feature leaderboards and performance metrics increase engagement in tasks and satisfaction in developing competencies ([? ?]). LLMs also offer many advantages for training in logistics. They facilitate real-time error analysis and provide personalized improvement suggestions, as well as the design of more realistic and adaptable training scenarios. These models can fluidly respond to the unique demands of logistics training, delivering tailored feedback and guidance. Operators trained with LLMs and serious games show quicker learning rates and better skill development ([?]). Most studies conducted so far in the academic field ([1?]) focuses on how serious games can function as teaching aids for logistics education. Their studies created a game model simulating genuine supply chain scenarios, tackling issues like storage, distribution, and transport. In a similar vein, ([?]) examined over 40 games, analyzing the influence of game-based learning on supply chain management, particularly how the complexity of game design affects learning and operational results. Although these studies have offered useful insights, they have predominantly aimed at training students in educational settings. This creates some constraints, as the absence of application in actual business environments may lead to games that fail to realistically portray logistical operations or that have not been tested for practical use. Expanding upon this previous research, the current study

advances by applying these methods (serious games, gamification, and large language models (LLMs)) within a more practical context.

3. Materials and Methods

In collaboration with Mecalux Software Solutions ([?]), researchers have developed the WarehouseGame Training Framework (see Figure 1), a gamification platform specifically designed to train and enhance logistical processes. This platform was aimed to explore the effects of gamification on training in logistical processes, specifically in the simulation of order picking. In a 3D virtual environment, the platform incorporated key gamification elements such as scoring systems, and engaging visual and narrative components. WarehouseGame Training Framework was conceived with the following guiding principles:

1. Realistic and Immersive Simulation.
2. Gamification.
3. Personalized Learning.



Figure 1. WarehouseGame Training Framework.

3.1. Realistic and Immersive Simulation

The solution proposed in this paper is designed to replicate warehouse operations with a high degree of realism, accurately representing workflows, equipment, and the common challenges encountered in daily operations. This immersive environment allows trainees to familiarize themselves with warehouse processes in a risk-free setting, facilitating a smoother and more effective transition to real-world tasks. Developed using the **Unity3D** engine ([?]) (see Figure 2), renowned for its advanced graphics and simulation capabilities, incorporating essential elements such as automated storage areas, picking stations, loading and unloading docks, and sorting zones. The goal is to enhance user immersion, making the training experience more engaging and effective while fostering a deeper understanding and better adaptation to real-world scenarios. The simulated warehouse was based on realistic customer cases provided by Mecalux Software Solutions. Key elements of warehouse work were included, such as radio frequency terminals, shelf tag scanning, and product barcode scanning, reflecting common tasks in these environments.

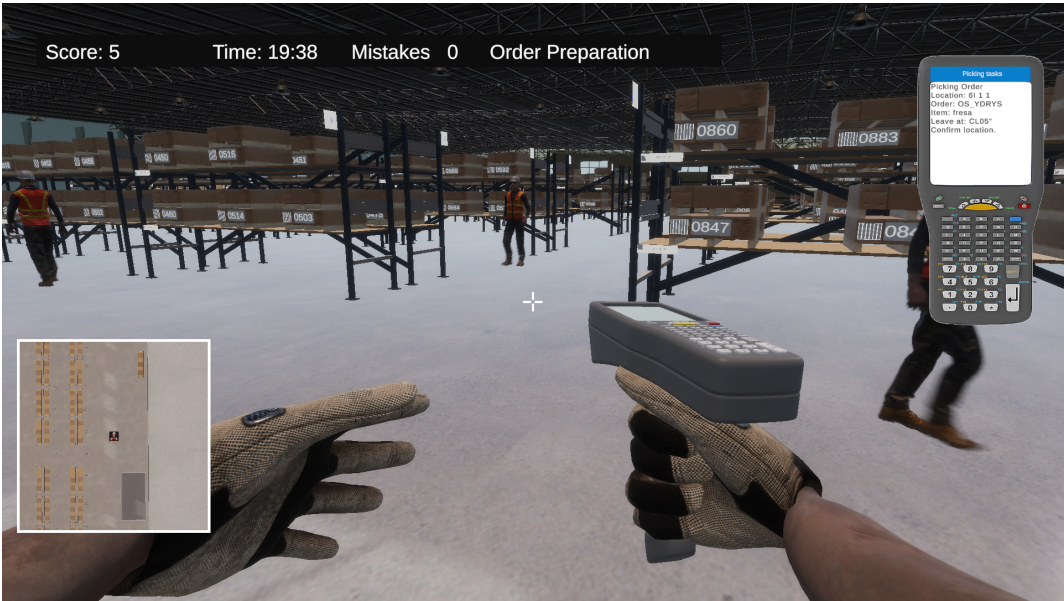


Figure 2. Unity engine.

A standout feature is the simulation of a real warehouse management system (WMS) on the radio frequency terminal. Players receive instructions through the terminal as if they were interacting with an actual WMS, enabling them to perform tasks as if they were working in a real warehouse. The simulated WMS software used in this framework is Easy WMS ([?]), a leading solution in warehouse management. The simulation is designed to encompass a wide variety of logistical tasks. The activities included order preparation (see Figure 3), material receiving, material placement or forklift operations. This approach ensures comprehensive skill development, enabling employees to perform various roles and assume multiple responsibilities within the warehouse.



Figure 3. Order picking

Each task category is designed with multiple difficulty levels, tailored to the player’s skills and progress. For example, in forklift operation tasks (see Figure 4), the initial challenges focus on basic movements, while more advanced levels require precise coordination. Similarly, order preparation tasks start with simple activities and progress to more complex scenarios, involving multiple simultaneous orders and time constraints.



Figure 4. forklift operation tasks

3.2. Gamification

WarehouseGame Training Framework has been designed to maximize player motivation and engagement through the Gamification. Research indicates that gamification enhances task engagement, skill acquisition, and knowledge retention ([? ? ?]). However, its use in logistics training remains underexplored, underscoring the need for strategies tailored to individual profiles to maximize real-world impact ([? ? ?]). Various gamification elements carefully integrated into the learning experience (see Figure 5). These elements aim not only to make the process more dynamic and entertaining but also to strengthen the technical and operational competencies required in a real logistics environment. This approach encourages continuous improvement, enabling users to track their progress and strive for better results in each task category, from material reception to forklift operation. Additionally, each challenge is designed with time constraints, fostering speed and precision in task execution—key aspects in the logistics context.



Figure 5. Gamification's elements

Among the standout features is a points and levels system that rewards players for their successes and penalizes their mistakes. At the end of each challenge, the player’s results are displayed (see Figure 6).

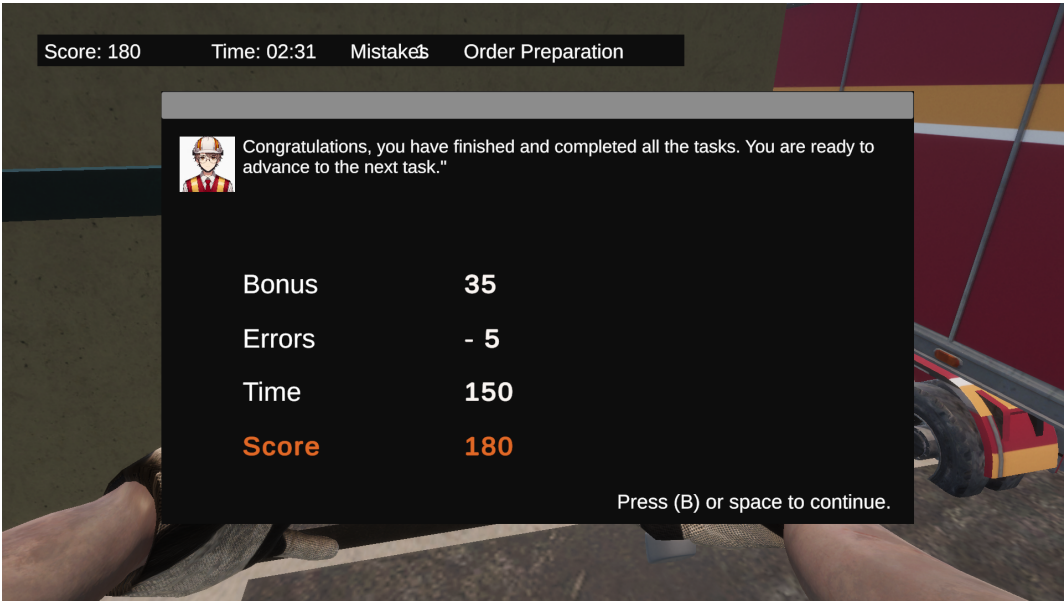


Figure 6. Player’s results

Another improvement have been incorporated: an expanded narrative (see Figure 7). The expanded narrative provides detailed explanations of the challenges and concepts related to warehouse operations, offering clearer and more engaging context that enhances the understanding of logistics processes.



Figure 7. Narrative

3.3. Personalized Learning

One of the key innovations is the ability to customize the training using Artificial Intelligence (AI). The incorporation of advanced AI models, such as **Large Language Models (LLMs)**, of this work is the feature that truly distinguishes this study. Through the simulation, players can select from various AI models.

1. **ChatGPT** ([?]): Developed by OpenAI using the GPT-4 architecture, stands out for its versatility and ease of use. It excels in tasks such as creative writing, coding assistance, and general inquiries. Its chain of thought reasoning capability allows it to generate detailed and contextually relevant responses. However, it has limitations in accessing up-to-date information, as it cannot perform real-time web searches. GPT-4 is the model that has been employed throughout the present study.
2. **DeepSeek** ([?]): Developed by the Chinese startup of the same name, is a language model based on a Mixture-of-Experts (MoE) architecture. Its latest version, DeepSeek-V3, features 671 billion parameters, activating 37 billion per token to enable efficient inference through Multi-head Latent Attention (MLA). The model has surpassed ChatGPT in downloads on the App Store and has been adopted by companies such as Great Wall Motor. DeepSeek-V3 model has been consistently utilized throughout the development of this research.
3. **Grok** ([?]): Developed by xAI, the artificial intelligence company founded by Elon Musk in 2023, is a chatbot designed to answer questions and provide broad insights. Its latest version, Grok 3, is noted for its advanced reasoning capabilities, outperforming competitors like ChatGPT. Additionally, xAI has integrated source citations into responses and introduced multimodal models, such as grok-2-vision-1212, which enhance accuracy and instruction-following capabilities across multiple languages. While Grok 3 had been released during the course of this study, it was not yet accessible via API. Consequently, this research was conducted using the previous version, Grok 2.

These models allow for the personalization of the training plan based on each player’s initial skill level. The process begins with an initial survey consisting of two questions per category (see Figure 8), aimed at determining the player’s level. Based on the responses provided, the AI evaluates the skill level and begins designing the challenges.

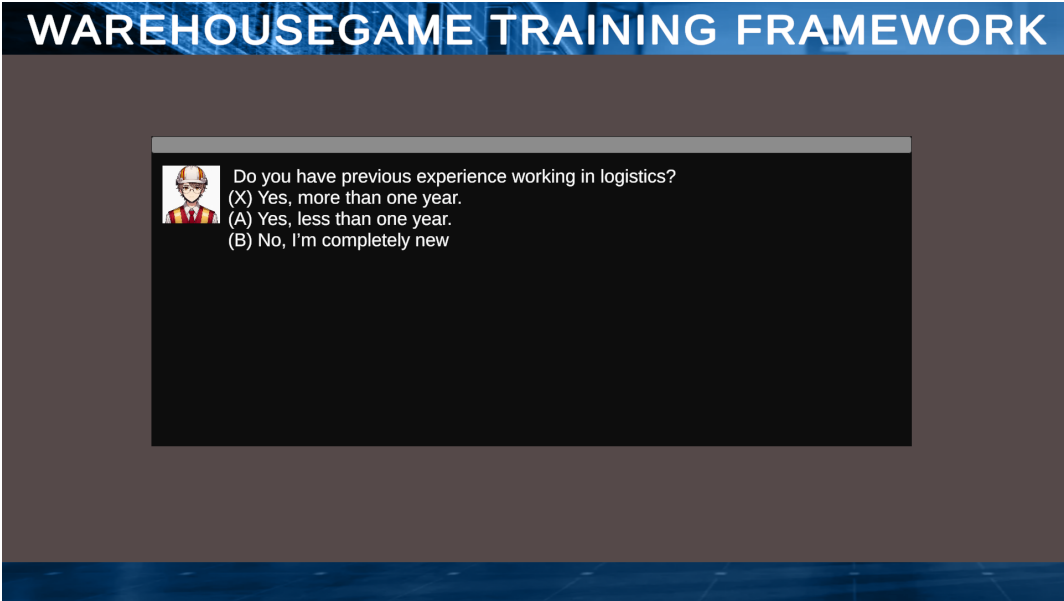


Figure 8. Initial Survey

Once the player’s skill level has been assessed, the artificial intelligence designs a series of progressive challenges aimed at guiding the user toward achieving expert proficiency in each category. Initially, the AI generates a challenge tailored to the player’s current level and, based on their performance, adjusts and redefines subsequent challenges. If the player successfully completes a challenge, the AI gradually increases the difficulty of the next one within the same category. For instance, if a player is classified as a beginner in the order preparation category, the AI will design a challenge appropriate for that level. If the player successfully completes the challenge, a new, more difficult challenge (Medium level) will be generated, and this process will continue until the player completes the expert-level challenge. Once the expert level is reached, the simulation will progress to the next

category, such as material reception. Additionally, the AI incorporates feedback mechanisms and dynamic difficulty adjustment. If the player makes mistakes during a challenge, the AI may provide additional assistance (see Figure 9) in the form of hints or reduce the complexity of the task. In cases where the player fails to complete the challenge successfully, the AI will generate a new version of the challenge at the same difficulty level, incorporating modifications that facilitate learning, such as increased guidance. This adaptive approach ensures a personalized and optimized progression for each player, promoting effective learning tailored to their performance.

A demonstration of the WarehouseGame Training Framework can be viewed in the following video [?].

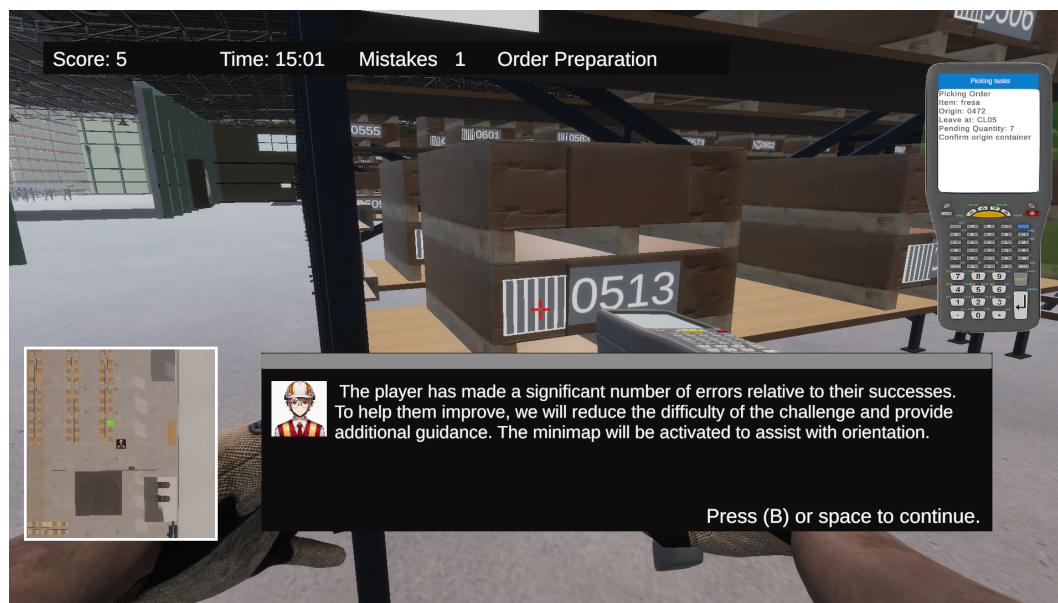


Figure 9. AI assistance

4. Implementation

To enable personalization, a communication protocol has been implemented using the APIs of LLMs. The AI models share a common communication model based on a REST architecture. In this architecture, the client sends HTTP requests to the API endpoints exposed by the respective servers. Any application that interacts with the API can function as a client, including chatbots, virtual assistants, or a Unity C# application such as the WarehouseGame Training Framework. Each platform provides extensive documentation, including sample code in multiple programming languages, simplifying the integration process. Before our software can communicate with any of these APIs, several preliminary steps must be completed:

1. **Account Creation:** A user account must be created on each platform.
2. **API Key Generation:** An API key must be generated to authenticate and secure communication between our software and the API.
3. **Credit Balance:** Since API usage is not free, it is necessary to add funds to the account. Each platform applies different pricing based on the model and number of requests. If the user lacks sufficient credit, the service will not process requests.

Once these steps are completed, our software can establish communication with the selected API and process the responses accordingly. During interaction, the application sends requests based on user input, processes the JSON response returned by the API, and interprets it to provide real-time feedback through the 3D interface. The communication flow is straightforward (see Figure 10):

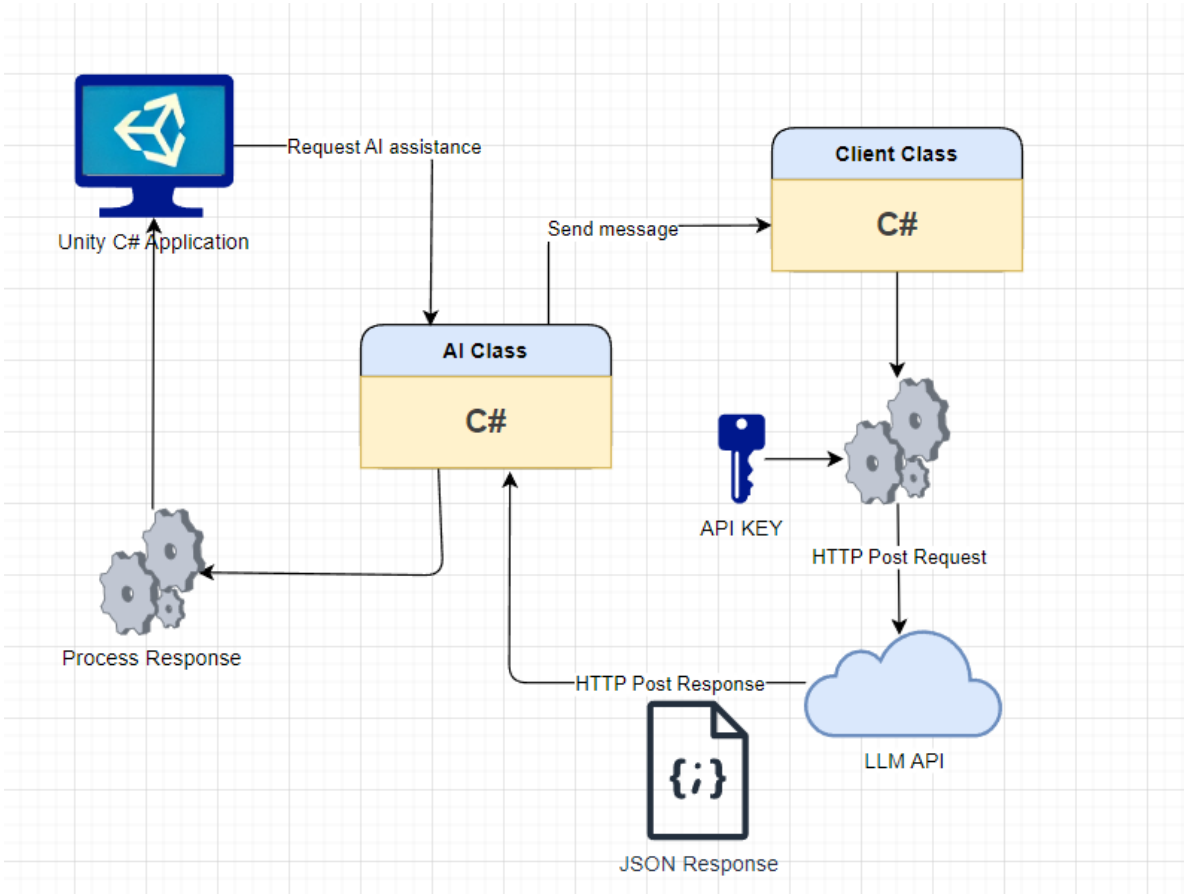


Figure 10. Communication architecture

1. **Request AI assistance:** From the Unity-based application, which serves as a 3D simulator, requests are sent to the AI for assistance in various aspects of training. These requests are handled by the corresponding class within the application, depending on the AI model used, ensuring efficient and adaptive communication with the selected API.
2. **Send message:** Once the request type has been identified, the system constructs the message to be sent to the AI. There are several types of messages. Each message consists of two parts: a description of the request’s objective (see Table A1) and a specification of the expected response format (see Table A2). The first part is a textual instruction directed at the AI, while the second part is a standardized JSON structure, ensuring that the Unity application can correctly interpret and process the response.
3. **HTTP Post Request:** Once the message is constructed, the next step is to send a POST request to the LLM server. This request must follow a specific structure to ensure that the server can correctly process it. The message must be encapsulated in a format that complies with the API’s requirements. Fortunately, as previously mentioned, all LLM models used share architectural similarities, making the POST request structure identical across all of them. The required format is as follows:

```
{
  "model": "",
  "messages": [
    {"role": "system", "content": ""},
    {"role": "user", "content": ""}],
  "temperature",
  "max_tokens"
}
```

Where:

- **model:** specifies the language model to be used.
- **messages:** contains the structured conversation.
- **role: 'system':** defines the AI's general behavior.
- **role: 'user':** represents the message sent by the user.

Once this structure is prepared, the HTTP client sends a POST request to the API endpoint: `https://server_llm/v1/chat/completions`. This endpoint is consistent across all models, with only the server name changing depending on the LLM provider.

4. **Http Post Response:** The server processes the request, executes the language model, and generates a response based on the conversation history and the defined parameters. The response is returned in JSON format with the following structure:

```
{
  "id": "id",
  "model": "model",
  "choices": [{
    "index": 0,
    "message": {
      "role": "role",
      "content": "content"
    }
  }]
}
```

The 'content' field contains the response generated by the AI. This response adheres to the JSON format previously specified in the request sent to the model. The specific JSON structure varies depending on the type of response, as detailed in Table A2.

5. **Process JSON Response:** Finally, the AI class in Unity deserializes the response and processes its content for interpretation and use within the simulation environment. At this stage, the system analyzes the AI-generated response and executes the appropriate actions based on the type of request. For instance, if the response pertains to an in-progress challenge assistance request, the system dynamically adjusts the help parameters to provide additional guidance to the user. If the request involves generating a new challenge, the system utilizes the AI-provided information to configure and present a tailored task suited to the player's needs. Moreover, the AI response is displayed to the user within the Unity 3D environment through the simulator interface, offering a detailed explanation that seamlessly integrates with the training context (see Figure 9). This ensures that users can fully understand the recommendations or adjustments made in real-time, thereby enhancing the learning experience.

The source code of the implementation developed and described in this work is available on Zendo [?], where it can be accessed for review and use.

5. Experimental Design for Evaluating LLM Performance.

This experiment aims to assess the performance of different Large Language Models (LLMs) within the WarehouseGame Training Framework, focusing on their ability to generate, adjust, and optimize logistics challenges based on user performance and skill level. Simulated user profiles, categorized by skill level (Beginner, Medium, Advanced and Expert) in order preparation (picking) tasks. Evaluation objectives:

1. **Evaluation of Difficulty Progression.**
2. **Evaluation of Model Adaptability.**
3. **Evaluation of Response Compliance.**
4. **Evaluation of Response Time.**

This methodology allows for a structured evaluation of LLMs, identifying the most effective models for adapting warehouse training scenarios.

5.1. Evaluation of Difficulty Progression

The LLMs were evaluated based on their ability to dynamically adjust the difficulty of challenges according to user performance. An effective progression system ensures a balanced learning curve, preventing both frustration from excessive difficulty and disengagement due to lack of challenge. The evaluation focused on the following criteria:

1. Generating challenges appropriate to the user's current skill level.
2. Gradually increasing difficulty when the user demonstrates consistent mastery.
3. Preventing difficulty escalation if the user has not yet corrected previous errors.

5.2. Evaluation of Model Adaptability

The adaptability of each LLM was assessed to determine its ability to personalize training experiences in real-time. A well-adapted model should ensure that challenges evolve logically based on user progress, promoting an effective and engaging learning process. The evaluation focused on the following criteria:

1. Dynamically adjusting challenges according to user performance.
2. Retain user history for logical progression.
3. Identifying recurring errors and adapting challenge complexity to prevent frustration and reinforce learning.

5.3. Evaluation of Response Compliance

The evaluation of each LLM focused on its ability to generate responses that strictly adhere to the predefined communication protocol. Ensuring compliance with the expected response format is essential for seamless integration with the WarehouseGame Training Framework and to prevent system errors. The assessment criteria included:

1. JSON Format Compliance: The response must strictly follow the expected JSON structure to ensure correct processing by the software.
2. Data Integrity: All required fields must be present and correctly structured within the response.
3. Error Handling: The model should not generate malformed responses or omit critical information that prevents interpretation by the system.

Maintaining response compliance ensures that the AI-generated outputs are functional, reliable, and seamlessly integrated into the training framework.

5.4. Evaluation of Response Time

The response speed of each LLM was analyzed to assess its ability to provide real-time feedback without perceptible delays. A fast response time is crucial for maintaining a seamless and engaging user experience, as excessive delays can disrupt the learning flow and negatively impact user interaction.

6. Results

6.1. Evaluation of Difficulty Progression

The difficulty progression of each LLM was evaluated based on its ability to gradually increase challenge complexity, with 50 iterations per level (see Table A3). **Grok** demonstrated the most structured and consistent difficulty progression, with orders (1 → 3) and tasks (5 → 20) scaling logically, multi-reference containers introduced at the Medium level, failures reduced (3 → 2), and assistance removed by advanced/expert levels. It demonstrated high consistency (72.22% Medium, 87.72% expert). However, 41.34% of advanced-level responses were misclassified as Medium, indicating significant level response errors that could disrupt difficulty alignment. **ChatGPT** provided satisfactory

progression, with orders (1 → 3), tasks (5 → 15), multi-reference containers at the advanced level, and assistance phased out by the expert level. It retained three failures at the advanced level (55.32%), potentially hindering mastery, and assistance until advanced (53.82%). Consistency was lower (36% beginner, 29.17% Medium), but no level response mismatches were observed, ensuring challenge alignment. **DeepSeek** displayed irregularities, with an abrupt task increase (10 → 12) at the Medium level, no multi-reference containers at this stage, and lenient time limits at the expert level (25s). It was highly consistent at beginner (100%) and expert (94%) levels but less so at Medium (38%) and advanced (60%) levels. No level response mismatches occurred, a notable strength. Based on the evaluation, **ChatGPT is the most suitable model**, with potential adjustments to failure tolerances and assistance removal to enhance training efficacy.

6.2. Evaluation of Model Adaptability

The analysis of the results demonstrates how each model adapted the difficulty when players reached the maximum number of allowed failures. Evaluation criteria encompassed reductions in orders and tasks, increases in allowed failures, and reactivity, quantified through the Average Error Ratio (Errors/Failures × 100) and Average Difference (Errors - Failures). Tables A4, A5, and A6 present the simulation results, detailing, for each model, level, and challenge, the adjusted challenge designed by the AI and the consistency percentage (% of agreement), based on 50 iterations per model/level. **ChatGPT** implemented substantial reductions in orders (from 2–3 to 1 at Medium and advanced levels) and tasks (from 15 to 5–12), increasing allowed failures to 10–12. However, it exhibited variable consistency (13.04–100%) and delayed reactivity (ratios: 128.43–424.69; differences: –0.84 to –7.22), indicating tardy adjustments. **DeepSeek** reduced tasks (from 12–15 to 8–10) and increased failures (to 5–8, 100% consistency in most cases) but retained orders at the Medium level (2 in 80.95% of cases), limiting adaptability. Its reactivity was moderate (ratios: 100–381; differences: 0 to –2.98). **Grok** balanced task reductions (from 15–20 to 10–15) and failure increases (to 3–5), with high consistency (100% at advanced). Its reactivity was optimal (ratios: 89.33–200; differences: 0.32 to –2), particularly at the advanced level (ratio: 101.75; difference: 0). Overall provides aggressive adjustments, but its slow reactivity compromises user experience, DeepSeek is consistent but limited in adaptability due to minimal order reductions and Grok integrates balanced adjustments with rapid reactivity, minimizing player frustration. **Grok is the most effective model, owing to its precise reactivity and well-calibrated adjustments, optimizing the learning experience.**

6.3. Evaluation of Response Compliance

The evaluation of response compliance revealed that all tested models successfully adhered to the predefined communication protocol. Each LLM generated responses that strictly followed the expected JSON structure, ensuring seamless integration with the WarehouseGame Training Framework. Throughout the assessment, no errors were detected in the interpretation of requests, and all models correctly understood the context of the queries. The generated responses consistently included all required fields in the expected format, demonstrating full compliance with data integrity and error-handling criteria. This result confirms the reliability of these models in structured data generation tasks, particularly within controlled environments that require strict adherence to predefined formats. The ability of each LLM to consistently provide well-formed responses reduces the risk of system errors and facilitates automated processing within the training framework. In summary, the tested models have demonstrated a high level of accuracy in response formatting and compliance, validating their suitability for structured interactions in gamified training applications.

6.4. Evaluation of Response Time

The response time of Large Language Models (LLMs) is a critical factor influencing the overall user experience, particularly in real-time training environments such as the WarehouseGame Training Framework. The response time directly affects the fluidity of interactions, as excessive delays may lead to user frustration and disrupt the learning process. Table A8 presents the response time statistics

(in milliseconds) for each LLM evaluated. Grok demonstrated the best performance in terms of response speed. With a mean response time of 1924.86 ms and a median of 1852.0 ms, the model responded consistently and quickly. Its standard deviation (299.42 ms) indicates minimal variability across interactions, and the 95% confidence interval (1883.10 ms to 1966.61 ms) confirms the model's stability. Furthermore, its minimum (1384 ms) and maximum (3235 ms) response times remained within acceptable thresholds, reinforcing its suitability for real-time interaction. ChatGPT provided moderate performance. Its mean response time was 5823.45 ms, with a median of 5578.0 ms. The standard deviation (1415.13 ms) suggests higher variability, occasionally resulting in longer waiting times. The 95% confidence interval (5626.12 ms to 6020.77 ms) shows general consistency, but the range between its minimum (3313 ms) and maximum (11956 ms) values highlights the potential for noticeable delays. While ChatGPT can still be considered viable for real-time training, these peaks in response time may interrupt the user experience in more dynamic scenarios. DeepSeek exhibited the longest response times among the evaluated models. With a mean of 10213.75 ms and a median of 10098.5 ms, the model consistently responded more slowly. Although the standard deviation (718.28 ms) was smaller than ChatGPT's, its base latency was significantly higher. The 95% confidence interval (10113.59 ms to 10313.91 ms) indicates consistent delay, and the minimum (8513 ms) and maximum (12836 ms) response times confirm performance issues that could compromise real-time usability in an interactive learning environment. When comparing the three models:

- Grok clearly outperformed the others, offering the lowest latency and highest consistency, making it ideal for seamless integration in real-time educational platforms.
- ChatGPT provided acceptable performance with occasional delays, suitable for applications where slight lags are tolerable.
- DeepSeek, while functional, presented response times that may hinder interactivity and user engagement in scenarios requiring prompt feedback.

Overall, **Grok emerged as the most efficient model in terms of response time**, delivering fast and reliable outputs suitable for interactive training applications. Its combination of low latency, minimal variance, and consistency underlines its advantage for real-time adaptive learning systems like WarehouseGame. While ChatGPT remains a strong alternative, further optimization would be required to match Grok's responsiveness. DeepSeek, on the other hand, would need substantial improvements in response latency to meet the demands of such environments.

7. Discussion

The comprehensive evaluation of three large language models (LLMs) (ChatGPT, DeepSeek, Grok) within the WarehouseGame Training Framework provides clear insights into their suitability for optimizing logistics simulation, considering four critical dimensions: difficulty progression, adaptability, response compliance, and response time. Each model exhibited distinct strengths and limitations that shape their applicability in this context. In difficulty progression, ChatGPT excelled by generating challenges consistently aligned with the requested level, ensuring a coherent experience across all skill levels, from beginner to expert. This consistency is vital for maintaining a clear learning curve. However, its lenient failure tolerances and prolonged assistance in advanced levels could hinder the development of autonomous skills. Grok delivered a well-structured progression, logically increasing complexity, but its reliability was undermined by alignment errors between requested and responded levels, particularly at the advanced stage. DeepSeek exhibited less refined scalability, with abrupt difficulty transitions that could disorient players, especially at Medium levels. Regarding adaptability, Grok proved superior in dynamically adjusting challenges based on user performance. Its rapid response to excessive errors, through balanced difficulty reductions and increased failure allowances, minimized frustration and fostered a fluid learning environment. ChatGPT, while implementing more aggressive adjustments, suffered from slower reactivity, potentially prolonging players' exposure to overly challenging tasks. DeepSeek, despite its consistency, was less effective due to insufficient order reductions at Medium levels, limiting its personalization capabilities. For response time, Grok emerged

as the clear leader, delivering fast and consistent responses essential for maintaining interactivity in a real-time training environment. ChatGPT performed acceptably but exhibited occasional delays that could disrupt fluidity in dynamic scenarios. DeepSeek, with significantly longer response times, was less suitable for applications requiring immediate feedback. **Grok is the most suitable model for gamified simulation**, excelling in adaptability and response time, and offering robust progression despite alignment issues. **ChatGPT is a strong alternative, particularly for lower levels**, but requires improvements in reactivity and rigor. **DeepSeek is not recommended** due to its irregular scalability and high latency, which compromise user experience. However, **a hybrid approach, using ChatGPT for lower levels and Grok for higher levels, could maximize training effectiveness**. ChatGPT would be ideal for beginner and Medium levels, where its perfect level alignment ensures a clear and coherent introduction to logistics tasks. As players progress to advanced and expert levels, Grok could take over, leveraging its superior adaptability and fast response times to manage more complex, personalized challenges, provided its alignment errors are addressed.

8. Future Work and New Research Directions

While this research has demonstrated the potential of LLM models to dynamically adjust difficulty in a gamified logistics environment, it is important to acknowledge certain limitations that should be addressed in future studies:

- Diversity of evaluated profiles: The study focused on a specific group of users, without considering a wide range of profiles with varying levels of experience and training in logistics. Evaluating LLM performance with a more diverse sample would validate their adaptability to different learning styles and competency levels.
- Exploration of other language models: The research was limited to a specific set of LLM models (Grok, ChatGPT, and DeepSeek). Alternatives such as LLaMA (Meta) or other emerging models, which could offer improvements in adaptability, efficiency, and personalization of the learning experience, were not assessed.
- Application to other warehouse tasks: The study concentrated on order picking, without extending its analysis to other key areas of the logistics environment, such as receiving, material sorting, or inventory management. Exploring the applicability of LLMs to these processes would allow for an evaluation of their impact across a broader range of warehouse operations.

Recognizing these limitations provides a foundation for future research and optimizations, enabling the enhancement of LLM integration in logistical environments and expanding their applicability to more complex and realistic scenarios. Future research should focus on refining the integration of LLMs into gamified logistics environments, improving their ability to personalize the learning experience based on real-time user performance. One of the primary aspects to optimize is the adjustment of error tolerance and the adaptive scalability of difficulty, ensuring smoother and more effective progression. Likewise, enhancing the structure and timing of visual and contextual aids could increase the effectiveness of the training process. Another relevant line of research is the use of virtual and mixed reality environments in combination with LLM models to enhance immersion and learning effectiveness. Simulation in a more realistic setting could improve the transfer of knowledge to practical scenarios, optimizing training in logistical environments. Another key direction for future research is the application of this system in a real-world setting, transferring the dynamics developed in the game to an operational warehouse with actual order-picking processes. This approach would allow for an evaluation of the impact of AI-driven gamification under real working conditions, validating its effectiveness in optimizing logistical tasks. Furthermore, incorporating a collaborative multiplayer environment within the game could strengthen learning through user cooperation, fostering teamwork dynamics and problem-solving in gamified logistics settings. Addressing these areas will contribute to enhancing the role of AI-powered gamification in workforce training, making logistics learning more efficient, engaging, and tailored to the individual needs of each user.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Juan José Romero Marras, Dr. Luis De la Torre Cubillo, Dr. Dictino Chaos García; development, testing and data collection: Juan José Romero Marras; analysis and interpretation of results: Juan José Romero Marras, Dr. Luis De la Torre Cubillo, Dr. Dictino Chaos García; Juan José Romero Marras wrote the manuscript with support from Dr. Luis De la Torre Cubillo and Dr. Dictino Chaos García. All authors reviewed the results and approved the final version of the manuscript..

Funding: This research was funded by the Spanish Ministry of Science and Innovation through a research project with reference PID2022-139187OB-I00.

Appendix A

Table A1. Message Descriptions and Prompt Message

Message Description	Prompt Request Example
Before starting, the AI is informed of the player’s level, and a challenge is requested.	<i>The player is about to undertake order preparation challenges. Their level in the Order Preparation category is Begginer. I need a challenge appropriate for their level so they can progress and reach expert level.</i>
Once the challenge is completed, the results are sent to the AI for evaluation and to generate the next challenge. Regardless of the AI model used, the messages sent remain the same for challenge design, adjustment, and evaluation.	<i>Analyze the results obtained by the player in the challenge. The player had to complete the following challenge: ‘Orders’: 1, ‘Tasks’: 5, ‘Multi-reference’: false, ‘Level’: ‘Beginner’, ‘Time’: 500, ‘Errors’: 2, ‘Assistance’: true, ‘MiniMap’: false. The results were: Errors: 5 and the Time spent: 10 minutes. Knowing that the player’s current level in the Order Preparation category is Beginner, I need you to determine whether they have successfully completed the challenge.</i>
If the player makes a mistake, the AI is notified with details about where the error occurred, the challenge the player was attempting, and a request for an assessment of the mistake.	<i>The player has just made an error while scanning the location of the container. The scanned label is incorrect. I want you to analyze whether it is necessary to assist the player. The assistance you can provide falls into three categories: A) Add more in-game assistance: In this case, you could activate the minimap and highlight the location to help the player orient themselves better. B) In addition to the above, if the player has made too many mistakes, reduce the difficulty of the challenge. This means adjusting the challenge to a lower level. C) Do nothing: If the number of mistakes is not high enough compared to the difficulty of the challenge and the player’s level, no adjustments are necessary.</i> <i>The player has just made a mistake picking the product from the container. The product and the quantity are not correct. They should have picked a quantity of 10 of the product picking.Stock. But that has not been the case. This may be due to several causes:1. They have placed the product in the wrong client container; they should have selected the client container CL01. 2. They have not picked the selected product. 3. They have not picked the full quantity for the indicated product. To perform the picking, the player must: 1.Select the client container using the F1, F2, or F3 keys (they must check the radio frequency terminal to know what to select). 2.Move the hand with the radio frequency terminal using the cursor to the container/product indicated in the task. 3. Select the box with the product (it will appear in violet when selected). 4. Confirm with the mouse. Next, I will provide you with the data of the challenge the player is currently performing. The player had to complete the following challenge:‘Orders’: 1, ‘Tasks’: 10, ‘Multi-reference’: false, ‘Level’: Beginner, ‘Time’: 15, ‘Mistakes’: 3, ‘Help’: true, ‘MiniMap’: true.</i>

Table A2. Message Descriptions and JSON Response

Message Description	JSON Response
Before starting, the AI is informed of the player's level, and a challenge is requested.	<pre>{ 'Orders', 'Tasks', 'MultiReference', 'Level', 'Time', 'Failures', 'Explanation', 'Help', 'MiniMap' }</pre>
Once the challenge is completed, the results are sent to the AI for evaluation and to generate the next challenge. Regardless of the AI model used, the messages sent remain the same for challenge design, adjustment, and evaluation.	<pre>{ 'OvercomeChallenge', 'AdjustChallenge', 'Orders', 'Tasks', 'MultiReference', 'Level', 'Time', 'Failures', 'Explanation', 'Help', 'MiniMap' }</pre>
If the player makes a mistake, the AI is notified with details about where the error occurred, the challenge the player was attempting, and a request for an assessment of the mistake.	<pre>{ 'OptionHelp' 'AdjustChallenge' 'Orders', 'Tasks', 'MultiReference', 'Level', 'Time', 'Failures', 'Explanation', 'Help', 'MiniMap' }</pre>

Table A3. Evaluation of Difficulty Progression

Model	Level	Level Response	Orders	Tasks	MultiRef	Time	Failures	Help	%
ChatGPT	Beginner	Beginner	1	5	0	15	3	1	36.00
		Beginner	2	5	0	20	5	1	16.00
		Beginner	1	5	0	15	5	1	14.00
		Beginner	2	5	0	15	3	1	10.00
		Beginner	2	5	0	20	3	1	10.00
	Medium	Medium	2	15	0	25	3	1	29.17
		Medium	2	10	0	20	3	1	22.92
		Medium	2	15	0	20	3	1	14.58
		Medium	2	10	0	20	3	1	10.42
		Medium	2	15	0	25	3	1	8.33
	Advance	Advance	2	15	1	25	3	1	31.91
		Advance	2	15	1	25	3	0	21.28
		Advance	2	15	1	25	2	1	19.15
		Advance	2	15	1	25	2	0	12.77
		Advance	3	15	1	25	3	0	2.13
	Expert	Expert	3	15	1	25	2	0	50.91
		Expert	3	15	1	25	3	0	10.91
		Expert	3	15	1	30	2	0	7.27
		Expert	2	15	1	25	2	0	7.27
		Expert	2	15	1	25	3	0	5.45
DeepSeek	Beginner	Beginner	1	5	0	10	3	1	100.0
	Medium	Medium	2	12	0	15	3	1	38.0
		Medium	2	12	0	15	3	1	30.0
		Medium	2	10	0	15	3	1	18.0
		Medium	2	10	0	15	3	1	14.0
	Advance	Advance	2	15	1	25	2	1	60.0
		Advance	2	15	1	25	2	0	26.0
		Advance	2	15	1	20	2	0	4.0
		Advance	3	15	1	25	2	1	4.0
		Advance	2	12	1	20	2	1	2.0
	Expert	Expert	3	15	1	25	1	0	94.0
		Expert	3	15	1	25	2	0	6.0
Grok	Beginner	Beginner	1	5	0	10	3	1	58.00
		Beginner	1	5	0	15	3	1	42.00
	Medium	Medium	2	15	1	25	3	1	72.22
		Medium	2	15	0	20	3	1	16.67
		Medium	2	10	1	20	3	1	5.56
		Medium	2	15	0	25	3	1	5.56
	Advance	Advance	3	15	1	25	2	0	58.67
		Medium	2	15	1	25	3	1	22.67
		Medium	2	15	1	25	2	1	8.00
		Medium	2	15	1	25	3	0	8.00
		Medium	2	15	1	25	2	0	2.67
	Expert	Expert	3	20	1	30	2	0	87.72
		Advance	3	15	1	30	2	0	10.53
		Medium	2	15	1	25	2	0	1.75

Table A4. Evaluation of Model Adaptability - ChapGPT

Model	Level	Orders	Taks	Failures	New Ordenes	New Taks	New Failures	%
ChatGPT	Beginner	1	5	3	1	3	5	90.91
		1	5	5	1	3	10	36.36
		1	5	5	1	3	6	27.27
		2	5	3	1	3	8	40.00
		2	5	3	1	3	10	40.00
		2	5	4	1	3	5	100.00
		2	5	5	1	3	8	45.45
		2	5	5	1	3	5	18.18
	Medium	2	10	3	1	5	10	57.14
		2	10	3	1	5	8	21.43
		2	12	3	1	8	8	50.00
		2	12	3	1	8	10	50.00
		2	15	3	1	10	10	50.00
		2	15	3	1	10	8	15.62
	Advance	2	15	5	1	10	5	100.00
		2	15	2	1	10	5	27.27
		2	15	2	2	12	6	18.18
		2	15	2	2	10	5	18.18
		2	15	3	2	10	8	26.09
		2	15	3	2	12	8	17.39
		2	15	3	1	10	10	13.04
		3	15	2	2	10	5	50.00
		3	15	2	2	10	10	50.00
		3	15	3	2	12	10	50.00
		3	15	3	2	10	12	50.00
	Expert	2	15	2	2	10	6	40.00
		2	15	2	1	10	6	20.00
		2	15	3	1	10	8	50.00
		2	15	3	2	10	10	50.00
		3	15	2	2	10	8	46.15
		3	15	2	2	12	8	15.38
		3	15	3	2	10	5	83.33
		3	15	3	2	10	6	16.67
		3	18	2	2	15	7	100.00

Table A5. Evaluation of Model Adaptability - DeepSeek

Model	Level	Orders	Taks	Failures	New Ordenes	New Taks	New Failures	%
DeepSeek	Beginner	1	5	3	1	3	5	100.00
	Medium	2	10	3	2	8	8	28.57
		2	10	3	2	8	7	19.05
		2	10	3	2	7	8	19.05
		2	10	3	1	8	8	14.29
		2	12	3	2	8	8	62.07
		2	12	3	1	8	8	37.93
	Advance	2	12	2	1	8	6	100.00
		2	15	2	1	10	6	58.70
		2	15	2	2	10	6	23.91
	Expert	3	15	1	2	10	5	100.00
		3	15	2	2	10	8	100.00

Table A6. Evaluation of Model Adaptability - Grok

Model	Level	Orders	Taks	Failures	New Ordenes	New Taks	New Failures	%
Grok	Beginner	1	5	3	1	3	5	66.00
		1	5	3	1	3	2	18.00
		1	5	3	1	3	4	16.00
	Medium	2	15	3	2	10	5	38.71
		2	15	3	1	10	5	29.03
		2	15	3	1	10	3	22.58
		3	15	2	2	10	5	94.00
		3	15	2	2	10	4	6.00
	Advance	2	15	2	2	10	3	100.00
		2	15	3	2	10	5	88.89
	Expert	2	15	3	2	10	4	5.56
		3	20	2	2	15	5	70.00
		3	20	2	2	15	3	26.00

Table A7. Relationship between actual user errors and the allowed error threshold per level

Model	Level	Avg Error Ratio	Std. Dev	Mean Difference	Std. Diff
ChatGPT	Beginner	128.43	35.73	-0.84	1.18
	Medium	220.82	33.76	-3.67	1.07
	Advance	312.93	77.30	-5.12	1.64
	Expert	424.69	74.73	-7.22	1.19
DeepSeek	Beginner	100.00	0.00	0.00	0.00
	Medium	154.00	18.92	-1.62	0.57
	Advance	193.00	17.53	-1.86	0.35
	Expert	381.00	52.38	-2.98	0.14
Grok	Beginner	89.33	18.37	0.32	0.55
	Medium	126.13	31.28	-0.48	0.71
	Advance	101.75	26.00	0.00	0.58
	Expert	200.00	14.29	-2.00	0.29

Table A8. Summary of Response Time Statistics

Model	Mean	Median	Std. Deviation	Minimun	Maximun	95% Confidence Interval
ChatGPT	5823.445	5578.0	1415.125	3313	11956	(5626.122, 6020.767)
DeepSeek	10213.75	10098.5	718.281	8513	12836	(10113.593, 10313.906)
Grok	1924.855	1852.0	299.421	1384	3235	(1883.104, 1966.605))

References

1. Kersten, W.; Blecker, T.; Ringle, C. Can Gamification Reduce the Shortage of Skilled Logistics Personnel? *TORÉ* **2024**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.