Design and Evaluation of an AI-Based Conversational Agent for Travel Agencies: Enhancing Training, Assistance, and Operational Efficiency

Pablo Vicente-Martínez * , Emilio Soria-Olivas , Inés Esteve-Mompó , Manuel Sánchez-Montañés , María Ángeles García Escrivà , Edu William-Secin

Posted Date: 19 December 2025

doi: 10.20944/preprints202512.1693.v1

Keywords: artificial intelligence; retrieval-augmented generation (RAG); travel agencies; operational efficiency; knowledge management; large language models (LLMs); conversational agent

 doi:10.20944/preprints202512.1693.v1

Article

Design and Evaluation of an AI-Based Conversational Agent for Travel Agencies: Enhancing Training, Assistance, and Operational Efficiency

Pablo Vicente-Martínez [1], Emilio Soria-Olivas [1], Inés Esteve-Mompó [2], Manuel Sánchez-Montañés [3], María Ángeles García Escrivà [4] and Edu William-Secin [5]

[1] IDAL, Department of Electronic Engineering, Universitat de València, Valencia, Spain
[2] SPV Sala Scala, Gran Canaria, Spain
[3] 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: pablo.vicmar8@gmail.com

Abstract

The tourism industry faces increasing pressure for agile, personalized services, yet travel agencies struggle with fragmented knowledge scattered across isolated systems and legacy formats. While Large Language Models (LLMs) are widely applied in customer-facing roles, their potential to enhance internal operational efficiency remains largely underexplored. This study presents the design and evaluation of an intelligent assistant specifically for travel agency operations, built upon a Retrieval-Augmented Generation (RAG) architecture using Gemini 2.0 Flash. The system integrates heterogeneous data sources, including structured product catalogs and unstructured documentation processed via Optical Character Recognition (OCR), into a unified interface comprising work assistance, interactive training, and evaluation modules. Results demonstrate information retrieval times not greater than 45 seconds, ensuring its daily usability, while maintaining 95% accuracy. Furthermore, the system democratizes tacit senior expertise and accelerates new employee onboarding. This research validates RAG architectures as a powerful solution to knowledge fragmentation, shifting the strategic AI focus from customer automation to employee empowerment and operational optimization.

Keywords: artificial intelligence; retrieval-augmented generation (RAG); travel agencies; operational efficiency; knowledge management; large language models (LLMs); conversational agent

1. Introduction

The tourism and hospitality industry is experiencing unprecedented digital transformation, driven by the imperative to deliver agile, personalized services in an increasingly competitive marketplace. As customer expectations evolve toward instantaneous, tailored experiences, travel agencies face mounting pressure to optimize their operational efficiency while maintaining service quality [1]. Within this context, artificial intelligence (AI) and large language models (LLMs) have emerged as disruptive technologies with the potential to fundamentally reshape how tourism businesses operate and interact with both customers and internal knowledge systems [2].

Recent advances in generative AI, particularly the development of increasingly sophisticated large language models, have opened new frontiers in natural language processing and information retrieval [3]. These models demonstrate remarkable capabilities in understanding context, generating human-like responses, and processing complex queries across multiple domains. However, their application in the tourism sector has been predominantly customer-facing, with limited exploration of how these technologies can enhance internal operational efficiency for travel agency staff [4,5].

 doi:10.20944/preprints202512.1693.v1

1.1. The Challenge of Fragmented Knowledge Management

Traditional travel agencies confront a critical operational challenge: the fragmentation and dispersion of essential business information across multiple formats, platforms, and systems. Product catalogs, client databases, supplier contacts, booking procedures, policy documents, and training materials typically exist in isolated silos—scattered across email repositories, shared drives, legacy systems, and even tacit knowledge held by experienced agents [6]. This information fragmentation creates significant inefficiencies, as agents must invest substantial time searching for specific data points, consulting colleagues, or navigating disconnected systems during client interactions.

The consequences of this fragmented knowledge architecture extend beyond mere inconvenience. Research indicates that information retrieval inefficiencies directly impact service quality, customer satisfaction, and agent productivity [6]. When agents cannot rapidly access accurate information, response times increase, errors multiply, and the personalized service that differentiates human agents from automated systems becomes compromised. Moreover, the onboarding of new staff becomes protracted and resource-intensive, as institutional knowledge remains largely undocumented and dispersed among veteran employees.

1.2. State of the Art: LLMs and RAG in Tourism

The application of LLMs in tourism and hospitality has gained considerable research attention since 2023. Multiple studies have demonstrated the efficacy of LLM-powered chatbots for customer service, with implementations showing improved response accuracy and customer engagement [8]. Chatbot systems leveraging models such as GPT and Claude have been deployed for itinerary planning, destination recommendations, and frequently asked question resolution [9].

However, a critical limitation of vanilla LLMs is their tendency to generate plausible but factually incorrect information. This fact, known as hallucinations, poses significant risks for business applications [10]. In the tourism context, where accurate information on pricing, availability, policies, and regulations is paramount, such unreliability poses significant operational risks. Furthermore, LLMs trained on general internet data lack access to proprietary, up-to-date information specific to individual agencies, making them insufficient for internal operational support unless complemented by additional mechanisms.

The emergence of Retrieval-Augmented Generation (RAG) architectures constitutes an important advancement in addressing these limitations [19]. RAG combines the generative capabilities of LLMs with external information retrieval systems, grounding model outputs in verified, domain-specific knowledge bases. By retrieving relevant documents before generation, RAG systems significantly reduce hallucinations while enabling LLMs to operate on current, proprietary data [20]. Recent implementations in healthcare, legal services, and enterprise knowledge management have demonstrated RAG's potential for specialized information access [11].

Within tourism research, RAG applications have begun to emerge, primarily focused on enhancing customer services. Studies have explored RAG-based systems for customer support [57], personalized travel recommendations and dynamic itinerary generation [12]. These systems demonstrate improved accuracy over traditional chatbots by anchoring responses in curated tourism databases and real-time availability information.

Despite these advances, a notable research gap persists: the application of RAG architectures to optimize internal operational workflows for travel agency staff remains largely unexplored. While AI designed for costumers has received substantial attention, the potential for RAG systems to serve as intelligent assistants for travel agents themselves, facilitating rapid access to product catalogs, contact databases, procedural documentation, and training resources, has not been adequately investigated. This gap is particularly significant given that agent efficiency directly influences service quality [13,14], and that internal knowledge management challenges are among the most pressing operational issues facing agencies today [15].

 doi:10.20944/preprints202512.1693.v1

Furthermore, existing literature has not sufficiently addressed the integration of RAG systems with structured data sources, such as product databases and contact management systems, nor explored how such systems can extend beyond information retrieval to support training, evaluation, and performance tracking for agency staff. The potential for RAG-based assistants to function not merely as query responders but as comprehensive operational support tools represents an underexplored frontier in tourism technology research.

1.3. Proposed Solution and Contribution

This paper addresses the identified gap by presenting a comprehensive intelligent assistant system designed specifically for travel agency operations, built upon a RAG architecture. The proposed solution integrates three core functional components: (1) an intelligent work assistant providing instant access to internal documentation, product catalogs, and contact databases through natural language queries; (2) an interactive training module enabling agents to engage with educational content via conversational interfaces; and (3) an evaluation and performance tracking system that generates insights into agent knowledge levels and identifies training needs.

The system leverages advanced OCR technology to digitize and process diverse document formats, converts textual information into vector embeddings stored in a specialized database, and employs language models to generate contextually appropriate responses. Deployed on cloud infrastructure to ensure scalability and availability, the assistant provides a unified interface for accessing fragmented internal knowledge, eliminating the need for agents to navigate multiple disconnected systems.

The primary contribution of this research is demonstrating how RAG-based architectures can be effectively adapted to enhance operational efficiency and service quality in travel agencies through internal knowledge management. Unlike previous work focused on customer oriented applications, this study validates the use of RAG for empowering agency staff, thereby improving the human element of service delivery. Additionally, the research contributes methodological insights regarding the integration of RAG with structured databases, the application of OCR for document processing in tourism contexts, and the design of comprehensive training and evaluation modules powered by generative AI.

By addressing the critical challenge of knowledge fragmentation through a controlled, proprietary-data-grounded approach, this work advances both the theoretical understanding of RAG applications in tourism and provides practical insights for technology adoption in travel agencies. The results demonstrate this tasks can be performed in seconds and establish a foundation for future research on AI-enhanced operational workflows in hospitality and tourism services.

 doi:10.20944/preprints202512.1693.v1

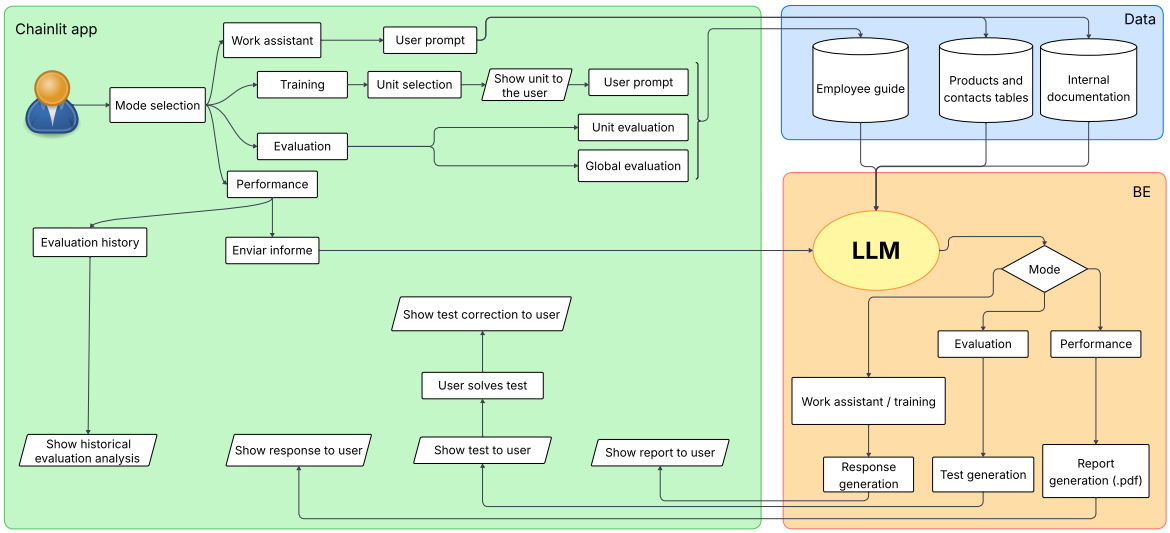

Figure 1. Functional overview illustrating the interaction between the user interface (Chainlit), the four core operational modules (Work Assistant, Training, Evaluation, Performance), and the backend data integration pipeline.

1.4. Paper Structure

The remainder of this paper is organized as follows: Section 2 describes the methodology, detailing the system architecture, RAG implementation, data processing pipeline, and evaluation framework. Section 3 presents the results, including quantitative performance metrics and qualitative feedback from pilot testing. Section 4 discusses the implications of findings, compares results with existing approaches, and addresses limitations. Finally, Section 5 concludes with key takeaways and directions for future research.

2. Materials and Methods

This section describes the methodological framework employed in the development and implementation of an intelligent assistant system for travel agency operations. The system leverages Retrieval-Augmented Generation (RAG) architecture to provide agents with efficient access to fragmented internal knowledge sources through natural language interaction. We detail the data sources utilized, system architecture, document processing pipeline, prompt engineering techniques, cloud deployment infrastructure, and evaluation methodology.

2.1. Data Sources

The intelligent assistant integrates heterogeneous data sources representing the typical information ecosystem of a travel agency. Data were categorized into structured and unstructured formats to enable comprehensive knowledge coverage while addressing diverse information retrieval requirements. The inclusion of structured data sources also facilitates the adoption of the tool by travel agencies, as it is common for such organizations to manage their products and services in formats like Excel or CSV files rather than in unstructured documents.

2.1.1. Structured Data

Structured data was organized in relational database formats and included:

- **Tourism product catalog:** A comprehensive database containing information on travel products across four categories: (1) flights (airline companies, origin-destination pairs, schedules, pricing, booking links, and customer ratings); (2) hotels (location, accommodation type, star category, pet policies, pricing, booking links, and ratings); (3) excursions (activity descriptions, organizing companies, destinations, ratings, and access information); and (4) car rental services (rental companies, service locations, pricing, and booking details). Each product record included observational fields

 doi:10.20944/preprints202512.1693.v1

for expert annotations, enabling agents to access specialized knowledge accumulated through operational experience.

- **Contact database:** A structured table containing client and supplier contact information, including company names, contact persons, positions, email addresses, telephone numbers, physical addresses, websites, and contextual notes regarding the nature of business relationships.

Crucially, implementing this structured data within a relational architecture enables the execution of precise, exhaustive searches across multiple records. This capability addresses a fundamental limitation of standard Vector Database RAG solutions [16]. In vector retrieval, the system requires specifying retrieval constraints: either a fixed number of records (top-k) or a minimum similarity threshold. The top-k approach is infeasible because the volume of relevant records is unknown *a priori* and predefining a big number of records to retrieve could be counterproductive [17] . Conversely, reliance on similarity thresholds is precarious; within high-dimensional embedding latent spaces, vectors often exhibit high density and close proximity [18], rendering static distance metrics unreliable for distinguishing between precise attribute matches and merely semantically related noise.

2.1.2. Unstructured Data

Unstructured textual data comprised internal documentation and training materials:

- **Internal Documentation:** Operational procedures, platform access credentials, return/cancellation /modification policies, destination information, strategic supplier data, and ad-hoc operational guidelines. These documents typically existed in diverse formats (PDF, Word, plain text) and lacked standardized structure.
- **Training and Evaluation Materials:** A centralized knowledge base organized thematically, covering travel product types, internal policies, client management procedures, frequently asked questions, and destination-specific information. This corpus served dual purposes: providing content for the training module and serving as a guide for test generation in the evaluation module.

All data were synthetically generated to faithfully represent real-world travel agency operations while ensuring privacy compliance. To guarantee ecological validity, we established direct collaboration with multiple travel agencies to acquire a reference corpus of actual operational documents. Following a comprehensive analysis of these artifacts, we reconstructed the dataset by replicating their structure and content flow. Crucially, we applied a rigorous anonymization protocol in which all sensitive information, including Personally Identifiable Information (PII) and proprietary commercial data, was systematically substituted with coherent fictitious values. This process preserved the complexity of the original documents without compromising confidentiality..

2.2. System Architecture

The intelligent assistant employs a RAG architecture that combines the generative capabilities of large language models with external information retrieval mechanisms [19], as shown in Figure 2. This approach grounds model outputs in verified, domain-specific knowledge while mitigating hallucination risks inherent in vanilla LLMs [20].

 doi:10.20944/preprints202512.1693.v1

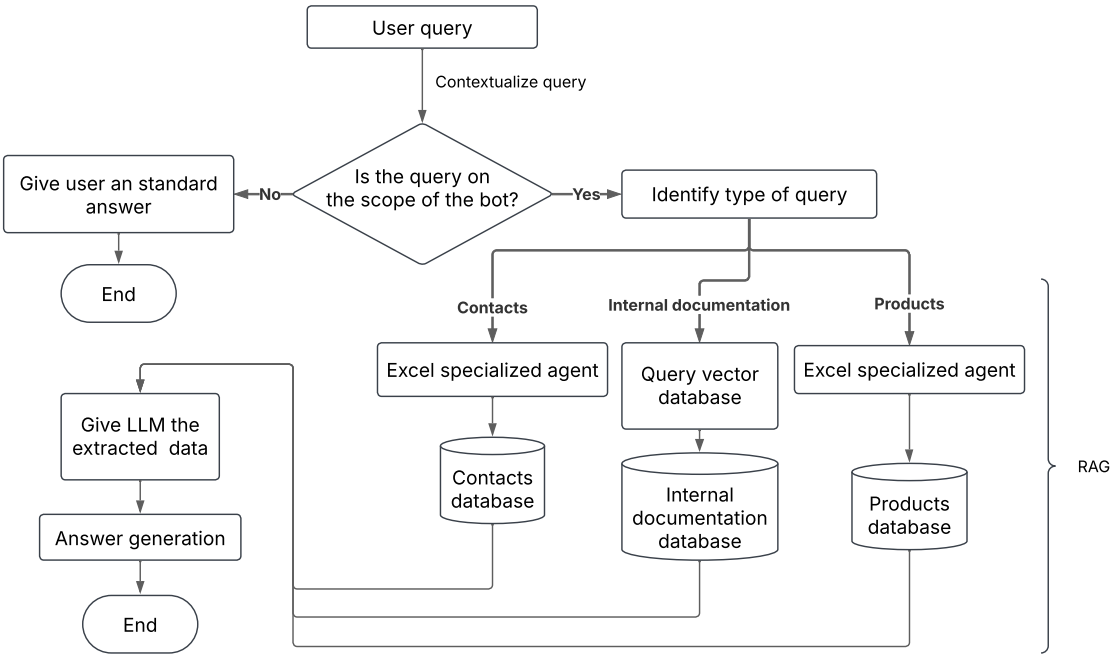

Figure 2. Flowchart demonstrating the RAG decision process, specifically identifying the routing mechanism that directs queries to either vector databases for unstructured documentation or specialized agents for structured product and contact data.

2.2.1. Large Language Model Selection

We selected Gemini 2.0 Flash (Google DeepMind) as the core language model [21]. This decision was informed by a comparative analysis across three key dimensions: performance, cost-effectiveness, and response latency. Table 1 presents a quantitative comparison with leading alternative models.

Table 1. Comparative analysis of large language models for conversational applications (June 2025 pricing).

Model	Input Cost (USD/1M tokens)	Output Cost (USD/1M tokens)	Latency
Gemini 2.0 Flash	0.10	0.40	Very High
GPT-4.1 mini	0.40	1.60	Medium
Claude 3.5 Haiku	0.80	4.00	High

Gemini 2.0 Flash demonstrated superior cost-performance characteristics, offering 4× cost reduction relative to GPT-4.1 mini and 8× reduction relative to Claude 3.5 Haiku, while maintaining competitive performance on standard benchmarks [21]. Critically, the model exhibited exceptionally low latency, which is essential for conversational applications where response delays negatively impact user experience and operational flow.

2.2.2. Embedding Generation and Vector Database

Textual information was transformed into dense vector representations to enable semantic similarity search [22]. The embedding pipeline consisted of four stages:

- **Document preprocessing:** Raw documents underwent cleaning and normalization to remove formatting artifacts while preserving semantic content. Unlike traditional NLP pipelines, we avoided aggressive normalization (e.g., stemming, lemmatization) to retain contextual nuances.
- **Chunking:** Documents were segmented into semantically coherent fragments using configurable strategies. We employed sliding-window chunking technique [23], balancing granularity require-

 doi:10.20944/preprints202512.1693.v1

ments with retrieval precision. Chunk size was empirically optimized to maximize relevance while maintaining sufficient context for coherent answer generation.

- **Vector encoding:** A pre-trained sentence-transformer model [24] converted each text chunk into a fixed-dimensional dense vector ($\mathbb{R}^d$, where d typically ranges from 384 to 3072 dimensions). These embeddings capture semantic relationships such that pieces of documents with similar meanings exhibit high cosine similarity in the embedding space.
- **Indexing:** Embeddings were stored in ChromaDB, a specialized vector database optimized for similarity search operations [25]. Each vector was associated with its source text, document metadata (origin, timestamp, category), and unique identifiers to facilitate retrieval and provenance tracking.

2.2.3. Retrieval Augmented Generation

The RAG pipeline operates through a multi-stage process when responding to user queries:

- **Routing and contextualizing:** System decides whether or not a query needs to be answered by information owned by the company. If it does classify it this way, the system selects whether it consults the internal documentation database, the products database, or the contacts database. There must exist a contextualization of the query to ensure that relevant documents are found when performing the vector search.
- **Query encoding:** If the database to consult is the vector database, user queries are transformed into embeddings using the same embedding model employed during indexing, ensuring consistent semantic representation across query and document spaces.
- **Similarity search:** The system performs Approximate Nearest Neighbor (ANN) search [26] to identify the k most relevant document chunks. We employed the Hierarchical Navigable Small World (HNSW) algorithm for efficient high-dimensional search, using cosine similarity as the distance metric:

$$\text{similarity}(q, d) = \frac{q \cdot d}{\|q\| \|d\|} \tag{1}$$

where q represents the query embedding and d represents a document embedding.

Context construction: The top-k retrieved chunks are extracted with their associated text. These chunks constitute the external knowledge context provided to the LLM. In case of consulting a structured database, what is given to the LLM is all the records that match the user query.

Prompt composition: The system constructs a structured prompt combining: (1) system instructions defining the assistant's role and behavioral constraints; (2) conversation history for contextual continuity; (3) retrieved document chunks or coincident records as grounding context; and (4) the user's current query.

Response generation: The LLM generates a response conditioned on the composed prompt, producing outputs grounded in retrieved information. This approach significantly reduces hallucination while enabling the model to synthesize information across multiple source documents.

2.3. Document Processing with Optical Character Recognition

To accommodate diverse document formats including scanned materials and image-based PDFs, we implemented an OCR-based preprocessing pipeline using MistralAI's OCR model [27].

2.3.1. OCR Model Selection

MistralAI OCR was selected based on three evaluation criteria: (1) recognition accuracy exceeded 99% in internal testing on travel agency documentation; (2) operational costs were competitive with established services (AWS Textract, Google Cloud Vision AI) while offering superior integration efficiency (evaluated on May 2025); and (3) API response latency was minimal, supporting real-time document upload workflows.

 doi:10.20944/preprints202512.1693.v1

2.3.2. OCR Processing Pipeline

The document digitization workflow comprised five stages:

1. **Text extraction:** The MistralAI OCR model performed character recognition on normalized images, extracting textual content with positional information.
2. **Post-processing:** Extracted text was refined using LLM-based error correction. Structured prompts directed the language model to identify and correct OCR artifacts (e.g., character substitution errors, formatting inconsistencies) while preserving semantic content.
3. **Format conversion:** Processed text was saved as plain text (.txt) files and subsequently converted to Markdown (.md) format to facilitate structural annotation and enhance chunking quality and model contextualization. In .md we can preserve the hierarchy of the text (titles, subtitles, numerations...).
4. **RAG integration:** OCR-processed documents entered the standard embedding and indexing pipeline, ensuring seamless integration with non-OCR content in the vector database.

2.4. Prompt Engineering Techniques

Optimal LLM behavior was achieved through systematic prompt engineering [28], employing multiple techniques tailored to specific interaction contexts:

- **Zero-Shot prompting:** For general queries that do not require external knowledge, the system provided task instructions without examples, relying on the model's pre-trained capabilities.
- **Few-Shot prompting:** Complex tasks requiring specific output formats included 1-3 complete input-output examples in the prompt to guide model behavior [3].
- **Chain-of-Thought (CoT) prompting:** For queries demanding multi-step reasoning, prompts explicitly instructed the model to articulate intermediate reasoning steps before providing final answers [29]. This technique significantly improved accuracy on complex information synthesis tasks.
- **Role assignment:** System prompts assigned the model a specific persona (e.g., "Act as an expert travel agent with comprehensive knowledge of agency operations...") to influence response tone, style, and content appropriateness [30].

These techniques were combined adaptively based on query classification, with prompt templates optimized through iterative refinement during system development.

2.5. Cloud Deployment Architecture

The system was deployed on Amazon Web Services (AWS) to ensure scalability, high availability, and continuous integration/deployment capabilities [31].

2.5.1. Containerization and Orchestration

The application was containerized using Docker [32], encapsulating the chatbot interface, backend logic, RAG pipeline, and dependencies in portable images. Container orchestration was managed through Amazon Elastic Container Service (ECS), enabling automated scaling, health monitoring, and resource allocation.

2.5.2. Load Balancing and Traffic Distribution

An Application Load Balancer (ALB) distributed incoming HTTP/HTTPS traffic across multiple container instances, providing fault tolerance and eliminating single points of failure. The load balancer performed health checks and automatically rerouted traffic away from unhealthy instances.

2.5.3. Data Persistence and Concurrency Management

User conversation histories were persisted in Amazon S3 [33], with each user session mapped to a unique object identifier. This architecture enabled: (1) conversation continuity across sessions; (2)

 doi:10.20944/preprints202512.1693.v1

concurrent user support through isolated conversation states; and (3) long-term storage for analytics and system improvement.

2.5.4. Continuous Integration and Deployment

A fully automated CI/CD pipeline was implemented using AWS CodePipeline [34], integrated with a GitHub repository. The pipeline executed three stages: (1) *Source*: monitoring the repository for code changes; (2) *Build*: AWS CodeBuild constructed updated Docker images and executed automated tests; (3) *Deploy*: successful builds triggered automatic deployment to ECS, updating running containers with zero downtime. Configuration and management were simplified using AWS Copilot CLI, which abstracted infrastructure complexity through intuitive command-line interfaces.

2.6. Evaluation Methodology

System efficacy was evaluated through quantitative time-reduction metrics and qualitative assessment of response quality.

2.6.1. System Response Latency Analysis

To evaluate the system's operational viability, we conducted a performance analysis focusing on response latency. We measured the average time elapsed between the user's query submission and the complete generation of the assistant's response. These measurements were categorized across the three primary functional domains: internal documentation queries, product catalog searches, and contact information retrieval. The objective was to determine if the system's processing speed remains within the thresholds required for real-time interaction during active client consultations.

2.6.2. Qualitative Evaluation

Response quality was assessed along three dimensions: (1) *accuracy*—factual correctness of information provided; (2) *relevance*—appropriateness of responses to user queries; and (3) *coherence*—linguistic quality and logical consistency of generated text. Evaluation involved systematic testing with representative queries spanning the system's functional scope, with detailed results presented in the Results section.

3. Results

This section presents the results of implementing the RAG-based intelligent assistant system for travel agency operations. We organize findings according to the four functional modules comprising the system: Work Assistant, Training Module, Evaluation Module, and Employee Performance. Results encompass both quantitative metrics—particularly time efficiency improvements—and qualitative assessments of functional capabilities and operational impact.

3.1. System Overview

The implemented system integrates four complementary modules within a unified conversational interface (Figure 3). The **Work Assistant** serves as the core operational support tool, providing natural language access to internal documentation, product catalogs, and contact databases through RAG architecture. The **Training Module** facilitates continuous learning by enabling interactive engagement with educational materials. The **Evaluation Module** allows agents to assess knowledge acquisition through diverse assessment formats with immediate feedback. Finally, the **Employee Performance** module provides consolidated analytics on evaluation results, supporting both individual progress tracking and organizational talent management. This modular architecture ensures comprehensive operational support while promoting professional development within the agency context.

 doi:10.20944/preprints202512.1693.v1

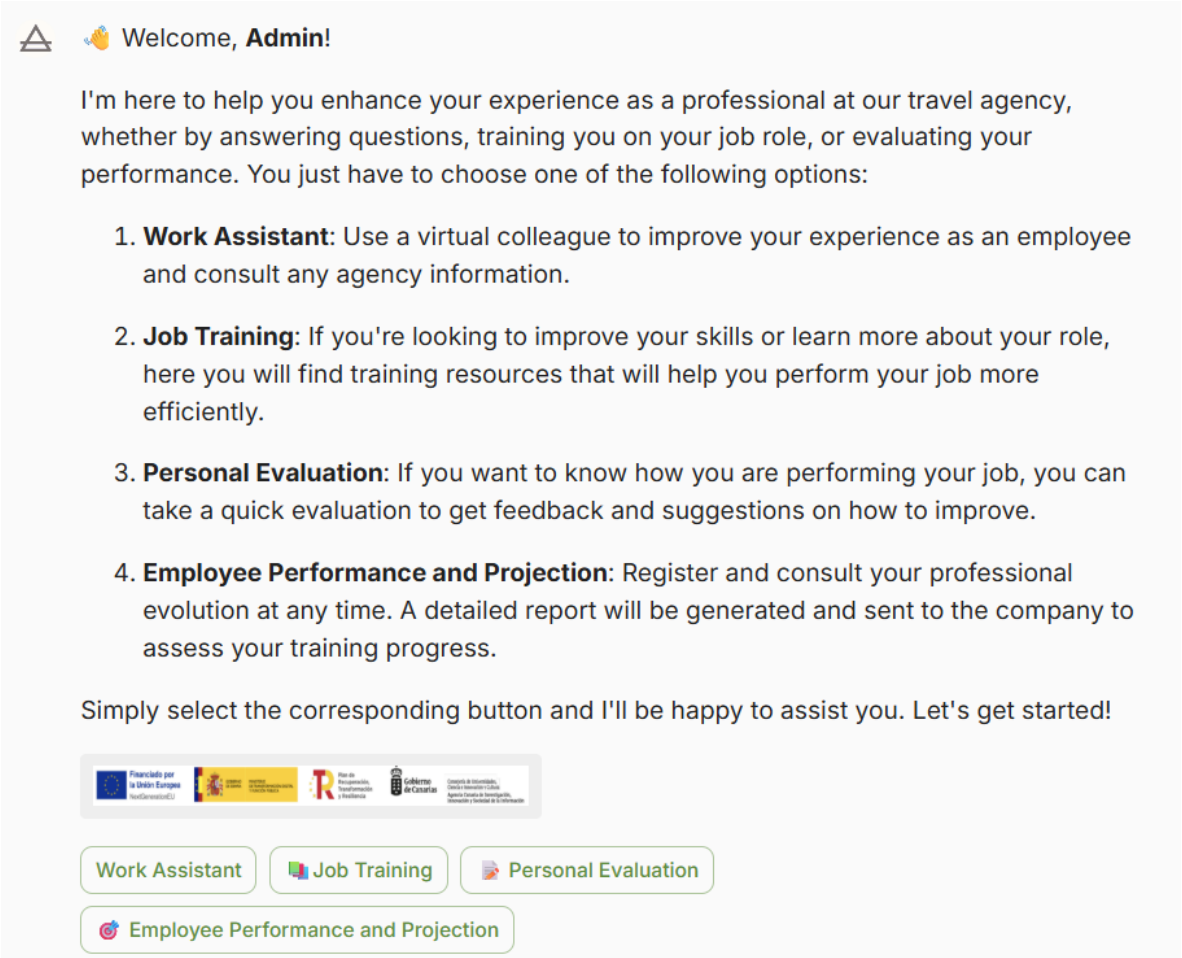

Figure 3. Home screen of the virtual assistant application for travel agents. The functionality of each module within the tool is explained to the user; additionally, they are identified by name at the start to provide greater personalization.

3.2. Work Assistant Results

The Work Assistant demonstrated substantial improvements in information retrieval efficiency across three primary functional domains: internal documentation queries, product catalog searches, and contact information retrieval. We present quantitative time metrics and qualitative functional assessments for each domain.

3.2.1. Internal Documentation Queries

The system enables agents to query policies, regulations, and operational procedures using natural language. The RAG architecture retrieves relevant document fragments from the vector database and generates contextualized responses grounded in official documentation.

Quantitative results: The intelligent assistant achieved an average response latency of 2–5 seconds for policy and procedure queries. In addition, the system maintained a 95% accuracy rate on questions related to internal documentation.

Qualitative results: Responses exhibited high accuracy, maintaining fidelity to source documentation while presenting information in accessible language. Critically, the assistant reduces interpretation errors that can occur when agents manually parse complex policy language under time pressure during client interactions.

Illustrative example: Consider a query regarding cancellation policies: "What is our agency's cancellation policy for hotel reservations?" The system retrieves relevant fragments from internal policy documents, identifying specific timeframes (e.g., cancellations more than 30 days prior receive 90% refund, 15-30 days receive 50% refund, less than 15 days incur full charges unless justified). The LLM

 doi:10.20944/preprints202512.1693.v1

synthesizes this information into a coherent response structured by timeframe, with clear explanation of refund percentages and exceptional circumstances. The agent receives accurate, immediately actionable information without consulting multiple documents or colleagues.

3.2.2. Product Catalog Queries

The system facilitates searches across structured tables containing tourism products: hotels, flights, excursions, and car rentals. Notably, product records include an "Observations" field containing expert annotations from experienced agents, effectively democratizing specialist knowledge across the workforce.

Quantitative results: For structured product searches, the system demonstrated response times ranging from 5 to 45 seconds, depending on query complexity.

Qualitative results: The system enables instantaneous multi-option comparison with access to expert insights previously siloed among senior agents, which from our point of view gives enormous value to the solution. The integration of observational notes (often containing critical information about seasonal availability, local partnerships, or service quality nuances) significantly enhances recommendation quality. Summing up, agents can effectively "consult all agency experts simultaneously," substantially improving sales conversation quality and conversion potential.

Illustrative example: For the query "What hotels do we recommend in Ecuador for families with children?", the system retrieves relevant hotel records filtered by destination (Ecuador) and enhanced by observational notes. A sample response might identify three properties, specifying for each: location, star category, family-friendly amenities (e.g., connecting rooms, children's activities, swimming pools), pricing range, and crucially, expert observations such as "Excellent relationship with hotel management; can arrange special rates for groups" or "Popular with returning clients; outstanding kids' club with bilingual staff." This level of detailed, contextualized information would be nearly impossible to compile manually within a client conversation time-frame.

3.2.3. Contact Information Retrieval

The system provides rapid access to client and supplier contact tables, extracting specific information (telephone, email, address, contact person) in structured format.

Quantitative Results: The assistant successfully retrieved specific contact details with an average latency of 2–15 seconds, presenting the data in a structured format immediately available to the agent..

Qualitative Results: Instantaneous contact access during client interactions eliminates workflow interruptions and maintains conversation flow. This is particularly critical for time-sensitive communications (e.g., confirming availability during active booking discussions) where retrieval delays can negatively impact client experience.

Illustrative Example: The query "What is the telephone number for the sales manager at Iberia?" prompts the system to identify the corresponding table entry for Iberia (airline company), locate the sales manager contact (Laura Fernández), and return: telephone (+34 960 111 500), email (atencion@iberia.com), position (Sales Manager), and contextual note ("Negotiation of group tariffs and agreements"). The structured presentation ensures the agent immediately obtains the required information without manual database navigation.

3.2.4. Overall Work Assistant Impact

Table 2 presents a quick view of task completion times across the three functional domains.

Table 2. Retrieval task completion times. The difference on time within a functional areaare due to task complexity.

Functional Area	Assistant Time
Internal Documentation	2 - 5 seconds
Product Catalog	5 - 45 seconds
Contact Information	2 - 15 seconds

 doi:10.20944/preprints202512.1693.v1

The Work Assistant consistently delivered information within seconds across all functional domains. This low latency confirms the system's capability to support real-time workflows without introducing operational delays. This capability could lead to some improvements in the day-to-day work:

1. **Enhanced operational efficiency** Intensive tasks are completed in seconds enabling higher client throughput without quality degradation.
2. **Liberation from repetitive tasks** Time that was previously spent on monotonous information search can now be reallocated to high-value activities, such as personalized service, relationship building, and sales closure.
3. **Improved service quality** Instantaneous access to expert knowledge during client interactions enables more informed recommendations and responsive service delivery, directly elevating the customer experience.

3.3. Training Module Results

The Training Module transforms passive information consumption into interactive learning experiences, facilitating continuous professional development for agency staff.

3.3.1. Functional Description

The module presents educational content organized in thematic blocks (Figure 4), covering topics such as travel product types, cancellation policies, client management procedures, frequently asked questions, and destination information. Within each theme, agents access structured content and engage with an interactive chat interface powered by RAG, enabling contextualized questions about the material.

Figure 4. Main view of the Training module, showing the organization of topics available to the agent.

3.3.2. Qualitative Results

The conversational learning paradigm offers several advantages over traditional training approaches:

1. **Active learning vs Passive learning:** Rather than passively reading documentation, agents actively engage through questioning, which enhances retention and comprehension. The system adapts to individual learning needs, providing elaboration on concepts requiring clarification.
2. **Learning personalization:** Agents control learning pace and depth, focusing on challenging concepts while moving efficiently through familiar material. This self-directed approach accommodates diverse learning styles and prior knowledge levels.
3. **Training Efficiency:** Contextualized responses drawn directly from training materials ensure consistency with organizational procedures while eliminating the latency of instructor availability. Complex concepts receive immediate clarification through examples and multi-perspective explanations.

 doi:10.20944/preprints202512.1693.v1

4. **Illustrative example:** Consider an agent reviewing cancellation policy training materials who asks: "What happens if a client cancels a non-refundable booking due to a medical emergency?" The system retrieves relevant policy sections addressing exceptional circumstances, explaining: (1) the standard non-refundable policy; (2) the exception process for documented medical emergencies; (3) required documentation (medical certificates); (4) approval workflow; and (5) precedent handling. This contextual depth—typically requiring instructor consultation—becomes immediately accessible.

3.3.3. Operational Impact

The Training Module contributes to organizational capability in three critical dimensions:

1. **Continuous learning culture** An accessible interface encourages ongoing skill development and knowledge refresh beyond formal training sessions, fostering a culture of perpetual improvement among agents.
2. **Onboarding acceleration** New employees can access comprehensive knowledge resources independently, significantly reducing time-to-productivity (T_{tp}) and decreasing the workload on mentors and senior staff.

$$T_{\text{tp, new}} < T_{\text{tp, old}}$$

3. **Knowledge currency and alignment** When policies or procedures change, updated training materials immediately become accessible through the conversational interface. This ensures workforce alignment with the most current standards and minimizes operational risk due to outdated information.

3.4. Evaluation Module Results

The Evaluation Module enables systematic knowledge assessment through diverse question formats, supporting both individual learning validation and organizational competency management.

3.4.1. Functional Description

Agents access two evaluation modalities: *Objective-Based Evaluation*, focusing on specific thematic areas, and *Global Evaluation*, comprehensively assessing knowledge across all training content. Question types include true/false items, multiple-choice questions, open-ended responses, and case-based scenarios requiring applied reasoning. Upon test completion, the system provides immediate feedback identifying correct responses, explaining errors, and highlighting knowledge gaps.

3.4.2. Qualitative Results

The evaluation functionality serves multiple organizational objectives:

1. **Learning validation:** Agents obtain objective measures of knowledge acquisition and content mastery, identifying topics requiring additional study before real-world application.
2. **Reinforcement tool:** The detailed feedback process—reviewing questions, analyzing incorrect responses, understanding correct answers—functions as a learning mechanism itself, reinforcing key concepts through spaced repetition and error correction.
3. **Competency gap identification:** Systematic evaluation results reveal knowledge deficits at both individual and collective levels. Aggregate performance patterns enable the agency to identify widespread comprehension challenges, informing targeted training interventions.
4. **Assessment diversity:** The inclusion of case-based scenarios requiring applied judgment mirrors real-world complexity, assessing not merely factual recall but practical decision-making capability under realistic constraints.

3.4.3. Organizational Impact

The Evaluation Module contributes to three key strategic outcomes within the organization:

 doi:10.20944/preprints202512.1693.v1

1. **Quality assurance (QA)** Systematic competency assessment ensures that agents meet all organizational knowledge standards and compliance requirements *before* being assigned client-facing responsibilities. This serves as a critical knowledge gate.
2. **Onboarding validation** New employees formally demonstrate their readiness through objective assessments. This provides clear, data-driven support for hiring decisions and confirms a satisfactory transition from training to active duty.
3. **Continuous improvement culture** Regular evaluation normalizes assessment as a developmental tool rather than a punitive measure. By focusing on identifying knowledge gaps for future training, the module fosters a proactive growth mindset among the workforce.

3.5. Employee Performance Results

The Employee performance module synthesizes evaluation data into actionable insights for both individual agents and agency management.

3.5.1. Functional Description

Individual agents access performance dashboards displaying key metrics: average scores, temporal progression through line graphs (Figure 5, and test-by-test performance breakdowns. These visualizations enable agents to track learning trajectories and identify improvement areas. Complementarily, the system generates comprehensive reports for agency management, providing detailed analyses identifying strengths, weaknesses, and personalized development recommendations for each staff member.

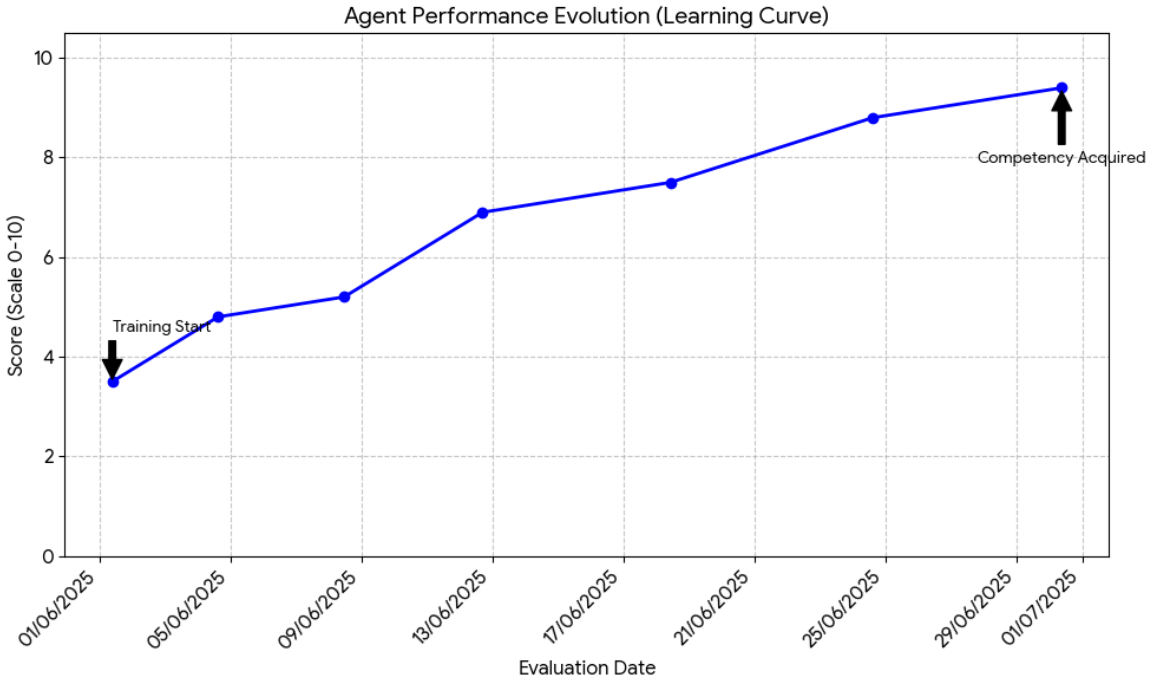

Figure 5. Longitudinal analysis of agent performance scores over a one-month training period. The upward trend demonstrates the effectiveness of the interactive training module in closing knowledge gaps and improving evaluation outcomes.

3.5.2. Qualitative Results

The performance tracking system delivers value through multiple mechanisms:

1. **Progress visualization:** Graphical representations of performance evolution provide concrete evidence of improvement, enhancing motivation and engagement with professional development activities.

 doi:10.20944/preprints202512.1693.v1

2. **Transparency and motivation:** Performance visibility creates accountability while enabling agents to take ownership of professional development. The detailed feedback and improvement tracking foster intrinsic motivation for continuous learning.
3. **Early intervention capability:** Performance monitoring enables early identification of struggling employees, allowing supportive interventions before competency gaps impact client service quality or sales performance.

3.5.3. Organizational Impact

The Performance module transforms learning assessment from episodic evaluation to continuous capability management. Key outcomes include:

1. **Data-driven development planning** Training investments are precisely targeted toward empirically identified needs rather than general assumptions. This ensures optimal resource allocation for maximum return on investment (ROI) in agent capability.
2. **Professional growth culture** Visible progress tracking and detailed, objective feedback position learning and development as a valued organizational priority, actively encouraging agents toward continuous professional growth.
3. **Quality consistency** Ongoing competency monitoring ensures consistent adherence to service quality standards across the entire workforce, regardless of the agent's tenure or current experience levels.

3.6. Overall Impact and Synthesis

The implemented RAG-based intelligent assistant system successfully achieved the foundational objectives established at project inception. The system's success can be synthesized across five core dimensions:

- **Knowledge centralization:** The system provides unified access to previously fragmented information across documentation repositories, product databases, and contact records, eliminating the need for agents to navigate multiple disconnected systems.
- **Operational efficiency enhancement:** Quantitative results demonstrate **95** − **99**% time reduction and 95% accuracy in information retrieval tasks (Table 2), translating to substantial productivity gains and enabling agents to serve more clients without sacrificing service quality.
- **Service quality improvement:** Instantaneous access to expert knowledge and comprehensive product information during client interactions enables more informed recommendations and responsive service delivery, potentially increasing conversion rates and customer satisfaction.
- **Continuous professional development:** The integrated Training, Evaluation, and Performance modules create a comprehensive learning ecosystem, supporting both initial onboarding and ongoing skill development throughout employee tenure.
- **Strategic knowledge management:** The system transforms tacit knowledge held by senior agents into accessible organizational assets through expert annotations in product catalogs and structured training content, reducing dependency on individual expertise and facilitating knowledge transfer.

Emergent benefits and resilience

Beyond achieving stated objectives, the system delivers additional strategic value through two emergent benefits:

- *Democratization of expertise:* Junior agents gain immediate access to senior agent knowledge through observational annotations and comprehensive training materials, significantly accelerating capability development.
- *Organizational resilience:* The codification of operational knowledge in accessible formats reduces vulnerability to staff turnover and facilitates overall business continuity.

 doi:10.20944/preprints202512.1693.v1

The results validate the hypothesis that RAG-based architectures effectively address knowledge fragmentation challenges in travel agency operations, delivering measurable efficiency improvements while supporting comprehensive talent development and organizational learning objectives.

4. Discussion

This study demonstrates the successful implementation of a RAG-based intelligent assistant that achieved low-latency responses and 95% accuracy in information retrieval tasks while supporting continuous professional development in travel agency contexts. The upcoming section interprets these findings, contextualizes them within recent literature, examines practical and theoretical implications, acknowledges limitations, and identifies future research directions.

4.1. Interpretation of Results

4.1.1. Operational Efficiency and Agent Work Transformation

The efficiency gains observed across all functional domains represent more than incremental improvement; they signify a fundamental transformation in the nature of agent work. Manual information retrieval previously dominated a significant portion of agent time allocation, often relegating relationship-building to residual time slots. The assistant's near-instantaneous responses invert this priority structure, enabling agents to dedicate cognitive resources to high-value activities: understanding nuanced client preferences, providing strategic travel advice, and cultivating long-term relationships [35].

This transformation aligns with theoretical frameworks positioning AI as augmentation rather than replacement of human capabilities. The system handles routine information retrieval, tasks where machines excel, while agents focus on empathy, persuasion, and complex problem-solving. These last-named tasks are where human judgment remains superior [36].

The efficiency gains also carry implications for business economics. Streamlined information retrieval enables higher client throughput without proportional workforce expansion, improving operational margins. Moreover, faster response times during client interactions may reduce booking abandonment rates—a critical metric in conversion optimization [37].

4.1.2. RAG Architecture as a Solution to Knowledge Fragmentation

The success of the RAG architecture in addressing knowledge fragmentation validates recent theoretical work positioning retrieval-augmented generation as superior to fine-tuning for dynamic knowledge domains [20,38]. Unlike fine-tuning, which embeds knowledge within model parameters, requiring expensive retraining for updates. RAG maintains knowledge in external, easily updatable repositories. This architectural choice proves particularly appropriate for travel agencies, where product availability, pricing, and policies change continuously.

The integration of structured (product databases) and unstructured (textual documentation) data sources within a unified conversational interface represents a methodological contribution extending existing RAG research, which typically focuses on single data modalities [20]. This heterogeneous data integration mirrors real organizational knowledge ecosystems, suggesting broader applicability beyond travel agencies.

The incorporation of OCR for digitizing legacy documentation further enhances practical utility. Many small and medium enterprises maintain critical information in scanned or paper formats. Our OCR pipeline demonstrates that even organizations lacking comprehensive digital archives can leverage RAG technologies, lowering adoption barriers for traditional businesses [53].

4.1.3. Training, Evaluation, and Performance Modules as Innovation in Professional Development

The training and evaluation modules represent a novel application of RAG to workplace learning, an area receiving limited attention in tourism AI research. While most LLM applications in tourism focus on customer-facing services [5,39], our system demonstrates RAG's potential for employee capability development.

 doi:10.20944/preprints202512.1693.v1

The transformation of passive documentation review into interactive dialogue addresses fundamental limitations of traditional training approaches. Adult learning theory emphasizes active engagement, immediate feedback, and self-directed learning as critical for knowledge retention [47]. The conversational interface operationalizes these principles: agents actively query content, receive immediate contextualized responses, and control learning pace and depth. This alignment with pedagogical best practices suggests potential learning outcome improvements beyond mere convenience.

The evaluation module's systematic competency assessment enables evidence-based talent management [56]. The immediate feedback mechanism serves dual purposes: validating learning for individual agents and generating aggregate data revealing organizational knowledge gaps. This data-driven approach to training needs assessment represents a significant advancement over traditional methods relying on manager intuition or periodic surveys.

The performance tracking system implements what organizational learning literature terms "learning analytics": systematic measurement of learning processes and outcomes to optimize educational interventions [46]. By visualizing progress trajectories and identifying struggling employees early, the system enables proactive rather than reactive talent development, potentially reducing turnover costs and improving service quality consistency.

4.2. Contextualization with Existing Literature

4.2.1. Comparison with LLM Applications in Tourism

Literature on LLMs in tourism and hospitality predominantly emphasizes customer oriented applications: chatbots for booking assistance [40], personalized itinerary generation [12], and sentiment analysis of reviews [41]. While valuable, this research focus leaves internal operational applications underexplored. Our work addresses this gap by demonstrating how RAG-based systems enhance agent productivity and organizational knowledge management, a perspective shift from AI as customer interface to AI as employee empowerment tool.

This distinction carries theoretical significance. Customer-facing chatbots must handle unpredictable queries, diverse communication styles, and potentially adversarial interactions. Employee-facing assistants operate in more constrained, domain-specific contexts with cooperative users sharing organizational vocabulary. These differing requirements suggest that employee-facing applications may achieve higher accuracy and reliability at lower development costs—an insight with implications for AI adoption prioritization in resource-constrained organizations [43].

Our findings complement recent work on RAG for enterprise knowledge management [11,42], extending it to tourism-specific contexts. While enterprise applications often focus on software development or legal compliance, travel agencies present unique challenges: highly dynamic information, integration of structured and unstructured data, and critical dependence on tacit expertise. Our successful implementation in this context demonstrates RAG's versatility across diverse organizational settings.

4.2.2. Empirical Validation of RAG Advantages

Theoretical literature increasingly argues for RAG superiority over fine-tuning in dynamic knowledge domains [19,20], but empirical validation in real-world operational contexts remains limited. Our study provides quantitative evidence supporting these claims: the system achieved exceptional time efficiency without model retraining, and updates to knowledge bases (product catalogs, policy documents) integrated seamlessly without system reconfiguration.

The cost-effectiveness dimension merits particular emphasis. Fine-tuning Gemini 2.0 Flash would require substantial computational resources and periodic retraining as information changes. In contrast, our RAG implementation required only database updates—a task manageable by non-technical staff. This operational simplicity democratizes AI adoption for small agencies lacking dedicated machine learning teams, addressing a critical barrier to AI diffusion in tourism SMEs [52].

The comparison with model selection alternatives (GPT-4.1 mini, Claude 3.5 Haiku) presented in our methodology demonstrates another key advantage: cost optimization through strategic model

 doi:10.20944/preprints202512.1693.v1

choice. By selecting Gemini 2.0 Flash based on cost-performance analysis, we achieved 4–8× cost reduction relative to alternatives while maintaining competitive quality—a practical consideration often overlooked in academic research favoring performance maximization over cost-effectiveness.

4.2.3. Implications for Digital Transformation in Tourism

Digital transformation literature in tourism emphasizes customer experience enhancement through technologies like AI, IoT, and blockchain [35,48]. Our work contributes a complementary perspective: digital transformation centered on employee enablement. This distinction matters because service quality ultimately depends on frontline employee capabilities—a reality sometimes obscured by technology-centric narratives.

The concept of "democratizing expertise" through observational annotations in product databases exemplifies this employee-centric approach. Senior agents' tacit knowledge, traditionally transferred through mentorship or accumulated through years of experience, becomes immediately accessible to junior colleagues. This knowledge externalization aligns with Nonaka and Takeuchi's SECI model [44], specifically the socialization-to-externalization transition: tacit knowledge held by individuals becomes explicit organizational knowledge accessible to all.

This capability addresses a persistent challenge in tourism organizations: high employee turnover and consequent knowledge loss. By codifying expertise in searchable, retrievable formats, the system reduces organizational vulnerability to staff departures and accelerates new employee onboarding—critical advantages in industries characterized by seasonal employment and limited training budgets [**?**].

4.3. Practical Implications

4.3.1. Implications for Travel Agencies

Travel agencies adopting similar systems may realize multiple competitive advantages:

Service differentiation: In markets where basic information is offered through online platforms, the ability to provide rapid, expert-informed, personalized advice differentiates human agents from automated alternatives. The system positions agents as knowledgeable consultants rather than mere transaction processors, justifying service fees and building client loyalty [50].

Operational scalability: Traditional agency growth requires proportional workforce expansion, which is a costly and slow process. The intelligent assistant enables productivity multiplication: agents handle higher query volumes and serve more clients simultaneously without quality degradation. This operational leverage improves profitability and enables growth without corresponding increases in labor costs.

Knowledge retention: The system transforms organizational knowledge from primarily tacit (residing in employees' minds) to explicit (codified in searchable databases). This transformation reduces dependency on individual experts and mitigates knowledge loss from employee turnover [49].

Training efficiency: The integrated training and evaluation modules reduce new employee time-to-productivity. Rather than shadowing experienced colleagues for weeks, new hires access comprehensive knowledge resources through conversational interfaces, potentially halving onboarding duration and cost.

4.3.2. Implications for the Tourism Sector

The architectural approach generalizes beyond travel agencies to other tourism organizations facing similar knowledge fragmentation challenges: hotels managing property information and local recommendations, tour operators coordinating complex multi-supplier itineraries, and destination marketing organizations synthesizing diverse stakeholder information. Dispersed, multi-format information requiring rapid retrieval characterizes the sector broadly [6].

The training and evaluation modules suggest pathways toward sector-wide professionalization. Tourism education often struggles with practice-theory integration; AI-powered training, grounded in real operational contexts, could bridge this gap improving workforce quality across the industry. [51].

 doi:10.20944/preprints202512.1693.v1

4.3.3. Implications for AI Adoption in SMEs

Small and medium enterprises often perceive advanced AI as accessible only to technology giants with substantial R&D budgets. Our implementation challenges this perception, demonstrating that SMEs can leverage state-of-the-art language models through cloud APIs, avoiding infrastructure investments and specialized employee requirements [52].

The total cost structure (primarily API calls, cloud hosting, and initial development) proves manageable for businesses processing hundreds to thousands of customer interactions monthly. This economic viability, combined with documented efficiency gains, provides a replicable template for SME AI adoption across sectors, potentially accelerating AI diffusion beyond large enterprises [54].

4.4. Theoretical Implications

4.4.1. Contribution to Knowledge Management Theory

The system operationalizes key concepts from knowledge management theory, particularly Nonaka and Takeuchi's SECI model of knowledge creation [44]. The RAG architecture facilitates multiple SECI transitions:

- **Socialization:** capturing expert agents' **tacit knowledge** through observational annotations.
- **Externalization:** codifying this knowledge in searchable text.
- **Combination:** synthesizing information from multiple sources in responses.
- **Internalization:** enabling knowledge acquisition through interactive training.

This research updates classic knowledge management theory for the AI age. We show that generative models are dynamic systems that actively manage knowledge, unlike the static repositories of the past. The key improvement is the conversational interface, which allows users to obtain the information they need intuitively and based on their context, making knowledge access far superior to manual searching[45].

4.4.2. Contribution to AI in Tourism Research

The study expands AI tourism research from predominantly customer-centric to employee-centric applications, identifying operational efficiency and workforce development as high-impact intervention points. This perspective shift has implications for research prioritization: while AI designed customer use attracts attention due to direct revenue connections, employee-facing AI may offer higher ROI through productivity multiplication and quality improvement [43].

The empirical validation of RAG in tourism contexts provides domain-specific evidence, complementing generic enterprise AI research. Tourism's unique characteristics (dynamic information, heterogeneous data sources, integration of structured and unstructured data, and dependence on tacit expertise) create requirements that may not generalize from other sectors. Demonstrating RAG effectiveness in this specific context advances the understanding of AI applicability across diverse organizational environments.

4.5. Limitations

4.5.1. Methodological Limitations

The quantitative evaluation, while demonstrating substantial efficiency gains, relies on estimated average completion times for manual tasks rather than controlled experimental measurements with multiple participants. Future research should employ rigorous time-motion studies comparing matched agent pairs (with and without system access) performing identical tasks under controlled conditions to validate reported improvements.

The study lacks direct measurement of user satisfaction, system usability, and long-term adoption rates. While efficiency gains appear substantial, actual adoption depends on agent trust in AI outputs, interface usability, and integration with existing workflows [55].

Most critically, we did not measure downstream business impacts: client satisfaction, booking conversion rates, revenue per transaction, or customer retention. While the logic chain connecting

 doi:10.20944/preprints202512.1693.v1

information access speed to service quality to business outcomes appears sound, empirical validation remains necessary. Future implementations should incorporate A/B testing methodologies to assess business-level impacts.

4.5.2. System Limitations

RAG architecture substantially reduces, but does not eliminate, hallucination risks. When retrieved documents provide ambiguous or contradictory information, LLMs may generate plausible-sounding but incorrect syntheses. Operational deployment requires human-in-the-loop validation for consequential decisions, adding friction that may reduce efficiency gains.

System effectiveness depends critically on the quality, completeness, and currency of the knowledge base. Outdated documentation or incomplete product catalogs degrades response accuracy. Maintaining these knowledge bases requires ongoing organizational commitment, a hidden operational cost that may challenge resource-constrained SMEs.

The system was implemented in Spanish for a Spanish-speaking market. International agencies serving a multilingual clientele would require cross-lingual adaptations, potentially including translation layers, although the model used is already multilingual. While technically feasible, this adds complexity and cost that our implementation did not address.

Integration with existing travel agency systems, such as Global Distribution Systems (GDS), Customer Relationship Management (CRM), or Property Management Systems (PMS), was not implemented. Full operational deployment would require API connections that enable the assistant to query real-time availability, pricing, and booking status. This integration complexity represents a significant barrier to real-world adoption not reflected in our prototype.

4.5.3. Generalization Limitations

The study examines a single implementation context: travel agency operations. While we argue that findings generalize to similar tourism organizations, empirical validation across diverse contexts (hotels, tour operators, destination marketing) remains necessary. Organizational size, operational complexity, and technological sophistication may moderate system effectiveness in ways that our single case study cannot reveal.

The use of synthetic data, designed to represent realistic operational scenarios, limits external validity. Real world implementations face messier data: inconsistent formatting, incomplete records, contradictory information, and legacy systems with incompatible data structures. System robustness to these real world imperfections requires validation with authentic organizational data under actual operational conditions.

4.6. Future Research Directions

4.6.1. Robust Empirical Evaluation

Future research should employ controlled experiments with representative agent samples, measuring not only efficiency but also user experience (satisfaction, trust, perceived usefulness) and business outcomes (client satisfaction, conversion rates, revenue impacts). Longitudinal studies tracking adoption over months or years would reveal whether initial enthusiasm translates to sustained usage and whether efficiency gains persist or diminish as novelty fades [55].

Comparative studies evaluating alternative architectural choices (RAG vs. fine-tuning vs. hybrid approaches) under controlled conditions would provide evidence-based guidance for system designers. Similarly, ablation studies isolating the contributions of individual components (embeddings quality, retrieval algorithms, and prompt engineering techniques) would identify optimization priorities.

4.6.2. System Extensions

Several promising extensions merit investigation: *Multimodal capabilities* integrating image and video understanding could enable queries about destinations using photos or video tours [58]. *Predictive analytics* leveraging conversation data could forecast customer needs, suggest proactive recom-

 doi:10.20944/preprints202512.1693.v1

mendations, or identify cross-selling opportunities. *Personalization* adapting response style and detail level to individual agent experience and its preferences could further enhance usability.

Integration with live booking systems would transform the assistant from an information provider to a transaction facilitator, enabling conversational booking workflows. This evolution from "answering questions" to "completing tasks" represents a fundamental capability expansion worthy of focused research [59].

4.6.3. Generalization and Validation

Implementing and evaluating the system across diverse tourism contexts would assess its ability of generalization and identify specific context adaptation requirements. Cross-sectoral applications in industries with similar knowledge fragmentation challenges (insurance, healthcare, legal services) could validate broader architectural principles.

Comparative studies across SMEs of varying sizes and technological maturity would reveal adoption barriers and success factors, informing implementation guidance for practitioners considering AI adoption [52].

4.6.4. Organizational Change and Adoption Research

The sociotechnical dimensions of AI adoption deserve focused attention. Research questions include: What organizational factors (leadership support, change management strategies, training approaches) predict successful adoption? How do power dynamics and job security concerns influence agent acceptance? What implementation strategies maximize adoption while minimizing resistance?

Longitudinal case studies examining cultural transformation accompanying AI implementation would reveal how such systems reshape organizational norms, workflow patterns, and professional identities [43].

4.6.5. Ethics and Trust Research

Critical questions surrounding AI ethics in service contexts require examination: Under what circumstances should agents override AI recommendations? How should responsibility be allocated when AI-informed decisions produce negative outcomes? What transparency mechanisms build appropriate trust levels—neither blind acceptance nor unwarranted skepticism?

The employment implications of productivity-multiplying AI merit serious consideration. While our framing emphasizes augmentation over replacement, business pressures may drive workforce reductions once efficiency gains materialize. Research examining how organizations navigate these tensions would inform policy discussions and ethical guidelines [54,60].

5. Conclusions

This study demonstrates that RAG-based intelligent assistants can dramatically improve operational efficiency in travel agencies while supporting continuous professional development. The results contribute to tourism AI research by shifting the focus from customer-facing to employee-facing applications, provide empirical validation of RAG advantages in dynamic knowledge domains, and offer a replicable template for SME AI adoption. However, acknowledged limitations such as particularly regarding empirical rigor, system robustness, or generalization, indicate substantial opportunities for future research to build upon this foundation. As the tourism industry navigates ongoing digital transformation, AI systems that empower rather than replace human workers offer promising pathways toward enhancing both business performance and employment quality.

Author Contributions: Conceptualization, Emilio Soria-Olivas, Manuel Sánchez-Montañés and Edu William-Secin; methodology, Pablo Vicente-Martínez; software, Pablo Vicente-Martínez; validation, Emilio Soria-Olivas, Manuel Sánchez-Montañés, María Ángeles García-Escrivà and Pablo Vicente-Martínez; formal analysis, María Ángeles García-Escrivà and Inés Esteve Mompó; investigation, Pablo Vicente-Martínez; resources, Pablo Vicente-Martínez; data curation, Pablo Vicente-Martínez; writing—original draft preparation, Emilio Soria-Olivas, Pablo

 doi:10.20944/preprints202512.1693.v1

Vicente-Martínez and Inés Esteve Mompó; writing—review and editing, Emilio Soria-Olivas and Manuel Sánchez-Montañés, Pablo Vicente-Martínez; visualization, Pablo Vicente-Martínez.; supervision, Emilio Soria-Olivas. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: 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.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Buhalis, D., & Leung, M. (2023). Smart hospitality: from smart cities and smart tourism towards agile business ecosystems in networked destinations. *International Journal of Contemporary Hospitality Management*, Vol. *35*. https://doi.org/10.1108/IJCHM-04-2022-0497
2. Gu, Shengyu. (2024). A Survey of Large Language Models in Tourism (Tourism LLMs). *Qeios*. https://doi.org/10.32388/8R27CJ
3. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *In Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 1877–1901). Curran Associates Inc.
4. Bulchand-Gidumal, J., William Secin, E., O'Connor, P., & Buhalis, D. (2023). Artificial intelligence's impact on hospitality and tourism marketing: exploring key themes and addressing challenges. *Current Issues in Tourism*, *27*(14), 2345-2362. https://doi.org/10.1080/13683500.2023.2229480
5. Alawami, A., Alawami, M., Obaid, A., Al-rushaydan, M., Boudjedra, M. L., Bou-djedra, B. E., Alharbi, M., Alzoori, H., Ab-osaif, S., Bshir, G., Alawami, H., & Abufawr, M. (2025). A Systematic Review of AI-Driven Innovations in the Hospitality Sector: Implications on Restaurant Management. *American Journal of Industrial and Business Management*, *15*, 30-40.https://doi.org/10.4236/ajibm.2025.151003.
6. Sabre. (July 22, 2025). *Information overload: Sabre's global travel agency survey finds content fragmentation is driving up agency costs and undermining customer experience* [Press release]. https://www.sabre.com/insights/releases/information-overload-sabres-global-travel-agency-survey-finds-content-fragmentation-is-driving-up-agency-costs-and-undermining-customer-experience(accessed on 30 October 2025)
7. Gretzel, U., Buhalis, D., & O'Connor, K. (2024). Technology-mediated tourism experiences: Designing for operational efficiency and service quality. *Annals of Tourism Research*, *105*, 103705.
8. Chen, W., Konar, R., & Kumar, J. (2024). The role of AI chatbots in transforming guest engagement and marketing in hospitality. In *Integrating AI-Driven Technologies Into Service Marketing* (pp. 595-620). IGI Global. https://doi.org/10.4018/979-8-3693-7122-0.ch029
9. Sigala, M., Ooi, K. B., Tan, G. W. H., Aw, E. C. X., Buhalis, D., Cham, T. H., ... & Ye, I. H. (2024). Understanding the impact of ChatGPT on tourism and hospitality: Trends, prospects and research agenda. *Journal of Hospitality and Tourism Management*, *60*, 384-390.
10. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A. & Fung, P. (2024). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1-38.
11. Neha, F., Bhati, D., & Shukla, D. K. (2025). Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review. *AI*, *6*(9), 226. https://doi.org/10.3390/ai6090226
12. Chen, A.; Ge, X.; Fu, Z.; Xiao, Y.; Chen, J. TravelAgent: An AI Assistant for Personalized Travel Planning. *arXiv* 2024, *preprint arXiv:2409.08069*. Available online: https://arxiv.org/abs/2409.08069 (accessed on 24 October 2025).
13. Heskett, J. L., T. O. Jones, G. W. Loveman, W. Earl Sasser, and L. A. Schlesinger. Putting the Service-Profit Chain to Work. *Harvard Business Review*, *72*(2), (March–April 1994): 164–174.

 doi:10.20944/preprints202512.1693.v1

14. Rachel W.Y. Yee, Andy C.L. Yeung, T.C. Edwin Cheng (2008). The impact of employee satisfaction on quality and profitability in high-contact service industries. *Journal of Operations Management*, *26*(5). 651-668. https://doi.org/10.1016/j.jom.2008.01.001.
15. Benckendorff, P. J., Sheldon, P. J. & Fesenmaier, D. R. (2019). *Tourism Information Technology* (3rd ed.). CABI.
16. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A. & Gurevych, I. (2021). BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. *1*.
17. Yoran, O., Wolfson, T. & Berant, J. (2023). Making Retrieval-Augmented Language Models Robust to Irrelevant Context. *International Conference on Learning Representations (ICLR)*.
18. Ethayarajh, K. (2019). How Contextual are Contextualized Word Representations? *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 2409–2419.
19. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.
20. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
21. Google DeepMind. (2024). *Gemini 2.0 Flash: Technical Report*. Available online: https://deepmind.google/technologies/gemini/ (accessed on 30 April 2025).
22. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3982-3992.
23. Kshirsagar, Aadit. (2024). Enhancing RAG Performance Through Chunking and Text Splitting Techniques. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 10*. 151-158. 10.32628/CSEIT2410593.
24. Günther, M., Ong, J., Mohr, I., Abdessalem, A., Abel, T., Akram, M. K., ... & Xiao, H. (2023). Jina embeddings 2: 8192-token general-purpose text embeddings for long documents. arXiv preprint arXiv:2310.19923.
25. ChromaDB. (2024). *ChromaDB: The AI-native open-source embedding database*. Available online: https://www.trychroma.com/ (accessed on 30 April 2025).
26. Malkov, Y. A., & Yashunin, D. A. (2018). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(4), 824-836.
27. MistralAI. (2024). *Mistral OCR: Document understanding at scale*. Available online: https://mistral.ai/news/mistral-ocr (accessed on 30 March 2025).
28. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*.
29. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824-24837.
30. Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493-498.
31. Amazon Web Services. (2024a). *Amazon Elastic Container Service: Developer Guide*. Available online: https://docs.aws.amazon.com/ecs/ (accessed on 30 April 2025).
32. Merkel, D. (2014). Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal*, *2014*(239), 2.
33. Amazon Web Services. (2024b). *Amazon S3: User Guide*. Available online: https://docs.aws.amazon.com/s3/ (accessed on 30 April 2025).
34. Amazon Web Services. (2024c). *AWS CodePipeline: User Guide*. Available online: https://docs.aws.amazon.com/codepipeline/ (accessed on 30 April 2025).
35. Knani, M., Echchakoui, S., Ladhari, R. (2022) Artificial intelligence in tourism and hospitality: Bibliometric analysis and research agenda. *International Journal of Hospitality Management*, Volume 107 https://doi.org/10.1016/j.ijhm.2022.103317.
36. Nyholm, Sven. (2023). Artificial Intelligence and Human Enhancement: Can AI Technologies Make Us More (Artificially) Intelligent?. *Cambridge quarterly of healthcare ethics : CQ : the international journal of healthcare ethics committees*. *33*. 1-13. 10.1017/S0963180123000464.

 doi:10.20944/preprints202512.1693.v1

37. Kandampully, J., Bilgihan, A., & Zhang, T. (2015). Customer Loyalty: A Review and Future Directions with a Special Focus on the Hospitality Industry. *International Journal of Contemporary Hospitality Management*, *27*, 379-414. https://doi.org/10.1108/IJCHM-03-2014-0151
38. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *Proceedings of ICLR 2024*.
39. Prasanna, A., P. P., Prasad, B. (2025).Conversational AI in Tourism: A systematic literature review using TCM and ADO framework. *Journal of Hospitality and Tourism Management*, *64*, 101310. https://doi.org/10.1016/j.jhtm.2025.101310.
40. Tuomi, A., Tussyadiah, I., & Hanna, P. (2025). Customized travel planning with generative AI: Opportunities and challenges. *Annals of Tourism Research*, *106*, 103756.
41. Zhang, M., & Li, J. (2024). Large language models for tourism sentiment analysis: A comparative study. *Tourism Management Perspectives*, *51*, 101234.
42. Khattab, O., Santhanam, K., Li, X. L., Hall, D., Liang, P., Potts, C., & Zaharia, M. (2024). Augmented language models: A survey. *Foundations and Trends in Machine Learning*, *17*(2), 123-231.
43. Raisch, S., & Krakowski, S. (2024). Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review*, *49*(1), 192-210.
44. Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford University Press.
45. Alavi, M., & Leidner, D. E. (2001). Knowledge management and knowledge management systems: Conceptual foundations and research issues revisited. *MIS Quarterly*, *1*(10), 107-136.
46. Alfredo, R., Echeverria, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D. & Martinez-Maldonado, R. (2024). human-centred learning analytics and AI in education: A systematic literature review. *Computers and Education: Artificial Intelligence*, *6*, 100215. https://doi.org/10.1016/j.caeai.2024.100215
47. Knowles, M. S., Holton, E. F., & Swanson, R. A. (2024). The adult learner: The definitive classic in adult education and human resource development (9th ed.). Routledge.
48. Gretzel, U., Sigala, M., & Xiang, Z. (2024). Smart tourism: Foundations and developments from a multidisciplinary perspective. *Electronic Markets*, *34*(1), 15-34.
49. Kravariti F, Jooss S, Tom Dieck MC, Fountoulaki P, Hossain F. (2025). Talent management in the hospitality and tourism industry: the role of societal and organisational culture. *International Journal of Contemporary Hospitality Management*, *37*(1), 260–278.
50. Neuhofer, B., Buhalis, D., & Ladkin, A. (2024). A Typology of Technology-Enhanced Tourism Experiences. *International Journal of Tourism Research*, *16*(4).
51. Sheldon, P. J., & Fesenmaier, D. R. (2024). Tourism education in the age of AI: Challenges and opportunities. *Journal of Teaching in Travel & Tourism*, *24*(1), 45-67.
52. Grünbichler, R., Salimbeni, S. (2024). Artificial Intelligence in Small and Medium-Sized Enterprises: Requirements and Barriers. In: Concli, F., Maccioni, L., Vidoni, R., Matt, D.T. (eds) Latest Advancements in Mechanical Engineering. ISIEA 2024. Lecture Notes in Networks and Systems, vol 1125. Springer, Cham.
53. Jose Antonio Clemente-Almendros, Dorina Nicoara-Popescu, Ivan Pastor-Sanz. (2024). Digital transformation in SMEs: Understanding its determinants and size heterogeneity. *Technology in Society*, *77*, 102483. https://doi.org/10.1016/j.techsoc.2024.102483.
54. Mossavar-Rahmani, F. & Zohuri, B. (2024). Artificial Intelligence at Work: Transforming Industries and Redefining the Workforce Landscape. *Journal of Economics & Management Research*. *5*, 1-4.
55. Venkatesh, V., Thong, J. Y., & Xu, X. (2024). Unified theory of acceptance and use of technology: A synthesis and the road ahead. *Journal of the Association for Information Systems*, *17*(5), 328-376.
56. Collings, D. G., Mellahi, K., & Cascio, W. F. (2018). Global talent management and performance in multinational enterprises: A multilevel perspective. *Journal of Management*, *45*(2).
57. Shih, Y. K., & Kang, Y. K. (2025). Design and Implementation of a Secure RAG-Enhanced AI Chatbot for Smart Tourism Customer Service: Defending Against Prompt Injection Attacks–A Case Study of Hsinchu, Taiwan. arXiv preprint arXiv:2509.21367. (accesed on 27 october 2025)
58. Li, J., Li, D., Savarese, S., & Hoi, S. (2024). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proceedings of ICML 2024*, 19730-19742.
59. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2024). ReAct: Synergizing reasoning and acting in language models. *Proceedings of ICLR 2024*.
60. Acemoglu, D., & Restrepo, P. (2024). Artificial intelligence, automation, and work. In *The economics of artificial intelligence: An agenda* (pp. 197-236). University of Chicago Press.

 doi:10.20944/preprints202512.1693.v1