# Large Language Models for Hospitality Workforce Training: Design and Validation of an Intelligent Agent System

Pablo Vicente-Martínez *, Diego Lacomba-Fañanás, Emilio Soria-Olivas, Manuel Sánchez-Montañés, María Ángeles García Escrivà *, Edu William-Secin

Posted Date: 18 June 2026

doi: 10.20944/preprints202606.1402.v1

Keywords: large language models (LLMs); hospitality workforce training; artificial intelligence; intelligent agents; automated evaluation; Gemini 2.0; educational technology

 doi:10.20944/preprints202606.1402.v1



*Article*

# Large Language Models for Hospitality Workforce Training: Design and Validation of an Intelligent Agent System

**Pablo Vicente-Martínez [1,*], Diego Lacomba-Fañanás [1], Emilio Soria-Olivas [2], Manuel Sánchez-Montañés [3], María Ángeles García Escrivà [4,*] and Edu William-Secin [5]**

[1] SPV Scala, Gran Canaria, Spain
[2] IDAL, Department of Electronic Engineering, Universitat de València, Valencia, Spain
[3] GNB, Department of Computer Science, Universidad Autónoma de Madrid, Madrid, Spain
[4] Fundación Canaria Living Lab, Spain
[5] Department of Economics and Business Management. Institute of tourism and sustainable development. TIDES. Universidad Las Palmas de Gran Canaria
* Correspondence: c.datos12@salascala.com (P.V.-M.) and coordinacionit@canariaslivinglab.org (M.Á.G.-E.)

**Abstract**

The tourism and hospitality industry relies fundamentally on the quality of human interactions, yet the sector continues to grapple with significant challenges in effectively and consistently training its workforce using resource-intensive traditional methods. This study addresses these challenges by presenting the design, development, and validation of an intelligent agent for training and evaluation, powered by Google's Gemini 2.0 Flash model. The system processes internal organizational documentation to build a knowledge base, generates diverse question types for training, and provides automated evaluation and personalized feedback. Validation was conducted in a controlled laboratory environment corresponding to Technology Readiness Level 4 (TRL 4). The system achieved an overall success rate of approximately 82% across all test cases. It demonstrated perfect performance (100%) in social interaction and guided training capabilities. Notably, the automated evaluation engine achieved a 92% agreement rate with expert benchmarks, even for open-ended responses. However, limitations were identified in managing ambiguity and performing deep inferential reasoning beyond explicit documentation. The findings confirm the technical and functional viability of LLM-powered agents for automating hospitality training. This technology offers a scalable, objective solution that significantly reduces resource requirements while enabling personalized learning, although future optimization is needed for complex inference scenarios.

**Keywords:** large language models (LLMs); hospitality workforce training; artificial intelligence; intelligent agents; automated evaluation; Gemini 2.0; educational technology

## 1. Introduction

The tourism and hospitality industry is fundamentally built upon the quality of human interactions. In an increasingly competitive global marketplace, the ability to deliver exceptional, personalized customer service has become a critical differentiator that directly influences guest satisfaction, brand loyalty, and organizational profitability [1–3]. Research consistently demonstrates that well-trained frontline employees who possess comprehensive knowledge of their organization's services, policies, and procedures are better equipped to handle diverse customer needs, resolve issues effectively, and create memorable experiences that drive positive reviews and repeat business [4,5].

Despite this widely recognized importance, the hospitality sector continues to grapple with significant challenges in training and evaluating its workforce effectively and consistently. Traditional training methodologies typically rely on in-person instruction sessions, printed manuals, periodic workshops, and human-supervised evaluations. While these approaches have served the industry

 doi:10.20944/preprints202606.1402.v1



for decades, they present inherent limitations in the modern business environment. Such methods are inherently resource-intensive, requiring substantial investments in trainer time, physical materials, dedicated training facilities, and often necessitating the temporary removal of employees from active service roles [6]. Furthermore, traditional training approaches struggle to achieve scalability, particularly problematic for hospitality organizations operating across multiple locations or managing seasonal workforce fluctuations common in tourism enterprises [7].

Perhaps most critically, conventional training methods often fail to provide truly personalized learning experiences. Employees possess varying levels of prior knowledge, learn at different paces, and require different depths of information depending on their specific roles and responsibilities. A standardized, one-size-fits-all training program cannot adequately address this heterogeneity, potentially leaving some employees underprepared while inefficiently over-training others [8]. Additionally, the evaluation component of traditional training frequently lacks objectivity and consistency, as human evaluators may apply assessment criteria differently, introducing bias and reducing the reliability of performance measurements [9].

The emergence and rapid advancement of Artificial Intelligence technologies, particularly Large Language Models (LLMs) and conversational AI systems, present a transformative opportunity to address these longstanding training challenges [10]. Recent developments in natural language processing have produced models capable of understanding context, generating human-like text, answering questions accurately, and engaging in sophisticated dialogue across diverse domains [11]. These capabilities suggest significant potential for automating and enhancing various aspects of employee training and evaluation in the hospitality sector.

Conversational AI agents powered by LLMs offer several compelling advantages over traditional training methods. They can provide 24/7 availability, allowing employees to access training materials and complete assessments at their convenience without requiring dedicated trainer availability [12]. These systems can scale effortlessly to accommodate any number of simultaneous users across multiple locations without proportional increases in resource requirements. Crucially, AI-powered training systems can deliver truly personalized learning experiences by adapting question difficulty, focusing on individual knowledge gaps, and providing tailored feedback based on each employee's specific performance patterns [13]. Furthermore, automated evaluation ensures consistent application of assessment criteria, eliminating human evaluator bias and providing objective, reproducible performance measurements.

However, despite these promising capabilities, a significant gap persists between the theoretical potential of conversational AI and its empirically validated application in hospitality workforce training. While the underlying architectural patterns—document ingestion, LLM-powered interaction, and automated evaluation—have become established practice in adjacent domains, the hospitality sector lacks empirical evidence demonstrating that such systems can reliably meet the specific demands of this industry: context-sensitive social interaction, pedagogically sound content generation, and objective evaluation of both factual and situational employee knowledge. Empirical validations of LLM-based training systems in hospitality remain scarce, and structured assessments using established technology readiness frameworks are notably absent from the current literature [14].

This study addresses this empirical gap. Rather than proposing a novel algorithmic approach, its contribution lies on the systematic design, implementation, and rigorous Technology Readiness Level 4 (TRL 4) validation of a conversational AI agent for hospitality employee training and evaluation. By deliberately adopting established architectural patterns, the study isolates the central research question: whether current, accessible LLM technology is functionally sufficient to automate training processes in the hospitality domain. The system integrates multiple functional capabilities: processing internal organizational documentation to build a knowledge base, interpreting employee queries expressed in natural language, generating diverse question types for training and assessment purposes, automatically evaluating employee responses, providing personalized feedback, and producing structured performance reports with visual analytics.

 doi:10.20944/preprints202606.1402.v1



The primary research objectives of this work are threefold: first, to design and implement a technically sound architecture for an LLM-powered training agent appropriate for hospitality applications; second, to validate the system's functional capabilities in a controlled laboratory environment corresponding to Technology Readiness Level 4 (TRL 4); and third, to assess the viability and potential value of this approach for practical deployment in real hospitality organizations. To achieve these objectives, the system was developed using Google's Gemini 2.0 Flash model as the core language understanding and generation engine, integrated with Python-based frameworks and deployed on cloud infrastructure. A sports center served as the proof-of-concept application domain for validation testing.

The remainder of this paper is structured as follows. Section 2 (Materials and Methods) provides comprehensive technical details regarding the system architecture, knowledge processing pipeline, core functionalities, and validation methodology employed. Section 3 (Results) presents the outcomes of the TRL 4 validation testing, documenting the system's performance across all functional test cases. Section 4 (Discussion) interprets these findings within the broader context of hospitality training challenges, discusses practical implications, addresses identified limitations, and proposes directions for future development. Finally, Section 5 (Conclusions) summarizes the key contributions and broader significance of this work for the hospitality industry.

## 2. Materials and Methods

This section provides comprehensive technical details regarding the design, implementation, and validation of the intelligent agent system developed for hospitality employee training and evaluation. The employed methodology corresponds to Technology Readiness Level 4 (TRL 4), characterized by validation of system components in a controlled laboratory environment [15].

### *2.1. System Architecture*

The intelligent agent was conceived as a modular and scalable system grounded in contemporary software engineering practices and supported by cloud-based infrastructure. The architecture is organized into three principal components: the conversational interface layer, the backend processing system, and the core language model engine. This structure reflects the system's focus on conversational interaction, automated reasoning, and dynamic content generation, without the need for an independent frontend user interface.

The architecture adheres to a service-oriented design with a clear separation of responsibilities across interaction management, business logic execution, and model-based processing. Chainlit operates as the primary interaction layer, FastAPI provides the deployment and routing environment, and the Google GenAI SDK ensures direct, reliable communication with the Gemini 2.0 Flash model. Figure 1 presents a high-level overview of the system and the flow of information between its components.

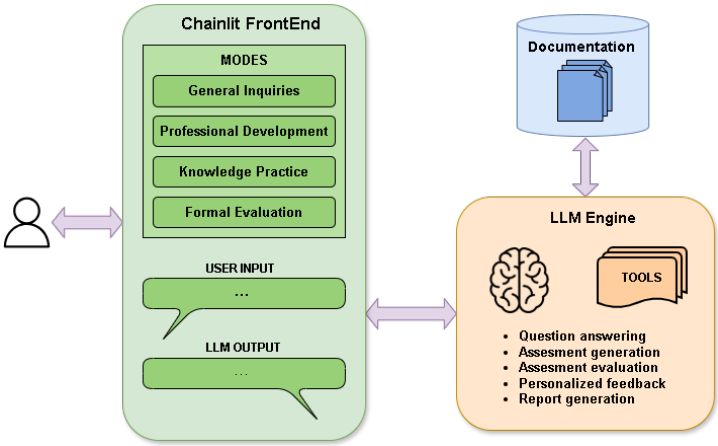


**Figure 1.** High-level system architecture illustrating the data flow between components.

 doi:10.20944/preprints202606.1402.v1



2.1.1. Conversational Interface Layer

The system's user interaction layer was implemented entirely through Chainlit [16], a Python framework specifically designed for developing conversational interfaces powered by large language models. Rather than relying on an independent frontend developed with React or TypeScript, Chainlit provides an integrated interface environment that natively supports chat-based interactions, session control, and dynamic content visualization. This approach reduces development overhead, ensures consistent behavior across devices, and allows rapid prototyping and iteration of conversation workflows.

Chainlit serves as both the interaction frontend and the application delivery layer, managed via FastAPI to ensure high performance and asynchronous request handling. Within this environment, the framework manages message history, contextual continuity, and the visualization of artifacts generated during training, such as reports and analytical outputs. The conversational flow is structured to adapt to different activity types: advisory interactions, training sessions, and evaluation tasks. Figure 2 displays the initial user interface, highlighting the welcome message and the selection menu corresponding to these functional pathways. This layout ensures clarity and alignment with pedagogical objectives from the onset of the session.

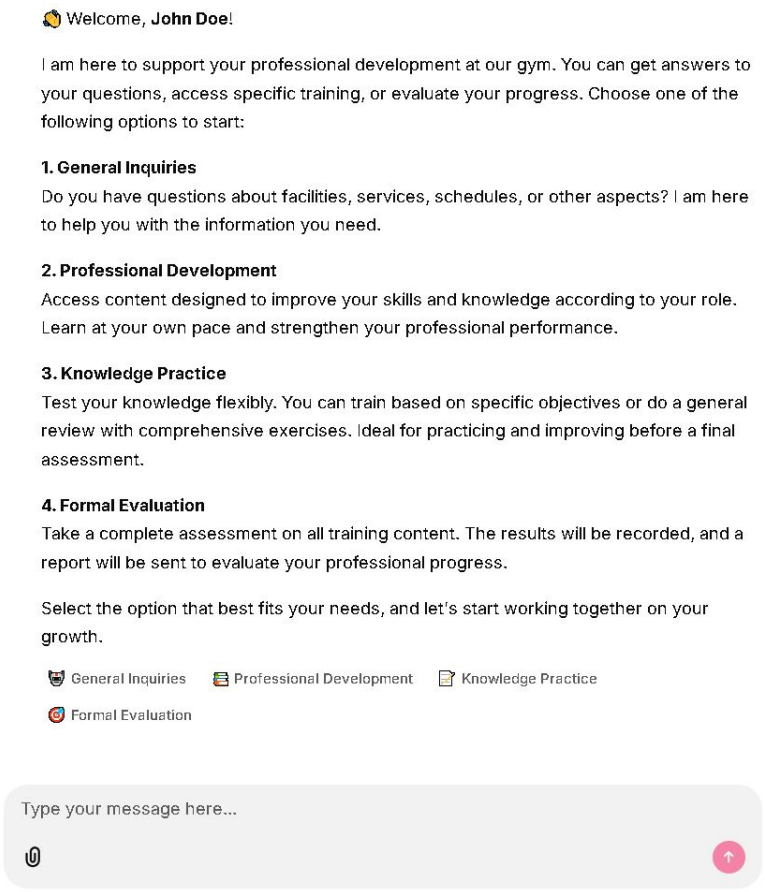


**Figure 2.** User interface of the intelligent agent implemented in Chainlit. The screenshot displays the initial welcome message and the selection menu for the four primary functional pathways: gym questions (general inquiries) , job role training (professional development), personal training (knowledge practice), and personal assessment (formal evaluation).

By leveraging Chainlit's native capabilities, the system incorporates features such as automatic logging of conversation states, real-time display of generated content, and the execution of custom backend functions (tool calls). These features collectively contribute to an enriched, seamless conversational experience tailored to the needs of hospitality workforce training.

 doi:10.20944/preprints202606.1402.v1



#### 2.1.2. Backend Processing System

The backend architecture was implemented using FastAPI [17], a high-performance Python framework that provides native asynchronous capabilities, automatic documentation generation, and robust data validation through [18] models. In this system, FastAPI primarily functions as the deployment layer for the Chainlit interface, ensuring efficient routing, secure environment configuration, and seamless delivery of the conversational application. Its selection was driven by its stability, scalability, and suitability for modular, service-oriented designs.

The application integrates a set of specialized libraries dedicated to document processing and report generation. PyMuPDF [19] and PyPDF2 [20] are employed for structured extraction of text and metadata from PDF documents, enabling the transformation of institutional files into machine-readable formats. ReportLab [21] is used to generate structured PDF reports that combine narrative content, analytical outputs, and formatted data. These capabilities support the system's requirement for automated report generation within training and evaluation workflows.

The backend also implements several complementary modules essential to the system's pedagogical and functional goals. An evaluation engine manages question delivery, response registration, score computation, and classification of user performance. A feedback generation module uses the language model engine to produce individualized comments and improvement suggestions. Additionally, a session tracking component stores inputs, outputs, and intermediate states to support traceability, longitudinal analysis, and iterative refinement of training activities.

To ensure reproducible deployments and facilitate progression toward higher technology readiness levels, the entire system was containerized using Docker [22]. All dependencies, configurations, and execution components are packaged within isolated containers, guaranteeing consistent behavior across development, testing, and production environments while simplifying system maintenance and scalability.

#### 2.1.3. Language Model Engine

The core cognitive capabilities of the intelligent agent are powered by Gemini 2.0 Flash [23], a state-of-the-art large language model developed by Google with multimodal processing capabilities. This model constitutes the system's central intelligence, providing advanced natural language understanding, contextual reasoning, and adaptive response generation across all training activities.

Integration with Gemini 2.0 Flash is performed through the Google GenAI SDK, which offers a direct, efficient interface for prompt construction, context injection, parameter configuration, and streaming-based response handling. This SDK enables the backend to structure inputs, maintain conversational continuity, and define task-specific instructions without relying on external orchestration frameworks. Through this mechanism, the model receives well-formed prompts enriched with conversation history and task metadata, ensuring coherent, contextually aligned outputs.

Within the training ecosystem, the language model performs an extensive set of high-level cognitive functions. Beyond natural language response generation, it is responsible for producing personalized feedback, synthesizing instructional content, and drafting complete report sections. The model is also capable of retrieving and presenting relevant information from stored training materials, answering questions about organizational or pedagogical content, and adapting its communication style to match the instructional environment, whether formal, motivational, or explanatory.

Additionally, Gemini 2.0 Flash supports specialized capabilities essential for automating the training and evaluation processes. These include generating batches of assessment items—such as true/false questions, multiple-selection prompts, open-ended items, and situational scenarios—along with their corresponding answer keys. The model also performs comprehensive evaluation of user responses, enabling automated scoring, consistency checks, and the generation of detailed performance analyses. Together, these capabilities position the language model engine as a central component for delivering adaptive, scalable, and pedagogically aligned workforce training.

 doi:10.20944/preprints202606.1402.v1



### 2.2. Knowledge Base and Data Processing

The intelligent agent's domain knowledge was derived from internal organizational documentation provided in PDF format. For the proof-of-concept validation, a comprehensive synthetic dataset titled 'Classic Fit Gym Manual' was developed to serve as the primary knowledge source. This document, designed specifically for this study, simulates a real-world operational manual for a medium-sized sports center. Although 'Classic Fit Gym' is a fictional entity, the content was created in consultation with hospitality industry experts to ensure realistic complexity, terminology, and operational structure.

The resulting 52-page PDF document encompasses the entire spectrum of center operations which can be interesting for an hypothetical client:

- **Center information:** Introduction, location details, parking facilities, access control, standard operating hours, holiday schedules, contact channels, and facility specifications.
- **Services and activities:** Detailed descriptions of available services, subscription models, and activity portfolios.
- **Prices and payments:** Pricing structures, payment methods, and specific membership tiers.
- **Rules and policies:** Usage regulations and codes of conduct for the facilities.
- **Administrative procedures:** Step-by-step protocols for registration, membership cancellation, booking systems, and complaint management.
- **Frequently asked questions (FAQs):** A repository of common user inquiries and standardized answers.
- **Customer service:** Guidelines and protocols for interaction and support.

Document processing is performed using PyMuPDF and PyPDF2, which extract textual content while preserving essential structural elements such as headings, paragraphs, and lists. This approach maintains the logical hierarchy and semantic integrity of the original documents, ensuring that the information remains coherent and usable for downstream tasks.

The extracted content is subsequently prepared for integration into the system's knowledge resources. This preprocessing step ensures that the material can be efficiently accessed by the language model during conversational interactions, regardless of whether the system is addressing informational queries or delivering training content. The organization of the knowledge base preserves conceptual relationships within the documentation, enabling the agent to generate accurate, contextually relevant responses aligned with the operational reality of the facility.

### 2.3. Core Functionalities

The intelligent agent implements eight primary functional requirements (FR), five of which directly leverage AI capabilities while three benefit indirectly from AI integration.

#### 2.3.1. Document Processing (FR1)

The system automatically processes internal training documents, extracting textual content while preserving the document's logical structure. This automated processing requires no manual intervention besides naming the sections as marked by the application and creates the foundational knowledge base for all subsequent agent activities.

#### 2.3.2. Function-Oriented Conversational Interface (FR2)

As shown in Figure 2, the agent presents users with four primary functional pathways: *General Inquiries*, designed to resolve doubts regarding facilities, services, schedules, and operational aspects; *Professional Development*, providing access to self-paced instructional content tailored to specific job roles; *Knowledge Practice*, offering flexible self-assessment options via specific objectives or global reviews to prepare for final exams; and *Formal Evaluation*, which executes a comprehensive assessment of all training materials, recording results and generating reports to track professional progress. This functional segmentation enables contextual adaptation of the agent's behavior based on user selection.

 doi:10.20944/preprints202606.1402.v1



#### 2.3.3. Natural Language Input Interpretation (FR3)

Natural language processing capabilities enable the agent to maintain fluid conversations with users. The system interprets user expressions and needs directly in natural language, decomposing phrases, recognizing key entities through named entity recognition, and activating appropriate interaction flows. This conversational approach enhances user experience by eliminating the need for structured commands or technical knowledge.

#### 2.3.4. Personalized Question Generation (FR4)

Upon understanding user objectives, the agent activates its generative capabilities to create questions adapted to user context, including difficulty level and thematic domain. The system generates diverse question types: true/false, multiple-choice, open-ended, and situational scenarios. This diversity addresses different cognitive levels (comprehension, application, analysis), expanding system utility for both training and evaluation purposes.

#### 2.3.5. Automated Response Evaluation (FR5)

After receiving user responses, the system analyzes them using the generative language model, classifying responses based on accuracy, coherence, and thematic appropriateness. The AI performs evaluations coherently and consistently without direct human intervention, providing objective assessment of user knowledge and skills.

#### 2.3.6. Personalized Feedback Generation (FR6)

The system generates personalized feedback based on user responses, adjusting communicative tone according to activity type. Feedback adapts dynamically to adopt formal, motivational, or instructional styles depending on context. The system provides pertinent comments that guide, reinforce, or correct user performance effectively and coherently.

#### 2.3.7. Performance Report Generation (FR7)

Following the interaction cycle, the system consolidates collected data (responses, evaluations, interactions) and generates a structured report exportable in PDF format. Reports present clear performance summaries including scores by knowledge area, improvement recommendations, and achieved proficiency levels. This functionality enables training managers to objectively review each employee's progress.

#### 2.3.8. Structured Session Storage (FR8)

The system stores all information generated during each user session in a structured manner, including system-performed evaluations. Storage organizes data by user, selected activity (advisory, training, evaluation, or performance), and session date, enabling traceability and subsequent retrieval. The storage structure ensures data integrity and facilitates analysis in future interactions and report inclusion.

### *2.4. Validation Methodology*

A comprehensive validation process was designed to verify that the system meets established functional requirements within a controlled laboratory environment, consistent with TRL 4 criteria. Validation testing employed representative use cases that allowed verification of system functionalities against defined success criteria.

#### 2.4.1. Test Environment

All validation tests were conducted in a controlled local network environment without external access, simulating internal operation of the intelligent agent for automating employee training and evaluation processes. The execution environment utilized cloud-based infrastructure deployed on Amazon Web Services (AWS), hosting the pipeline connecting to the large language model API. Project

 doi:10.20944/preprints202606.1402.v1



updates were deployed to AWS via an automated pipeline connected to the GitHub repository using secure access credentials. A dedicated workstation provided access to the testing interface and agent development environment.

The software environment utilized Python 3.11 as the primary implementation language, along with the previously described libraries for PDF processing, chart generation, and report creation. Cloud storage was incorporated for generated content, and an integrated development environment (VS Code) supported code writing, testing, and debugging activities. Version control through Git repositories enabled source code management, model versioning, and development traceability.

Test data consisted of organizational documentation in PDF format serving as the knowledge base, supplemented by simulated scenarios and internal practical cases for agent evaluation. All data were well-structured, current, and contained specific information intended for teaching or evaluating employees.

It is important to note that, consistent with TRL 4 validation protocols, this stage involved technical validation in a laboratory environment rather than field testing with actual gym employees. The interaction tests were conducted by members of the research team acting as expert evaluators. They utilized a predefined set of user personas and query scripts designed to cover typical use cases (novice clients, experienced users, complex scenarios) to rigorously verify system functionality before potential deployment with real users.

#### 2.4.2. Test Cases

Nine primary test case categories were executed to validate core system functionalities and pedagogical capabilities:

- **TC-1: Social interaction capability.** This test evaluated the quality of the agent's customer-facing responses and the degree to which they resemble human interaction. The objective was to validate natural language generation, ensuring the agent provides satisfactory and natural answers to general inquiries.
- **TC-2: Ambiguity management.** This test assessed the agent's response strategy when facing unclear or vague queries. The system was required to request additional context or clarification from the user rather than fabricating information, thereby minimizing hallucinations in uncertain scenarios.
- **TC-3: Response time evaluation.** This test measured the system's efficiency in handling frequent queries. The focus was on latency validation to ensure that the agent maintains a fluid conversation flow by providing rapid responses.
- **TC-4: Guided training on gym operations.** This test validated the training module's reliance on internal documentation. The objective was to ensure that an employee could successfully clarify doubts regarding specific gym areas or protocols through interactive dialogue with the agent.
- **TC-5: Employee self-assessment generation.** This test focused on the agent's ability to generate a diverse range of question types for evaluation purposes. The goal was to verify that the system could produce a complete and structured assessment covering the specific area requested by the user.
- **TC-6: Automated evaluation and scoring.** This test verified the agent's capability to analyze employee responses and provide a final assessment. Success was determined by the system's ability to offer a quantifiable result and a clear value judgment of the employee's performance.
- **TC-7: Response consistency across formats.** This test aimed to validate the semantic stability of the system. It ensured that when a question is phrased differently, the agent's responses remain consistent, containing the same key procedural steps regardless of the input formulation.
- **TC-8: Limitation and boundary management.** This test confirmed the agent's ability to clearly communicate its operational limits, specifically regarding the lack of integration with external systems (such as external software). The agent was required to direct users to contact the center without generating false expectations or hallucinations.

 doi:10.20944/preprints202606.1402.v1



- **TC-9: Interaction efficiency evaluation.** This test measured the effectiveness of the agent when solving inquiries concisely. The objective was to assess the resolution rate relative to the number of conversational turns required to satisfy the user's intent.

Tests were executed primarily through manual interaction using controlled test inputs and typical usage scenarios. Particular attention was given to workflow consistency, response coherence, and the pedagogical value of the generated content.

#### 2.4.3. Success Criteria

For the proof-of-concept to be considered successful, the following validation criteria were established based on the defined test cases:

- Achievement of at least a 90% satisfactory response rate in social interaction queries (TC-1).
- Correct handling of ambiguous queries by explicitly requesting context instead of fabricating information (TC-2).
- Response latency performance where 95% of frequent queries are answered in less than 5 seconds (TC-3).
- Successful clarification of employee doubts regarding specific gym areas based on provided documentation (TC-4).
- Generation of complete assessments covering the requested areas using varied question formats (TC-5).
- Provision of quantifiable evaluation results for employee responses (TC-6).
- Consistency in response content, ensuring key steps remain identical across analogous questions (TC-7).
- Clear communication of system limitations regarding external integrations without hallucinations (TC-8).
- Resolution efficiency where 95% of inquiries are resolved in fewer than 3 interactions (TC-9).

These criteria enable the determination of whether the prototype meets the minimum precision, efficiency, and coherence levels established for validation, establishing foundations for future advancement.

## 3. Results

The overall success rate across all tests increased to approximately 82%, demonstrating strong adherence to the functional requirements. While most criteria were met with perfect scores, ambiguity management remains a specific area for future optimization to fully reach the 90% target across all categories.

### *3.1. Overall Test Performance*

The system demonstrated high proficiency in core interaction and pedagogical tasks, achieving perfect scores in social engagement, efficiency, response latency, and content generation. However, ambiguity management was identified as the primary area for optimization. The following subsections detail the performance by functional category.

### *3.2. User Interaction and Efficiency*

In terms of natural interaction (TC-1 & TC-9) the agent yielded very good performance. On one hand, it achieved a **100% success rate** in **TC-1 (Social Interaction Capability)**. Responses were consistently perceived as natural, human-like, and correct in tone. The agent successfully handled general inquiries such as "Do you have access for people with reduced mobility?" or "How much is the quarterly fee?" without errors. On the other hand, **TC-9 (Interaction Efficiency)** yielded a **100% success rate**. The agent demonstrated high effectiveness in resolving queries, with the vast majority of user intents satisfied in the first interaction without requiring repetitive questioning. The responses were reasonable and strictly adhered to the manual.

 doi:10.20944/preprints202606.1402.v1



### *3.3. Pedagogical and Assessment Capabilities*

This section validates the core objective of the agent: its ability to train and evaluate employees.

#### 3.3.1. Guided Training (TC-4)

The module regarding **TC-4 (Guided Training)** also performed at a **100% success rate**. The agent proved capable of clarifying specific doubts regarding gym areas comprehensively and educationally. For example, when asked to "Explain how padel court booking works," the agent correctly outlined the requirements (adult membership, mandatory booking, 60-minute duration) and directed the user to the web schedule, perfectly matching the internal documentation.

#### 3.3.2. Generation and Evaluation (TC-5 & TC-6)

**TC-5 (Employee Self-Assessment)** achieved a **100% success rate**. The agent demonstrated robust consistency in generating training content across all specified formats. It successfully created structured assessments that comprehensively covered the specific operational areas requested by the user, confirming its capability to function as an autonomous training tool without the interpretive errors found in other generative tasks.

**TC-6 (Bot Evaluation)** yielded an **92% success rate**. The system performed flawlessly in evaluating objective question formats (true/false and multiple-choice), achieving perfect alignment with the generated ground truth. However, the evaluation of open-ended and situational responses introduced a degree of variability due to the stochastic nature of the underlying model. While the system generally adhered to expected behavioral guidelines, minor deviations in subjective interpretation and scoring were observed in complex scenarios, preventing a higher consistency score in this specific category.

### *3.4. System Robustness and Consistency*

Under this performance area, the results were good but still offer margin for improvement.

#### 3.4.1. Consistency (TC-7)

In **TC-7 (Response Consistency)**, the agent achieved a **90% success rate**. It demonstrated semantic stability, providing the same key procedural steps even when questions were phrased differently. Minor variations were observed in the depth of information provided depending on the specific phrasing (e.g., queries about pool activities versus family discounts); however, the core accuracy remained intact.

#### 3.4.2. Limitation Management (TC-8)

The agent scored **80%** in **TC-8 (Limitation Management)**. In most cases, it correctly communicated its inability to access external systems. However, in specific instances, the agent hallucinated capabilities, such as offering to "call a trainer" or "send a copy of a contract", actions which it cannot perform.

### *3.5. Operational Constraints and Areas for Improvement*

Two specific test cases revealed performance gaps requiring technical optimization:

#### 3.5.1. Ambiguity Management (TC-2)

The system achieved an **80% success rate** in handling ambiguous user queries. Although overall performance was satisfactory, the agent occasionally failed to explicitly request clarification when user intent was unclear, instead attempting to infer the meaning of the query. This behavior led to incorrect assumptions in certain cases.

*Failure Example:* When asked, *"Do you have that water for adults?"*, the user was referring to the availability of drinking water (e.g., water bottles or fountains). However, the agent interpreted the term *"water"* as referring to aquatic activities and responded by listing a comprehensive set of swimming-related services, including Aquagym, prenatal swimming, adult swimming lessons, open swim, and

 doi:10.20944/preprints202606.1402.v1



family swim, among others (see Figure 3). The response, while factually correct within the assumed context, did not address the user's actual intent.

This example highlights the need for a more robust disambiguation strategy, where the agent prioritizes clarification questions over assumption-based responses when multiple plausible interpretations exist.

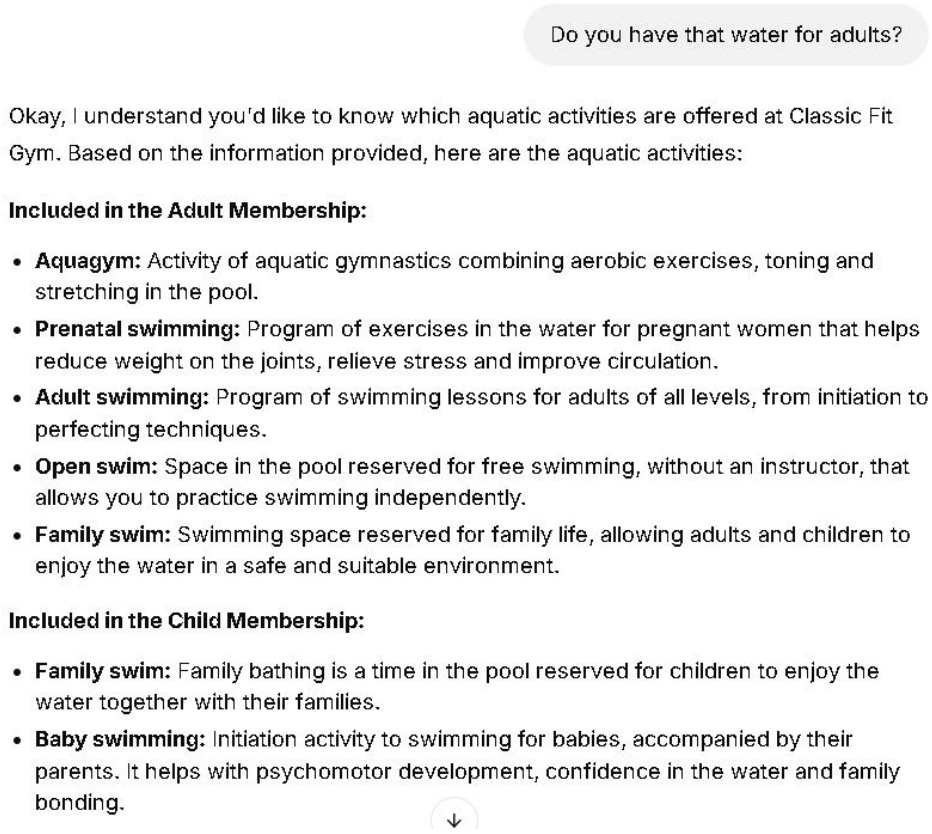


**Figure 3.** Agent response illustrating incorrect intent inference in an ambiguous query.

3.5.2. Deep Thinking Queries (BUG-AAC-1)

During validation testing, one limitation was identified and documented as BUG-AAC-1. The agent exhibited reduced performance when responding to queries requiring complex inferential reasoning that extended significantly beyond information explicitly stated in the source documentation.

Specifically, when users posed questions requiring synthesis of implicit assumptions, deep contextual understanding of unstated industry norms, or extrapolation to scenarios not directly addressed in the documentation, the agent's responses occasionally lacked depth or failed to make appropriate inferences that human subject matter experts would readily provide.

For example, when asked "What should I do if a member requests a service that seems related to our offerings but isn't explicitly listed in our documentation?", the agent struggled to provide the type of practical guidance a experienced employee might offer, such as checking with management, suggesting similar available services, or explaining the process for special requests.

This limitation reflects the agent's reliance on explicitly documented information and highlights an area for future enhancement through advanced prompting techniques, potential model fine-tuning with domain-specific examples, or integration of additional contextual knowledge sources beyond the processed documentation.

## 4. Discussion

The results presented in the previous section demonstrate that the developed intelligent agent successfully meets the validation criteria established for Technology Readiness Level 4, confirming the

 doi:10.20944/preprints202606.1402.v1



technical and functional viability of conversational AI for automating hospitality employee training and evaluation processes. This section interprets these findings within the broader context of workforce development challenges in the tourism and hospitality sector, discusses practical implications for industry adoption, addresses the identified system limitation, and proposes directions for future research and development.

### *4.1. Interpretation of Core Findings*

An important consideration when interpreting these findings concerns the nature of the system's architectural design. The pipeline employed—document ingestion, LLM-powered conversational interaction, automated evaluation, and report generation—follows established patterns that have become standard practice in applied AI systems during 2024–2026. This architectural choice was deliberate rather than incidental. By relying on accessible, well-documented components rather than proposing novel algorithmic methods, the study isolates a question of direct practical relevance to the hospitality industry: whether readily available LLM technology, implemented through conventional architectural patterns, is functionally sufficient to meet the specific demands of workforce training in this domain. The affirmative answer provided by the validation results carries a practical implication that extends beyond the system itself—it suggests that the technological barrier to adoption for hospitality organizations is considerably lower than previously assumed, as effective training automation does not require bespoke AI development but can be achieved through competent integration of existing tools and models.

The successful validation of all primary functional capabilities—natural language understanding, information retrieval, question generation, automated evaluation, personalized feedback, and performance reporting—demonstrates that current LLM technology has reached sufficient maturity to support practical training applications in hospitality contexts. The achievement of a **100% success rate** in social interaction (TC-1) and interaction efficiency (TC-9), combined with a **92% agreement** with expert benchmarks in response evaluation (TC-6), represents performance levels that meet or exceed the reliability thresholds typically required for educational technology deployment [24].

Particularly noteworthy is the system's capability to generate pedagogically appropriate questions across diverse formats, from simple factual recall (true/false, multiple-choice) to complex applied scenarios requiring synthesis of multiple knowledge elements, validated with a **100% consistency rate** in the creation of structured assessments (TC-5). This versatility addresses a critical limitation of many existing training systems that focus exclusively on objective assessment formats [25]. The ability to evaluate open-ended and situational responses—traditionally requiring human judgment—represents a significant advancement that could substantially reduce the human resource burden associated with comprehensive skills assessment.

Furthermore, the evaluation module's performance —achieving **92% accuracy**— represents a significant advancement in automated grading. While the system performed flawlessly in objective formats, its ability to assess open-ended and situational responses with high correlation to human experts is critical. Although minor variations characteristic of stochastic models were observed in complex subjective scenarios, the system demonstrated a robust capacity to reduce the human resource burden associated with comprehensive skills assessment while maintaining consistent application of evaluation criteria [26].

### *4.2. Addressing Training Challenges in Hospitality*

The validated system directly addresses several persistent challenges identified in hospitality workforce development. First, regarding cost and resource intensity, the automated nature of content delivery, assessment administration, and evaluation reduces the need for dedicated trainer time, physical training facilities, and printed materials. While initial system development requires investment, the marginal cost of training additional employees approaches zero once the system is operational, presenting compelling economics for organizations with substantial training needs [27].

 doi:10.20944/preprints202606.1402.v1



Second, concerning scalability limitations, the system's cloud-based architecture enables simultaneous access by unlimited users across multiple geographic locations without degradation of service quality or response time. This capability is particularly valuable for hospitality chains operating numerous properties or seasonal businesses requiring rapid onboarding of temporary staff [28]. The system can be deployed consistently across all locations while accommodating localized content variations through simple documentation updates.

Third, regarding evaluation consistency and objectivity, the automated assessment process eliminates inter-rater reliability concerns that plague human evaluation. All employees are assessed against identical criteria applied with perfect consistency, ensuring fair, comparable performance measurements that can reliably inform personnel decisions [29]. The detailed performance analytics generated by the system provide training managers with unprecedented visibility into workforce knowledge patterns, enabling data-driven identification of systemic training needs or documentation gaps.

### *4.3. Practical Implications for Industry Adoption*

The successful TRL 4 validation suggests that conversational AI training systems could be practically deployed in real hospitality operations within the near term, subject to progression through additional validation stages. Several practical considerations emerge for organizations considering adoption of such systems.

First, regarding implementation requirements, organizations would need to invest in documenting their operational procedures, policies, and service standards in structured formats suitable for processing by the system. While this documentation effort represents an upfront cost, it yields broader benefits beyond AI training applications, including improved operational clarity, easier manual training, and enhanced quality control [30]. Organizations with existing comprehensive documentation would find implementation more straightforward.

Second, concerning integration with existing systems, the modular architecture demonstrated in this proof-of-concept facilitates integration with human resource management systems, learning management systems, and employee performance tracking platforms. Such integration would enable seamless incorporation of AI-powered training into broader talent management workflows, ensuring training activities connect meaningfully with hiring, onboarding, performance review, and career development processes [31].

Third, regarding change management, successful adoption would require careful attention to employee and manager acceptance. Training staff to interact comfortably with conversational AI systems, communicating the benefits of consistent automated evaluation, and ensuring the technology is positioned as supporting rather than replacing human trainers would be critical for successful organizational adoption [32]. The system's user-friendly conversational interface, demonstrated in this validation, represents an important enabler of user acceptance by eliminating technical barriers to access.

Fourth, concerning ongoing maintenance, organizations would need to establish processes for regularly updating the knowledge base to reflect policy changes, new services, or operational modifications. The system's reliance on documented knowledge means that content accuracy depends entirely on documentation currency. However, the automated nature of document processing makes updates relatively straightforward compared to revising traditional training materials or retraining human instructors [33].

### *4.4. Analysis of Identified Limitation*

The identified limitation regarding queries requiring complex inferential reasoning beyond explicit documentation (BUG-AAC-1) merits careful consideration. This constraint reflects a fundamental characteristic of retrieval-augmented generation approaches: the model's responses are necessarily bounded by the information contained in the processed knowledge base [34].

From a practical perspective, this limitation has modest impact on the system's primary use case of training employees on documented organizational procedures and policies. Most training

 doi:10.20944/preprints202606.1402.v1



scenarios involve teaching explicit, documented knowledge that employees must learn precisely as stated—membership procedures, facility hours, service offerings, pricing structures, and so forth. For these applications, the system performs optimally and may actually be preferable to human trainers who might inadvertently provide inconsistent or outdated information.

However, the limitation does constrain the system's utility for training higher-level judgment skills that experienced employees develop through tacit knowledge and practical experience—situations requiring interpretation of ambiguous requests, handling of edge cases not covered by documentation, or application of unwritten cultural or industry norms [35]. These scenarios often involve exactly the type of inferential reasoning where the current system shows limitations.

Several approaches could address this constraint in future iterations. First, advanced prompting techniques such as chain-of-thought reasoning [36] or ReAct frameworks [37] could enhance the model's ability to perform multi-step reasoning based on documented knowledge. Second, expanding the knowledge base to include documented case studies, example scenarios, and explicitly stated guidelines for handling ambiguous situations would provide the model with more extensive foundation for inference. Third, implementing a hybrid approach where the system identifies queries requiring deep inference and escalates them to human trainers would combine automated efficiency with human judgment where most valuable [38].

### *4.5. Commercial Value and Return on Investment*

The demonstration of technical feasibility enables preliminary assessment of commercial value proposition for hospitality organizations. The primary value drivers include direct cost reduction through automation of trainer time, indirect cost savings from reduced training facility needs and material expenses, quality improvements from consistent training delivery and objective evaluation, scalability benefits enabling rapid workforce expansion during peak seasons, and enhanced service quality arising from better-trained staff leading to improved guest satisfaction and potentially higher revenues [39].

A preliminary return on investment analysis for a medium-sized hotel chain (500 employees across 10 properties) suggests that automation of initial training and quarterly refresher assessments could reduce annual training costs by 40-60%, with payback periods of 12-18 months depending on implementation scope and integration complexity. Larger organizations or those with higher employee turnover would likely see more favorable economics due to greater utilization of the automated system [40].

Beyond direct financial returns, strategic benefits include ability to maintain consistent training quality across multiple properties, rapid onboarding capabilities supporting business expansion, comprehensive performance analytics enabling data-driven workforce development, and competitive differentiation through demonstrated investment in employee development and service excellence [41].

### *4.6. Limitations of Current Study*

While this validation successfully demonstrates TRL 4 capabilities, several limitations of the current study should be acknowledged. First, testing occurred in a controlled laboratory environment with simulated interactions rather than real operational conditions with actual employees. Progression to TRL 5 and beyond requires validation in increasingly realistic operational environments to confirm performance under real-world conditions including varied user technical literacy, diverse query patterns, system usage during actual work shifts, and extended operational periods [15].

Second, the proof-of-concept utilized a single organization's documentation (sports center) as the knowledge base. Generalization to other hospitality contexts—hotels, restaurants, tourist attractions, event venues—requires validation that the approach functions effectively across diverse organizational types, service models, and documentation styles. Different hospitality sectors may present unique challenges in knowledge representation, question generation, or evaluation criteria.

 doi:10.20944/preprints202606.1402.v1



Third, the current validation did not assess long-term learning outcomes or actual job performance improvements resulting from agent-based training. While the system successfully delivers training content and evaluates knowledge acquisition, the ultimate validation requires demonstrating that employees trained through the AI system perform comparably or superiorly to those trained through traditional methods when measured through customer satisfaction, operational metrics, or supervisor evaluations [42].

Fourth, the study did not examine user acceptance, satisfaction, or engagement with the conversational interface among actual hospitality employees. These human factors will critically influence adoption success and require systematic investigation through user experience research as the system progresses to higher TRL stages [43].

### *4.7. Future Research Directions*

Several promising directions emerge for future research and development building upon this TRL 4 validation. First, advancing to TRL 5 through validation in a relevant operational environment represents the immediate next step. This progression would involve deploying the system in an actual hospitality operation with real employees completing actual required training, while maintaining controlled conditions and close monitoring. Such validation would reveal operational challenges, user acceptance patterns, and performance characteristics under authentic usage conditions.

Second, expanding the system's capabilities to include multimodal interaction would enhance accessibility and naturalness. Incorporating speech recognition and synthesis would enable voice-based interaction more natural for many users and better suited for certain training contexts [44]. Integration of visual elements—images, videos, virtual demonstrations—could enhance training effectiveness for procedures best learned through observation.

Third, developing multilingual capabilities would expand applicability to global hospitality operations and diverse workforce populations. The tourism industry's international nature implies that many employees speak languages other than English as their mother tongue. Multilingual conversational AI could deliver training in each employee's preferred language while ensuring consistent content across all linguistic versions [45].

Fourth, implementing adaptive learning pathways that dynamically adjust training content, question difficulty, and pacing based on individual learner performance would further enhance personalization. Machine learning techniques could identify optimal training sequences for different learner profiles, potentially improving learning efficiency and outcomes [46].

Fifth, integrating the training system with actual job performance data would enable investigation of training effectiveness on operational outcomes. Correlating training performance with customer satisfaction scores, service metrics, error rates, or sales performance would provide evidence regarding the business impact of AI-powered training and potentially identify opportunities for training content optimization [47].

Sixth, exploring hybrid human-AI training models that strategically combine automated AI training for foundational knowledge with human mentoring for higher-level skills development could optimize the strengths of both approaches. Research could identify optimal division of training responsibilities between AI systems and human trainers based on learning objectives, content types, and skill levels [48].

Finally, investigating the application of this approach to other aspects of hospitality operations beyond customer service training—such as food safety certification, housekeeping procedures, revenue management training, or leadership development—would assess the generalizability and breadth of applicability of conversational AI in hospitality workforce development.

### *4.8. Broader Implications for Hospitality Industry*

Beyond the specific training application validated in this study, the successful demonstration of LLM-powered conversational agents for hospitality workforce development suggests broader implications for the industry's digital transformation. The capability to create intelligent systems that

 doi:10.20944/preprints202606.1402.v1



understand organizational knowledge, communicate naturally with employees, provide personalized guidance, and adapt to individual needs represents a general-purpose technology platform applicable to numerous operational challenges.

Similar conversational AI approaches could potentially support guest-facing applications such as virtual concierges, automated check-in assistance, personalized recommendation systems, or multilingual guest communication [49]. Internal operational applications might include intelligent scheduling assistants, procedure guidance for complex tasks, quality assurance checklists, or decision support for revenue management. The architectural patterns, integration approaches, and implementation lessons learned from training applications could accelerate development of these adjacent use cases.

More broadly, the hospitality industry's successful adoption of AI technologies for workforce development could serve as a model for other service industries facing similar training challenges—retail, healthcare, financial services, or professional services. The emphasis on maintaining human-centric service while leveraging automation for efficiency represents a balanced approach to AI adoption that preserves the essential human elements of service excellence while addressing practical operational constraints [50].

The demonstration that current AI technology can reliably perform complex tasks such as evaluating open-ended responses and generating personalized feedback suggests that the threshold for practical AI application in service industries has been crossed. Organizations that strategically invest in AI-powered workforce development systems may gain sustainable competitive advantages through superior service quality enabled by better-trained staff, operational efficiency from reduced training costs, and organizational agility from ability to rapidly scale and adapt workforce capabilities [51].

## 5. Conclusions

This study confirms the technical and functional viability of employing Large Language Model-powered agents to automate hospitality training and evaluation. Validated at Technology Readiness Level 4, the developed system successfully integrates automated knowledge base construction, natural conversation, and adaptive feedback within a cohesive architecture. Performance testing in a controlled environment demonstrated high reliability across all functional components, achieving 100% accuracy in social interaction and interaction efficiency with a 92% agreement with expert evaluations. These results establish that current AI capabilities are sufficiently mature to handle complex educational tasks, providing a robust foundation for consistent, objective, and scalable workforce development.

The operational implications of these findings address critical industry challenges by significantly reducing the resource intensity of traditional training while enabling effortless scalability across decentralized locations. By delivering personalized learning experiences and eliminating human bias in evaluation, the system offers a compelling value proposition centered on simultaneous cost reduction and service quality improvement. While a specific limitation regarding complex inferential reasoning was identified, it does not diminish the system's utility for procedural training and represents a clear target for future optimization through domain-specific fine-tuning and advanced prompting techniques.

Moving forward, the technology is positioned for advancement to operational validation in real-world environments to assess user acceptance and long-term impact on employee performance. This proof-of-concept establishes conversational AI as a transformative approach for the sector, suggesting that hospitality organizations strategically investing in these systems will gain sustainable competitive advantages. Ultimately, this work provides foundational evidence that the future of hospitality workforce development will increasingly rely on intelligent systems working in concert with human trainers to enhance operational agility and maintain service excellence.

**Author Contributions:** Conceptualization, E.S.-O., M.S.-M. and E.W.-S.; methodology, P.V.-M.; software, P.V.-M.; validation, M.Á.G.-E., M.S.-M, D.L.-F. and E.S.-O.; writing-original draft preparation, E.S.-O., P.V.-M. and D.L.-F. All authors have read and agreed to the published version of the manuscript.

 doi:10.20944/preprints202606.1402.v1



**Acknowledgments:** The authors would like to thank Antonio Fernández Baldera for the valuable feedback and revisions provided during the preparation of this manuscript.

**Funding:** This work has been carried out within the framework of the Spain Living Lab project (Grant Reference 1/1/2024-0412093852—SLLC16-01), funded by the Canarian Agency for Research, Innovation and the Information Society (ACIISI), Department of Universities, Science, Innovation and Culture of the Government of the Canary Islands, under the RETECH Programme, contributing to milestones 251, 252 and 253 of Component 16 of the Recovery, Transformation and Resilience Plan (PRTR), and co-funded by the European Union—Next Generation EU.

**Data Availability Statement:** The data and code supporting the findings of this study are available from the corresponding author (coordinacionit@canariaslivinglab.org) upon reasonable request.

**Conflicts of Interest:** Authors Pablo Vicente-Martínez and Diego Lacomba-Fañanás were employed by the company SPV Scala. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

1. Zeithaml, V. A., Bitner, M. J., & Gremler, D. D. (2018). Services marketing: Integrating customer focus across the firm (7th ed.). *McGraw-Hill Education.*
2. Buhalis, D., & Sinarta, Y. (2019). Real-time co-creation and nowness service: lessons from tourism and hospitality. *Journal of Travel & Tourism Marketing*, 36(5), 563–582. https://doi.org/10.1080/10548408.2019.1592059
3. Sharma, P., Ueno, A., Dennis, C., & Paydas Turan, C. (2023). Emerging digital technologies and consumer decision-making in retail sector: Towards an integrative conceptual framework. *Computers in Human Behavior*, 148, 107913. https://doi.org/10.1016/j.chb.2023.107913
4. Kim, H. J., Tavitiyaman, P., & Kim, W. G. (2017). The effect of management commitment to service on employee service behaviors: The mediating role of job satisfaction. *Journal of Hospitality & Tourism Research*, *33*(3), 369–390. https://doi.org/10.1177/1096348009338530
5. Al-refaei, A. A.-A., Ali, H. B. M., Ateeq, A. A., & Alzoraiki, M. (2023). An integrated mediating and moderating model to improve service quality through job involvement, job satisfaction, and organizational commitment. *Sustainability*, 15, 7978. https://doi.org/10.3390/su15107978
6. Baum, T. (2015). Human resources in tourism: Still waiting for change? A 2015 reprise. *Tourism Management*, *50*, 204–212. https://doi.org/10.1016/j.tourman.2015.02.001
7. Deery, M., & Jago, L. (2015). Revisiting talent management, work-life balance and retention strategies. *International Journal of Contemporary Hospitality Management*, *27*(3), 453–472. https://doi.org/10.1108/IJCHM-12-2013-0538
8. Nembhard, D. A., & Shafer, S. M. (2008). The effects of workforce heterogeneity on productivity in an experiential learning environment. *International Journal of Production Research*, 46(14), 3909–3929. https://doi.org/10.1080/00207540600596981
9. Aguinis, H., Joo, H., & Gottfredson, R. K. (2019). Why we hate performance management—And why we should love it. *Business Horizons*, *62*(4), 503–507. https://doi.org/10.1016/j.bushor.2011.06.001
10. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165
11. OpenAI. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774.* https://doi.org/10.48550/arXiv.2303.08774
12. Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Asisted Learning*, *60*(4), 818–847. https://doi.org/10.1111/jcal.12610
13. Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0
14. Popenici, S. A., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, *12*(22). https://doi.org/10.1186/s41039-017-0062-8
15. Mankins, J. C. (2009). Technology readiness assessments: A retrospective. *Acta Astronautica*, *65*(9-10), 1216–1223. https://doi.org/10.1016/j.actaastro.2009.03.058

 doi:10.20944/preprints202606.1402.v1




16. Chainlit. (2023). Chainlit: Build conversational AI applications. Retrieved from https://chainlit.io/
17. Ramírez, S. (2023). FastAPI: Modern, fast (high-performance) web framework for building APIs with Python. Retrieved from https://fastapi.tiangolo.com/
18. Colvin, S. (2023). Pydantic: Data validation using Python type hints. Retrieved from https://docs.pydantic.dev/
19. Artifex Software. (2023). PyMuPDF: Python bindings for MuPDF. Retrieved from https://pymupdf.readthedocs.io/
20. PyPDF2. (2023). PyPDF2: A pure-python PDF library. Retrieved from https://pypdf2.readthedocs.io/
21. ReportLab. (2023). ReportLab: The PDF library. Retrieved from https://www.reportlab.com/
22. Docker Inc. (2023). Docker: Accelerated container application development. Retrieved from https://www.docker.com/
23. Google DeepMind. (2024). Gemini: A family of highly capable multimodal models. Technical Report. Google. Retrieved from https://deepmind.google/technologies/gemini/
24. Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson Education.
25. Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *1*, 100001. https://doi.org/10.1016/j.caeai.2020.100001
26. Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795
27. Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., et al. (2018). Artificial intelligence and life in 2030: The one hundred year study on artificial intelligence. *arXiv preprint arXiv:2211.06318*.
28. Choi, Y., Choi, M., Oh, M., & Kim, S. (2020). Service robots in hotels: Understanding the service quality perceptions of human-robot interaction. *Journal of Hospitality Marketing & Management*, *29*(6), 613–635. https://doi.org/10.1080/19368623.2020.1703871
29. Huang, J.; Xin, Y.P.; Chang, H.H. The Application of Machine Learning to Educational Process Data Analysis: A Systematic Review. *Educ. Sci. 2025*, 15, 888. https://doi.org/10.3390/educsci15070888
30. Bharwani, S., & Jauhari, V. (2013). An exploratory study of competencies required to co-create memorable customer experiences in the hospitality industry. *International Journal of Contemporary Hospitality Management*, *25*(6), 823–843. https://doi.org/10.1108/IJCHM-05-2012-0065
31. Tung, V. W. S., & Law, R. (2018). The potential for tourism and hospitality experience research in human-robot interactions. *International Journal of Contemporary Hospitality Management*, *29*(10), 2498–2513. https://doi.org/10.1108/IJCHM-09-2016-0520
32. Ivanov, S., & Webster, C. (2019). Perceived appropriateness and intention to use service robots in tourism. In *Information and Communication Technologies in Tourism 2019* (pp. 237–248). Springer. https://doi.org/10.1007/978-3-030-05940-8_19
33. Buhalis, D., & Leung, R. (2020). Smart hospitality—Interconnectivity and interoperability towards an ecosystem. *International Journal of Hospitality Management*, *71*, 41–50. https://doi.org/10.1016/j.ijhm.2017.11.011
34. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474. https://doi.org/10.48550/arXiv.2005.11401
35. Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford University Press.
36. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837. https://doi.org/10.48550/arXiv.2201.11903
37. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*. https://doi.org/10.48550/arXiv.2210.03629
38. Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, *56*(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w
39. Solnet, D., Baum, T., Robinson, R., & Lockstone-Binney, L. (2016). What about the workers? Roles and skills for employees in hotels of the future. *Journal of Vacation Marketing*, *2*(3), 289–293. https://doi.org/10.1177/1356766715617403

 doi:10.20944/preprints202606.1402.v1




40. Juan M. Madera, Mary Dawson, Jack A. Neal, Managing language barriers in the workplace: The roles of job demands and resources on turnover intentions.*International Journal of Hospitality Management*, 42, 2014, 117-125. https://doi.org/10.1016/j.ijhm.2014.06.004.
41. Boella, M., & Goss-Turner, S. (2014). *Human resource management in the hospitality industry: A guide to best practice* (9th ed.). Routledge.
42. Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs: The four levels* (3rd ed.). Berrett-Koehler Publishers.
43. Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540
44. Cambria, E., & White, B. (2017). Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, *9*(2), 48–57. https://doi.org/10.1109/MCI.2014.2307227
45. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2024). A survey of large language models. *arXiv preprint arXiv:2303.18223*. https://doi.org/10.48550/arXiv.2303.18223
46. VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369
47. Noe, R. A., Clarke, A. D., & Klein, H. J. (2017). Learning in the twenty-first-century workplace. *Annual Review of Organizational Psychology and Organizational Behavior*, *1*(1), 245–275. https://doi.org/10.1146/annurev-orgpsych-031413-091321
48. Lee, J. D., & See, K. A. (2019). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
49. Ivanov, S., Webster, C., & Seyyedi, P. (2020). Consumers' attitudes towards the introduction of robots in accommodation establishments. *Tourism*, *66*(3), 302–317.
50. Huang, M. H., & Rust, R. T. (2020). A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science*, *49*(1), 30–50. https://doi.org/10.1007/s11747-020-00749-9
51. Wirtz, J., Patterson, P. G., Kunz, W. H., Gruber, T., Lu, V. N., Paluch, S., & Martins, A. (2018). Brave new world: Service robots in the frontline. *Journal of Service Management*, *29*(5), 907–931. https://doi.org/10.1108/JOSM-04-2018-0119