UNIVERSIDAD EUROPEA VALENCIA

Máster de Formación Permanente en Inteligencia Artificial

TRABAJO DE FIN DE MÁSTER

Diseño de un sistema inteligente de agente conversacional para consultas académicas en entornos universitarios

DESIGN OF AN INTELLIGENT CONVERSATIONAL AGENT SYSTEM FOR ACADEMIC QUERIES IN UNIVERSITY ENVIRONMENTS

Kevin Bryan Andrés Abalos Ramirez

HECTOR ESPINÓS MORATÓ

Curso académico 2024-2025

Resumen

Este trabajo presenta el diseño, implementación y evaluación de un agente conversacional orientado a tareas para el acceso a información académica y administrativa en educación superior. Partimos del problema de la fragmentación informativa en múltiples plataformas institucionales y de la sobrecarga de los canales tradicionales de atención. Metodológicamente, se adoptó CRISP-DM y se propuso una arquitectura modular con tres pilares: (i) ingesta y normalización documental (segmentación en chunks, metadatos y embeddings), (ii) recuperación semántica mediante Retrieval-Augmented Generation (RAG) sobre `ChromaDB` y `OpenAIEmbeddings`, y (iii) generación con un LLM (`ChatOpenAI gpt-4o-mini`) orquestado con `LangChain` (patrón ReAct) y expuesto vía interfaz web en `Chainlit`. El sistema prioriza trazabilidad y transparencia al citar explícitamente la fuente de cada respuesta. La validación se realizó con un benchmark de preguntas frecuentes institucionales en un escenario de chat testing. Los resultados muestran una precisión del 82,4 %, cobertura del 91,7 %, First Turn Resolution del 76,3 % y latencia mediana de 6,2 s, evidenciando viabilidad técnica y experiencia de uso fluida para un despliegue inicial. La propuesta se alinea con la transformación digital universitaria y promueve accesibilidad y equidad en el acceso a la información.

Palabras claves: agente conversacional, recuperación aumentada por generación (RAG), procesamiento de lenguaje natural (PLN), educación superior.

Abstract

This paper presents the design, implementation, and evaluation of a task-oriented conversational agent for accessing academic and administrative information in higher education. We start from the problem of information fragmentation across multiple institutional platforms and the overload of traditional service channels. Methodologically, CRISP-DM was adopted and a modular architecture with three pillars was proposed: (i) document ingestion and normalisation (segmentation into chunks, metadata and embeddings), (ii) semantic retrieval using Retrieval-Augmented Generation (RAG) on ChromaDB and OpenAIEmbeddings, and (iii) generation with an LLM (`ChatOpenAI gpt-4o-mini`) orchestrated with `LangChain` (ReAct pattern) and exposed via a web interface in `Chainlit`. The system prioritises traceability and transparency by explicitly citing the source of each response. Validation was performed using a benchmark of institutional frequently asked questions in a chat testing scenario. The results show an accuracy

of 82.4%, coverage of 91.7%, First Turn Resolution of 76.3% and median latency of 6.2 s, demonstrating technical feasibility and a smooth user experience for initial deployment. The proposal is aligned with the university's digital transformation and promotes accessibility and equity in access to information.

Keywords: conversational agent, retrieval-augmented generation (RAG), natural language processing (NLP), higher education.

Índice

1. Introducción **1**

1.1. Contexto y motivación 1

1.2. Problema y finalidad 2

1.3. Objetivos del TFM 4

1.4. Estructura del documento 4

2. Marco teórico **6**

2.1. Sistemas conversacionales en educación 6

2.1.1. Chatbots 8

2.1.2. Asistentes virtuales 12

2.1.3. Agentes de IA 14

2.2. Procesamiento de lenguaje natural 17

2.2.1. Evolución del PLN: 17

2.2.2. Transformers 18

2.2.3. Grandes modelos de lenguaje 20

2.2.4. Modelos multimodales 22

2.3. Técnicas de adaptación de modelos 24

2.3.1. Fine-tuning 24

2.3.2. Generación aumentada con recuperación (RAG) 25

2.3.3. Discusión técnica 26

2.4. Chatbots en educación superior 27

2.4.1. Modelado matemático de chatbots universitarios 28

2.4.2. Integración con modelos de lenguaje 28

2.4.3. Aplicaciones en universidades 29

3. Metodología **31**

3.1. Diseño metodológico 31

3.2. Participantes 31

3.3. Instrumentos 31

3.4. Procedimiento 32

3.5. Análisis de datos 33

4. Resultados **34**

4.1. Ingesta y preprocesamiento de la información 34

4.2. Construcción del agente conversacional 35

4.3. Evaluación del sistema 38

4.4. Análisis cualitativo de las respuestas 39

5. Discusión **42**

6. Conclusiones **46**

7. Limitaciones y futuras líneas de investigación **48**

Referencias **52**

1. Introducción

1.1. Contexto y motivación

Las universidades han atravesado una acelerada transición hacia lo digital impulsada por los eventos mundiales ocurridos en los últimos años, lo que no sólo ha generado la búsqueda de mayor eficiencia en la gestión administrativa, sino también por el cambio en las demandas de estudiantes y docentes, quienes esperan un acceso más ágil y flexible a la información y a los servicios educativos (Mohamed Hashim, Tlemsani y Matthews, 2022). En este contexto, los entornos virtuales de aprendizaje, las plataformas administrativas, los repositorios académicos y los sistemas de gestión de contenidos constituyen elementos esenciales en la infraestructura digital universitaria (Mohamed Hashim, Tlemsani y Matthews, 2022). Sin embargo, pese a los avances en digitalización, el acceso a la información sigue presentando desafíos significativos para estudiantes, docentes y personal administrativo, especialmente en lo que respecta a la localización rápida, coherente y precisa de información dispersa en múltiples sistemas y plataformas.

La fragmentación de los recursos digitales y la ausencia de interfaces unificadas generan una sobrecarga cognitiva en los usuarios, quienes deben navegar entre sitios web, portales institucionales, bases de datos y documentos para satisfacer necesidades informativas puntuales (A. L. Rodrigues y Pereira, 2024). Esta situación se agrava para los nuevos miembros de la comunidad universitaria —tanto estudiantes de recién ingresados como profesorado recién incorporado—, quienes enfrentan una curva de aprendizaje elevada para familiarizarse con el entorno institucional altamente cambiante.

Paralelamente, los avances en inteligencia artificial (IA), especialmente en el campo del procesamiento del lenguaje natural (PLN), han abierto nuevas oportunidades para el diseño de interfaces inteligentes que permitan una interacción más natural y eficiente con los sistemas de información. Los chatbots inteligentes o agentes conversacionales, diseñados para comprender y generar respuestas en lenguaje natural, se consolidan actualmente como herramientas innovadoras que favorecen un acceso más ágil y contextualizado a la información en distintos escenarios, incluyendo la educación superior (Dawood, 2024).

Drake (2011) subraya que el éxito de los estudiantes en la universidad depende en gran medida de la calidad de la relación que mantienen con sus asesores académicos. Estos desempeñan un papel clave al orientar sobre normativas, procedimientos y recursos institucionales, aunque la elevada demanda de consultas puede sobrecargar su capacidad de atención. En este contexto, los

agentes conversacionales se presentan como una alternativa eficaz para aliviar dicha carga, pues permiten a los estudiantes acceder de manera ágil a información relevante. Diversos estudios han demostrado que su implementación en entornos universitarios contribuye a optimizar los recursos disponibles, reducir los tiempos de búsqueda y mejorar tanto la experiencia del usuario como los procesos de toma de decisiones académicas (Akiba y Fraboni, 2023).

En este marco, el presente TFM se enmarca en la necesidad de diseñar un sistema de agente conversacional que actúe como intermediario inteligente entre el usuario y la infraestructura informativa de la universidad. Este sistema buscará no solo mejorar el acceso a la información, sino también fortalecer la percepción de modernidad, eficiencia y accesibilidad institucional, contribuyendo así a una experiencia universitaria más inclusiva, personalizada y satisfactoria.

1.2. Problema y finalidad

Uno de los principales retos que enfrentan las universidades en la actualidad no radica únicamente en la generación de conocimiento, sino también en garantizar que la comunidad académica pueda acceder a la información institucional de manera rápida y confiable. Procesos como la inscripción en asignaturas, la consulta de calendarios académicos, la tramitación de becas o la solicitud de certificados suelen implicar una búsqueda extensa en diferentes plataformas, lo cual provoca demoras, errores y, en muchos casos, frustración en los estudiantes (Preetha et al., 2023). La dispersión de la información administrativa entre distintos sistemas y departamentos genera ineficiencias y aumenta la carga de trabajo para el personal universitario encargado de dar soporte.

Tradicionalmente, estas necesidades han sido atendidas mediante canales de atención presenciales, correos electrónicos o consultas directas a las secretarías académicas. Sin embargo, estos mecanismos no siempre logran absorber la elevada demanda, especialmente en períodos críticos como el inicio de semestre, lo que produce tiempos de espera prolongados y una experiencia poco satisfactoria para el estudiante (Majorana et al., 2022). Esta situación se traduce en una pérdida de eficiencia institucional y afecta negativamente la percepción de la calidad del servicio universitario.

La magnitud del problema se observa con claridad en experiencias recientes de educación superior. En Deakin University, el asistente “Genie” registró más de 85 000 conversaciones con 10 677 estudiantes activos poco después de su despliegue, ilustrando el volumen de dudas operativas que, de otro modo, saturan canales tradicionales de secretaría y soporte (Deakin

University, 2019). En el ámbito de admisiones, Georgia State University documentó que su agente "Pounce" redujo el *summer melt* en un 21,4 %, y que solo el 1 % de los mensajes requirió intervención humana, lo que sugiere una capacidad real de absorción de demanda y de mejora en tiempos de respuesta institucionales (Mainstay, 2020). Estos resultados confirman que un agente conversacional bien orquestado puede aliviar cuellos de botella administrativos de alto impacto estacional (p. ej., matrícula y becas) y sostener latencias compatibles con interacción natural, reforzando la pertinencia de una solución RAG en universidades con repositorios normativos dinámicos.

En este contexto, los avances en sistemas conversacionales basados en inteligencia artificial ofrecen una alternativa viable y escalable. Los chatbots permiten responder de forma inmediata a consultas frecuentes, automatizar procesos administrativos básicos y guiar a los estudiantes a través de trámites recurrentes sin necesidad de intervención humana constante (Klopfenstein et al., 2017). Su capacidad para funcionar de manera ininterrumpida y atender múltiples solicitudes simultáneamente contribuye a reducir la sobrecarga en el personal administrativo, al mismo tiempo que mejora la experiencia del usuario final.

La justificación de este TFM, por tanto, se fundamenta en la necesidad de modernizar los canales de comunicación entre la universidad y su comunidad mediante la incorporación de un agente conversacional inteligente. Este sistema no solo busca optimizar los tiempos de respuesta y la eficiencia administrativa, sino también reforzar la accesibilidad, la equidad y la satisfacción de los estudiantes en su interacción con la institución. Asimismo, su implementación se alinea con las tendencias globales de digitalización y automatización de servicios en educación superior, posicionando a la universidad como una institución innovadora y adaptada a las demandas de la sociedad digital actual (Al-Ghonmein, Al-Moghrabi y Alrawashdeh, 2023).

Además, la implementación de agentes conversacionales en el ámbito universitario se encuentra alineada con los Objetivos de Desarrollo Sostenible (ODS) propuestos por la Agenda 2030 de Naciones Unidas. En particular, contribuye al ODS 4: Educación de calidad, al facilitar un acceso inclusivo y equitativo a la información académica y administrativa, reduciendo las barreras de comunicación que enfrentan los estudiantes. Asimismo, guarda relación con el ODS 9: Industria, innovación e infraestructura, al promover la digitalización de servicios y la innovación en los procesos internos de las instituciones de educación superior. Finalmente, se vincula con el ODS 10: Reducción de las desigualdades, ya que el uso de chatbots permite atender de manera homogénea a todos los estudiantes, independientemente de sus condiciones socioeconómicas o

de sus competencias digitales, contribuyendo a mejorar la equidad en el acceso a los recursos universitarios (`https://www.un.org/sustainabledevelopment/`).

1.3. Objetivos del TFM

El propósito central de este TFM es el diseño y desarrollo de un sistema de agente conversacional inteligente que facilite el acceso a información académica y administrativa en un entorno universitario. Dicho sistema busca optimizar los procesos de búsqueda de información, mejorar la experiencia del usuario y apoyar la transformación digital de la institución educativa.

Objetivo general: Diseñar e implementar un agente conversacional basado en técnicas de PLN y recuperación de información, orientado a resolver consultas académicas y administrativas de los estudiantes en tiempo real, mejorando la accesibilidad, la eficiencia y la satisfacción en la interacción con la universidad.

Objetivos específicos:

- Diseñar una arquitectura modular y escalable para el agente conversacional universitario, asegurando su integración con sistemas institucionales y la posibilidad de futuras extensiones.
- Seleccionar e integrar un modelo de IA generativa de última generación (GPT-4o-mini) como núcleo del agente conversacional, garantizando su adaptación al dominio universitario.
- Implementar una interfaz de usuario basada en entorno web que permita consultas académicas y administrativas mediante interacción en lenguaje natural, priorizando la trazabilidad y la transparencia de las respuestas.
- Evaluar experimentalmente la efectividad del sistema a través de un escenario de *chat testing* con recuperación aumentada (RAG), utilizando un conjunto de preguntas frecuentes institucionales como *benchmark*. La evaluación se realizará con métricas objetivas como precisión de respuestas correctas, cobertura de consultas y latencia de interacción.

1.4. Estructura del documento

El presente TFM se organiza en seis capítulos principales. En el Capítulo 1 se expone la introducción, incluyendo el contexto, la motivación, el problema de investigación, la justificación y los objetivos del estudio. El Capítulo 2 desarrolla el marco teórico, revisando los fundamentos

de la IA, el PLN y el uso de agentes conversacionales en el ámbito universitario. El Capítulo 3 describe la metodología utilizada para el diseño e implementación del sistema, especificando la arquitectura propuesta y las tecnologías empleadas. El Capítulo 4 presenta el desarrollo e implementación del agente conversacional, detallando los módulos, la integración de datos y la interfaz de usuario. En el Capítulo 5 se analizan los resultados obtenidos, tanto en pruebas técnicas como en la evaluación con usuarios reales. Finalmente, el Capítulo 6 recoge las conclusiones, las limitaciones del trabajo y las líneas de investigación futura.

2. Marco teórico

2.1. Sistemas conversacionales en educación

El desarrollo de los sistemas conversacionales en el ámbito universitario responde a la necesidad de integrar soluciones basadas en inteligencia artificial (IA) capaces de gestionar interacciones complejas con los estudiantes y el personal académico. A diferencia de los portales web tradicionales o las interfaces gráficas de usuario, estos sistemas buscan ofrecer una experiencia de comunicación más natural, reduciendo la brecha entre el lenguaje humano y los entornos digitales (Følstad, Araujo, Law et al., 2021). Esta aproximación resulta especialmente relevante en la educación superior, donde la información institucional se encuentra dispersa y donde las demandas de los usuarios requieren respuestas rápidas, fiables y personalizadas.

De manera general, los sistemas conversacionales se dividen en dos modalidades principales. En primer lugar, los sistemas de diálogo abierto (*open-domain*) se caracterizan por su capacidad de sostener conversaciones de carácter general, sin estar restringidos a un objetivo concreto, lo que favorece interacciones más naturales y un mayor enriquecimiento en la dimensión social, acompañar al estudiante en su vida académica o incluso proporcionar apoyo motivacional y emocional (Gao, Galley y L. Li, 2019). Por otro lado, los sistemas de diálogo orientado a tareas (*task-oriented*) se centran en resolver problemas concretos con un alto grado de exactitud, como consultar plazos de matrícula, verificar requisitos de becas o como poder acceder a certificados oficiales (Følstad, Araujo, Papadopoulos et al., 2020). En el contexto universitario, esta segunda modalidad es la más relevante, ya que la mayoría de las interacciones están vinculadas a procesos administrativos y académicos donde la precisión, la trazabilidad y la integración con sistemas institucionales resultan críticas.

La arquitectura típica de un sistema conversacional académico integra tres componentes principales: (i) comprensión del lenguaje natural (NLU) que identifica la intención del usuario y extrae entidades relevantes (como *slots* o *keywords*); (ii) el módulo de gestión del diálogo, cuya función es mantener actualizado el estado de la conversación (Dialog State Tracking, DST) y decidir la acción a ejecutar en cada turno. Para ello puede apoyarse en políticas de decisión basadas en reglas predefinidas, modelos probabilísticos (MDP/POMDP) o enfoques de aprendizaje por refuerzo, lo que permite equilibrar la coherencia del intercambio con la eficiencia en la resolución de la tarea; y (iii) generación de lenguaje natural, que construye la respuesta al usuario, ya sea mediante selección de frases predefinidas (*retrieval-based*) o

generación automática (*generative-based*) con modelos neuronales (Gao, Galley y L. Li, 2019; McTear, 2022).

Desde un punto de vista técnico, los sistemas conversacionales han evolucionado desde modelos basados en reglas y patrones hasta arquitecturas que integran aprendizaje automático profundo y representaciones semánticas avanzadas, lo que ha permitido superar limitaciones históricas como la escasa capacidad para mantener el contexto conversacional, la dificultad para escalar a dominios amplios y la rigidez de las respuestas (Yu et al., 2011). El mayor avance en esta transición ha sido la introducción de arquitecturas basadas en *Transformers*, que gracias a mecanismos de autoatención permiten modelar dependencias de largo alcance y dotar a los sistemas de una mayor flexibilidad y coherencia en la interacción (Bird, Ekárt y Faria, 2023).

En el contexto educativo, esta transición tecnológica no implica únicamente mejoras en la calidad de las interacciones, sino también en la capacidad de integración con ecosistemas institucionales. En el ámbito de la educación superior, un agente conversacional debe ir más allá de la simple detección de intenciones y la producción de respuestas en lenguaje natural. Su diseño debe contemplar la integración con repositorios académicos, entornos virtuales de aprendizaje (LMS) y plataformas administrativas, de manera que las interacciones resulten no solo relevantes desde el punto de vista semántico, sino también precisas, verificables y con trazabilidad institucional garantizada (Følstad, Araujo, Law et al., 2021; C. Rodrigues et al., 2022).

La evaluación de su rendimiento constituye un aspecto esencial. Más allá de la exactitud en la detección de intenciones, es necesario medir la cobertura de consultas, la confianza percibida en las respuestas y la claridad de la interacción. Estas métricas, complementadas con evaluaciones de usabilidad y accesibilidad, permiten determinar si el sistema realmente aporta valor pedagógico e institucional (Al-Jaf et al., 2024; Wollny et al., 2021).

Finalmente, los sistemas conversacionales en educación deben enfrentar desafíos específicos relacionados con la ética y la privacidad. El tratamiento de información académica de carácter sensible —como historiales de matrícula, expedientes estudiantiles o trámites de becas— requiere que la arquitectura del sistema integre mecanismos robustos de autenticación, gestión de permisos y adherencia a marcos legales de protección de datos (Hasal et al., 2021; Tran et al., 2025). Asimismo, garantizar la trazabilidad de las interacciones y la posibilidad de auditar las respuestas generadas constituye un requisito fundamental para preservar la confianza tanto de los estudiantes como del personal universitario.

En suma, los sistemas conversacionales en educación superior no deben concebirse únicamente como herramientas de interacción lingüística, sino como infraestructuras inteligentes que combinan procesamiento de lenguaje natural, gestión del diálogo, recuperación de información y seguridad institucional. Esta visión integral asegura que su implementación contribuya a la eficiencia administrativa y a la calidad educativa, alineándose con los procesos de transformación digital que atraviesan las universidades en la actualidad.

2.1.1. Chatbots

Los chatbots constituyen una de las aplicaciones más consolidadas de los sistemas conversacionales y representan un paso intermedio entre la automatización tradicional de servicios y la interacción inteligente mediada por agentes de IA. Un chatbot puede definirse como un sistema computacional orientado a reproducir interacciones conversacionales con personas, ya sea mediante texto o voz, con la finalidad de proporcionar respuestas, apoyar en la realización de tareas específicas o acompañar al usuario en procesos previamente estructurados (Shawar y Atwell, 2007).

Históricamente, los primeros chatbots como *ELIZA* (1966) o *PARRY* (1972) se basaban en reglas simbólicas y técnicas de emparejamiento de patrones. Aunque limitados en comprensión semántica, demostraron la viabilidad de establecer un intercambio de tipo conversacional entre humano y máquina (Collby, 1975; Weizenbaum, 1966). Posteriormente, con el auge de los enfoques estadísticos y del aprendizaje automático en la década de 1990, se incorporaron técnicas de clasificación y modelos probabilísticos que permitieron avances en el reconocimiento de intenciones y la recuperación de información. Actualmente, la introducción del aprendizaje profundo y de arquitecturas de tipo *transformer* ha revolucionado el campo, otorgando a los chatbots capacidades de comprensión contextual y generación de lenguaje de alta naturalidad (Gao, Galley y L. Li, 2019; Vinyals y Le, 2015).

Como se detalló en la sección sobre sistemas conversacionales, un chatbot moderno integra módulos de NLU, gestor de diálogo y generación de lenguaje natural, los cuales permiten identificar intenciones, gestionar el flujo conversacional y construir respuestas adaptadas al contexto. Esta integración técnica ha permitido ampliar su funcionalidad y precisión en dominios complejos como la educación superior.

A nivel de clasificación funcional, los chatbots pueden dividirse en tres categorías principales. Dentro de la clasificación funcional, los chatbots pueden dividirse en tres enfoques principales.

Los sistemas *retrieval-based* recuperan la respuesta más adecuada de un conjunto previamente definido, lo que garantiza un mayor control sobre la exactitud de la información. En contraste, los *generative-based* utilizan modelos de lenguaje para producir respuestas nuevas, adaptadas al contexto de la conversación. Finalmente, los modelos híbridos combinan ambas aproximaciones, buscando un equilibrio entre precisión y flexibilidad (Serban et al., 2015). En el caso de entornos universitarios, los enfoques híbridos se perfilan como los más adecuados, pues garantizan que las respuestas relacionadas con normativas, plazos o requisitos provengan de fuentes verificadas, mientras que el componente generativo se reserva para interacciones menos críticas, como orientación general o acompañamiento en procesos de aprendizaje.

El despliegue de chatbots en instituciones de educación superior plantea, no obstante, desafíos técnicos y éticos significativos. Uno de los desafíos más críticos radica en la posibilidad de que el sistema produzca respuestas erróneas o no fundamentadas —conocidas como "alucinaciones"—, lo que puede comprometer la fiabilidad percibida por el usuario y disminuir su confianza en la herramienta. Para mitigar este problema, investigaciones recientes recomiendan el uso de técnicas de Retrieval-Augmented Generation (RAG), donde el modelo de lenguaje se apoya en documentos oficiales recuperados en tiempo real, garantizando tanto precisión como actualizaciones dinámicas (P. Lewis et al., 2020).

Recientes estudios han concentrado su atención en el uso de RAG para consultas administrativas en entornos universitarios. Por ejemplo, Z. Chen et al. (2024) presentan un chatbot basado en RAG para simplificar los procesos de admisión universitaria, usando documentos institucionales para mejorar precisión y trazabilidad en respuestas sobre requisitos académicos y plazos . Asimismo, Neupane et al. (2024) desarrollaron BARKPLUG V.2, un sistema conversacional universitario orientado a responder consultas sobre recursos del campus, con alta calificación de usabilidad (SUS) y puntuación RAGAS de 0,96. Finalmente, Nguyen y Quan (2024) describen URAG, una arquitectura híbrida RAG optimizada para agentes de admisión, que logró resultados comparables a modelos comerciales, validada en un estudio de caso real.

Otro aspecto crucial es la integración segura con los sistemas académicos existentes (LMS, ERP, SIGA), que requiere autenticación robusta, control de accesos y mecanismos de auditoría para asegurar la trazabilidad de cada interacción. Asimismo, la dimensión ética cobra relevancia en el manejo de datos sensibles de los estudiantes, lo que obliga a cumplir estrictamente normativas como el Reglamento General de Protección de Datos (RGPD) en Europa (Hasal et al., 2021).

Un aspecto fundamental para garantizar la eficacia de los chatbots en el ámbito universitario es la capacidad de escalabilidad y mantenimiento del sistema. A diferencia de los entornos comerciales, donde los dominios de aplicación suelen estar más acotados, en la universidad las normativas, calendarios y procesos administrativos cambian con frecuencia. Esto exige que el chatbot no sea una herramienta estática, sino un sistema vivo capaz de actualizar sus repositorios de conocimiento y entrenar continuamente sus modelos de NLU con datos institucionales actualizados. En este sentido, los enfoques de *continual learning* y el uso de pipelines de aprendizaje automático automatizados (MLOps) se perfilan como estrategias clave para mantener la vigencia y la precisión de las respuestas en el tiempo (Al-Jaf et al., 2024).

La evaluación del rendimiento de un chatbot académico no puede limitarse a métricas técnicas tradicionales como la exactitud o la cobertura. En contextos educativos, resulta imprescindible incorporar indicadores de impacto en la experiencia del usuario, como el nivel de satisfacción percibida, la tasa de resolución en el primer turno (*First Turn Resolution, FTR*), el tiempo medio de respuesta y la accesibilidad para estudiantes con distintas competencias digitales. La Tabla 1 sintetiza algunos de los criterios de evaluación más relevantes en el ámbito universitario, diferenciando entre métricas técnicas y métricas orientadas a la experiencia del usuario.

Tabla 1: Métricas de evaluación de chatbots en educación superior.

Categoría	**Métrica y descripción**
Técnicas	*Precisión:* proporción de respuestas correctas respecto al total. *Cobertura:* porcentaje de consultas que el sistema puede gestionar. *Tiempo de respuesta:* latencia media entre la consulta y la respuesta. *Tasa de error de NLU:* frecuencia con que las intenciones o entidades se clasifican incorrectamente.
Experiencia del usuario	*Satisfacción percibida:* evaluación subjetiva del usuario respecto a la utilidad del chatbot. *FTR:* proporción de consultas resueltas en el primer turno de interacción. *Engagement:* grado de continuidad en el uso del sistema en distintos escenarios académicos. *Accesibilidad:* facilidad de uso para estudiantes con diferentes niveles de alfabetización digital.

Fuente: elaboración propia.

Finalmente, la correcta integración del chatbot con los sistemas universitarios constituye un factor determinante para su éxito. Más allá de responder consultas de carácter general, un chatbot académico debe poder conectarse de manera segura con sistemas de gestión del aprendizaje (LMS), plataformas de gestión administrativa (ERP académico), bases de datos de bibliotecas digitales o sistemas de información de gestión académica (SIGA). El uso de interfaces de conexión estandarizadas (como APIs REST o GraphQL) junto con esquemas de autenticación federada (OAuth2, SSO) no solo asegura la trazabilidad de las interacciones, sino que también habilita al agente conversacional para desempeñar funciones proactivas, tales como generar recordatorios individualizados o comprobar automáticamente requisitos en procesos administrativos (Tran et al., 2025).

La Tabla 2 sintetiza los principales módulos funcionales de un agente conversacional universitario moderno, junto con sus tecnologías asociadas y las funciones que desempeñan dentro del sistema.

Tabla 2: Componentes funcionales de un agente conversacional universitario

Módulo	**Tecnologías asociadas**	**Función principal**
Comprensión del lenguaje natural (NLU)	Transformers, embeddings contextuales (BERT, RoBERTa)	Detectar intenciones del usuario y extraer entidades clave
Gestión del diálogo	POMDP, reglas, aprendizaje por refuerzo	Mantener el estado de la conversación y seleccionar acciones
Generación de lenguaje natural (NLG)	Plantillas, modelos generativos (GPT, T5)	Construir respuestas en lenguaje natural adaptadas al contexto
Módulo de recuperación (en RAG)	Búsqueda semántica, vectores, FAISS	Localizar documentos relevantes en tiempo real
Motor de personalización	Perfiles de usuario, aprendizaje supervisado	Adaptar respuestas y sugerencias al historial del estudiante
Gestión del contexto persistente	Bases de datos, almacenamiento de sesiones	Recordar interacciones anteriores del usuario
Integración con sistemas institucionales	APIs REST, GraphQL, SSO, OAuth2	Consultar y actualizar datos en LMS, ERP, SIGA

Fuente: elaboración propia.

2.1.2. Asistentes virtuales

Los asistentes virtuales representan una evolución conceptual y técnica de los chatbots hacia sistemas conversacionales más avanzados, caracterizados por su capacidad de aprendizaje adaptativo, razonamiento contextual y soporte multimodal. Mientras que los chatbots tradicionales se diseñan principalmente para ejecutar interacciones acotadas y responder a consultas predefinidas, los asistentes virtuales integran técnicas de procesamiento de lenguaje natural (PLN), aprendizaje automático profundo y razonamiento simbólico, con el objetivo de ofrecer un soporte continuo, personalizado y escalable (Laranjo et al., 2018). En el ámbito universitario, esta diferencia resulta esencial, dado que los estudiantes y el personal administrativo requieren no solo respuestas inmediatas a consultas puntuales, sino también acompañamiento proactivo en procesos académicos que se extienden a lo largo del ciclo formativo.

Desde una perspectiva formal, los asistentes virtuales pueden modelarse como agentes inteligentes que resuelven problemas de decisión secuencial en entornos dinámicos. Una aproximación ampliamente empleada consiste en representarlos como Procesos de decisión de Markov parcialmente observables (POMDP), donde el sistema debe seleccionar acciones óptimas a partir de observaciones incompletas del estado real de la interacción (Kanwal y Farooq, 2021). En este marco, la política óptima π^* se define como aquella que maximiza la recompensa acumulada esperada:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right],$$

donde s_t denota el estado latente de la interacción en el instante t, a_t la acción elegida por el asistente, $R(s_t, a_t)$ la función de recompensa asociada y γ el factor de descuento. Este enfoque probabilístico resulta adecuado para modelar tareas conversacionales en contextos universitarios, donde el sistema debe razonar bajo incertidumbre sobre intenciones ambiguas o información incompleta.

A nivel arquitectónico, los asistentes virtuales extienden la estructura típica de los chatbots (NLU, gestor de diálogo y NLG) mediante módulos adicionales que refuerzan su capacidad de adaptación y razonamiento. Entre ellos destacan: (i) un gestor de contexto persistente, capaz de mantener la coherencia de la conversación a lo largo de múltiples sesiones y canales; (ii) un motor de razonamiento sobre grafos de conocimiento y ontologías institucionales, que permite inferir relaciones y resolver consultas más allá de la recuperación literal de información (Braun

et al., 2017); y (iii) componentes multimodales, que habilitan interacciones en voz, texto e interfaces gráficas, enriqueciendo la experiencia del usuario (Radziwill y Benton, 2017). Esta arquitectura híbrida, que combina representaciones neuronales (basadas en *transformers*) con estructuras de conocimiento simbólico, constituye un paradigma emergente en la investigación de IA conversacional.

Un aspecto distintivo de los asistentes virtuales es su capacidad de personalización dinámica. El sistema no solo procesa consultas aisladas, sino que aprende de los patrones de interacción de cada usuario, ajustando su comportamiento en función del historial y las preferencias. Este mecanismo se formaliza como un problema de optimización supervisada, donde para cada usuario u se ajusta un vector de parámetros θ_u que modela sus necesidades. El proceso de adaptación busca minimizar la pérdida:

$$\mathscr{L}(\theta_u) = \frac{1}{N}\sum_{i=1}^{N} \ell(\hat{y}_i(\theta_u), y_i),$$

donde $\hat{y}_i$ es la predicción del asistente y y_i la respuesta esperada en el contexto académico. Gracias a esta aproximación, el sistema puede generar recomendaciones individualizadas, tales como itinerarios académicos sugeridos, alertas sobre plazos administrativos o recursos de aprendizaje adaptados.

Otro componente esencial es la proactividad. A diferencia de los sistemas reactivos, los asistentes virtuales son capaces de anticiparse a las necesidades de los usuarios mediante modelos predictivos. En entornos universitarios, esta capacidad se traduce en sistemas de recomendación que identifican asignaturas adecuadas según el rendimiento previo, o en mecanismos de predicción de abandono (*dropout prediction*) que permiten emitir alertas tempranas (Yadegaridehkordi et al., 2019). Desde un punto de vista técnico, esta funcionalidad requiere un sistema de monitorización en tiempo real que detecte eventos críticos en los sistemas institucionales y active rutinas conversacionales adaptadas al contexto.

La evaluación de asistentes virtuales requiere un marco de métricas que vaya más allá de los indicadores clásicos de exactitud o latencia. En este sentido, se incluyen medidas como la *Task Completion Rate* (TCR), que cuantifica la proporción de procesos académicos completados con éxito mediante el asistente, y el *User Adaptation Score*, que mide el grado de personalización alcanzado en las interacciones (Braun et al., 2017). Estas métricas, complementadas con estudios de usabilidad y satisfacción del usuario, permiten evaluar tanto la eficacia técnica como el

impacto institucional de la herramienta.

Una de las principales distinciones entre chatbots y asistentes virtuales radica en la dimensión de la personalización. Mientras que los primeros tienden a ofrecer respuestas homogéneas a partir de repositorios de información comunes, los asistentes virtuales son capaces de adaptar su comportamiento en función del usuario, teniendo en cuenta preferencias individuales, historial de consultas y progresión académica. En universidades, esta capacidad se traduce en ofrecer recomendaciones personalizadas sobre asignaturas, alertas sobre plazos administrativos relevantes o sugerencias de recursos de aprendizaje complementarios (Villegas-Ch et al., 2021). Para lograr este nivel de adaptación, resulta fundamental la integración con sistemas de información estudiantil (SIS), plataformas de gestión del aprendizaje (LMS) y repositorios bibliográficos, lo que implica abordar retos asociados a interoperabilidad, sincronización de datos y cumplimiento normativo en protección de la privacidad.

Los asistentes virtuales representan un avance significativo respecto a los chatbots tradicionales, al incorporar una mayor complejidad en su arquitectura, mayores capacidades de razonamiento y un nivel más profundo de personalización. En el ámbito universitario, estas características les permiten funcionar como herramientas de apoyo inteligentes, permanentes y adaptativas, beneficiando tanto la gestión administrativa como los procesos de enseñanza-aprendizaje. No obstante, su despliegue implica desafíos técnicos relevantes, entre ellos la escalabilidad de la infraestructura, la adecuada gobernanza de los datos institucionales y la sostenibilidad de los modelos de entrenamiento, aspectos que deben ser contemplados de forma crítica en cualquier iniciativa de implementación.

2.1.3. Agentes de IA

Los agentes de IA constituyen sistemas computacionales autónomos diseñados para percibir su entorno, procesar información y ejecutar acciones con el objetivo de maximizar una medida de desempeño definida. A diferencia de los chatbots y asistentes virtuales, que operan principalmente en marcos conversacionales delimitados o bajo supervisión, los agentes de IA integran capacidades avanzadas de razonamiento, aprendizaje automático y adaptación contextual, lo que les permite tomar decisiones complejas en tiempo real y actuar sin intervención humana continua (Wooldridge, 2009).

Desde un punto de vista formal, un agente puede representarse como una función de mapeo:

$$f : P^* \rightarrow A$$

donde P^* es el conjunto de secuencias de percepciones y A el conjunto de acciones posibles. En entornos dinámicos e inciertos, esta formulación se extiende a procesos de decisión de Markov (MDP) y POMDP, donde el agente debe seleccionar políticas $\pi(s)$ que maximicen la recompensa esperada a lo largo del tiempo (Sutton, Barto et al., 1998). El uso de aprendizaje por refuerzo profundo (Deep Reinforcement Learning, DRL) ha potenciado la capacidad de los agentes para aprender estrategias óptimas directamente a partir de experiencias de interacción con el entorno.

Históricamente, la evolución de los agentes de IA se remonta a los sistemas expertos de la década de 1980, basados en reglas simbólicas y representaciones lógicas. Con el avance de los enfoques estadísticos y del aprendizaje automático, fue posible construir agentes con la capacidad de gestionar escenarios de incertidumbre y de ajustarse de manera progresiva a diferentes dominios de aplicación. La convergencia reciente de arquitecturas neuronales profundas y modelos de lenguaje preentrenados (como GPT o BERT) ha posibilitado agentes híbridos con capacidades tanto de razonamiento simbólico como de procesamiento contextual del lenguaje natural (McLean et al., 2023; Tunstall, Von Werra y Wolf, 2022).

En cuanto a su clasificación, los agentes de IA suelen distinguirse en: (i) agentes reflejos simples, que responden a estímulos inmediatos bajo esquemas condición–acción; (ii) agentes basados en modelos, que mantienen una representación interna del entorno para razonar sobre él; (iii) agentes orientados a objetivos, que planifican secuencias de acciones para alcanzar metas definidas; y (iv) agentes basados en utilidad, que seleccionan acciones evaluando el valor esperado de distintos estados futuros (Al-Jaf et al., 2024). Este marco se amplía con los sistemas multiagente (MAS), donde múltiples agentes colaboran o compiten entre sí para resolver problemas distribuidos de gran escala, como la gestión académica y administrativa en universidades (Albrecht, Christianos y Schäfer, 2024).

En el ámbito universitario, los agentes de IA ofrecen un potencial transformador. Mientras que un chatbot se orienta principalmente a resolver consultas específicas, un agente de IA posee la capacidad de desempeñar un rol más amplio como orquestador de procesos, integrando información distribuida en diferentes sistemas, automatizando procedimientos administrativos y anticipando de manera proactiva las demandas de los estudiantes. Ejemplos de aplicación incluyen agentes que: (i) integrados con plataformas de e-learning, recomiendan asignaturas

personalizadas según el historial académico; (ii) predicen riesgo de abandono estudiantil mediante análisis de patrones de interacción; (iii) optimizan la asignación de aulas y recursos docentes; o (iv) coordinan múltiples servicios institucionales bajo un esquema multiagente. En estos escenarios, la capacidad de percepción, planificación y acción autónoma diferencia al agente de IA de otros sistemas conversacionales más limitados.

La Tabla 3 presenta una comparación entre chatbots, asistentes virtuales y agentes de IA, destacando las diferencias en autonomía, nivel de complejidad y ámbitos de aplicación educativa. En ella se observa cómo los agentes representan el estadio más avanzado dentro del espectro de sistemas conversacionales implementados en el ámbito universitario.

Tabla 3: Comparación entre chatbots, asistentes virtuales y agentes de IA.

Criterio	**Chatbots**	**Asistentes virtuales**	**Agentes de IA**
Dominio	Resolución de consultas puntuales.	Acompañamiento personalizado y proactivo.	Gestión autónoma de objetivos múltiples en entornos complejos.
Autonomía	Reactivos, dependen de entradas explícitas.	Proactivos, anticipan necesidades.	Autónomos, planifican y ejecutan acciones en función de metas.
Tecnologías de soporte	PLN (NLU, NLG), reglas y ML.	PLN + grafos de conocimiento, razonamiento automático, multimodalidad.	PLN + planificación automática, aprendizaje por refuerzo, sistemas multiagente.
Gestión de contexto	Limitada a una sesión.	Persistente en múltiples sesiones.	Dinámica, con razonamiento a largo plazo y coordinación interagente.
Aplicaciones universitarias	FAQs y trámites básicos.	Tutoría académica personalizada, recordatorios inteligentes.	Optimización de recursos, predicción de abandono, gestión coordinada de procesos institucionales.

Fuente: elaboración propia.

No obstante, la implementación de agentes de IA en entornos académicos plantea retos técnicos y éticos adicionales. En términos de ciberseguridad, se requiere garantizar autenticación

robusta (SSO, OAuth2), control granular de accesos y mecanismos de auditoría que permitan trazar cada acción ejecutada por el agente. Asimismo, la transparencia y explicabilidad de las decisiones (XAI) resultan críticas para asegurar la confianza institucional, evitando la percepción de los agentes como "cajas negras" que toman decisiones opacas. Finalmente, la escalabilidad y el mantenimiento continuo del conocimiento (continual learning) se configuran como requisitos indispensables para mantener la vigencia de los agentes en contextos universitarios altamente dinámicos (Al-Jaf et al., 2024).

2.2. Procesamiento de lenguaje natural

2.2.1. Evolución del PLN:

El PLN constituye una de las áreas más relevantes de la inteligencia artificial, cuyo objetivo es dotar a las máquinas de la capacidad de comprender, interpretar y generar lenguaje humano en diferentes contextos. Su evolución histórica refleja un tránsito desde enfoques simbólicos y basados en reglas hacia modelos estadísticos, y finalmente hacia arquitecturas neuronales profundas que hoy sustentan los grandes modelos de lenguaje (LLMs) (Basha et al., 2023).

En su primera etapa, durante las décadas de 1950 y 1960, predominaban los sistemas simbólicos, fundamentados en gramáticas formales y reglas de reescritura. Ejemplos paradigmáticos incluyen los parsers sintácticos de Chomsky y sistemas como ELIZA, que operaban mediante emparejamiento de patrones. Aunque útiles en dominios restringidos, estos enfoques resultaban poco escalables debido a la necesidad de definir manualmente extensas bases de reglas (Weizenbaum, 1966).

A partir de la década de 1980, el campo experimentó un giro hacia métodos estadísticos, impulsados por la disponibilidad de grandes corpus textuales y el auge de la lingüística computacional. Modelos como los n-gramas permitieron capturar regularidades probabilísticas en el uso del lenguaje, mejorando tareas de modelado del habla y desambiguación léxica (Yu et al., 2011). El uso de modelos ocultos de Markov (HMM) y clasificadores como Naive Bayes y Máquinas de Vectores de Soporte (SVM) consolidó esta fase, posibilitando avances en etiquetado morfosintáctico, reconocimiento de voz y traducción automática (Hristea, 2025).

La siguiente etapa estuvo marcada por la incorporación del aprendizaje profundo. El empleo de redes neuronales recurrentes (RNN) y sus variantes (LSTM, GRU) permitió modelar dependencias a largo plazo en secuencias de texto, mejorando notablemente tareas como el reconocimiento del habla y la generación automática (Alonso et al., 2024). Sin embargo, estas

arquitecturas presentaban limitaciones en el procesamiento paralelo y la captura de dependencias de largo alcance, lo que abrió paso a modelos más sofisticados.

El cambio decisivo en el procesamiento del lenguaje natural ocurrió con la aparición de los mecanismos de atención y, más tarde, con la arquitectura Transformer, la cual reemplazó las estructuras recurrentes por esquemas de autoatención altamente paralelizables (Tunstall, Von Werra y Wolf, 2022). Este cambio revolucionó el PLN, posibilitando la aparición de modelos preentrenados a gran escala como BERT y GPT, que consolidaron un nuevo paradigma basado en el preentrenamiento y la adaptación a tareas específicas mediante *fine-tuning*. Con ello, el campo del PLN pasó de sistemas especializados y dependientes de reglas a modelos generales capaces de transferir conocimiento entre dominios con una eficacia sin precedentes.

La evolución del PLN refleja una progresiva abstracción desde reglas explícitas hacia representaciones distribuidas y jerárquicas del lenguaje. Este tránsito no solo ha incrementado la precisión en tareas clásicas como clasificación de texto, traducción y extracción de información, sino que ha habilitado la construcción de agentes conversacionales y asistentes inteligentes con un grado de naturalidad y contextualización anteriormente inalcanzable. Así, el PLN constituye el núcleo metodológico sobre el cual se sustentan los sistemas conversacionales descritos en las secciones anteriores.

2.2.2. Transformers

La introducción de la arquitectura *Transformer* supuso un punto de inflexión en el procesamiento de lenguaje natural, al reemplazar las limitaciones de las redes recurrentes (RNN, LSTM) mediante un mecanismo de autoatención totalmente paralelo (Tyukin et al., 2024). Este cambio permitió modelar dependencias de largo alcance en secuencias textuales de manera más eficiente y escalable, constituyendo la base sobre la que se han construido los modelos de lenguaje más influyentes de la última década.

El núcleo matemático de los Transformers es el mecanismo de *self-attention*, mediante el cual cada token de entrada puede ponderar dinámicamente su relación con los demás elementos de la secuencia. Formalmente, dado un conjunto de vectores de entrada $X \in \mathbb{R}^{n \times d}$, se definen tres proyecciones lineales aprendibles:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V,$$

donde $W^Q, W^K, W^V \in \mathbb{R}^{d \times d_k}$ son matrices de parámetros. El cálculo de atención se expresa

como:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

lo que permite asignar pesos adaptativos a cada token en función de su relevancia contextual. Este proceso se aplica de forma paralela en varios cabezales de atención (*multi-head attention*), lo que permite al modelo capturar de manera simultánea diferentes patrones de dependencia y tipos de relaciones semánticas presentes en la secuencia.

El entrenamiento de los Transformers se fundamenta en el aprendizaje no supervisado a gran escala, con estrategias de preentrenamiento que marcan la diferencia entre modelos bidireccionales y autoregresivos. El modelo BERT se entrena mediante la técnica de *masked language modeling* (MLM), en la cual se oculta un subconjunto de tokens de la secuencia de entrada y el sistema debe predecirlos utilizando el contexto bidireccional disponible (Devlin et al., 2019). Sea $X = (x_1, \ldots, x_n)$ la secuencia original y $M \subseteq \{1, \ldots, n\}$ el conjunto de posiciones enmascaradas, la función de pérdida correspondiente se expresa como:

$$\mathscr{L}_{MLM} = -\sum_{i \in M} \log P(x_i \mid X_{\setminus M}),$$

lo que permite al modelo aprender dependencias contextuales amplias y simétricas en el texto.

En contraste, GPT (*Generative Pretrained Transformer*) sigue un paradigma autoregresivo, en el cual cada token se predice condicionalmente en función de los tokens precedentes. Formalmente, la optimización se plantea como:

$$\mathscr{L}_{GPT} = -\sum_{t=1}^{n} \log P(x_t \mid x_{<t}),$$

lo que dota al modelo de una notable capacidad para generar secuencias textuales coherentes y naturales en tareas de redacción y diálogo (Radford, Narasimhan et al., 2018). Así, la diferencia entre BERT y GPT refleja sus roles arquitectónicos: mientras BERT, basado en un encoder, se orienta hacia la comprensión profunda del lenguaje, GPT, fundamentado en un decoder, está diseñado principalmente para la generación fluida de texto.

El impacto de BERT y GPT ha sido decisivo en la revolución del PLN. La posibilidad de adaptar estos modelos a dominios específicos mediante *fine-tuning* ha disminuido de manera significativa la necesidad de contar con grandes volúmenes de datos etiquetados. Gracias a

ello, se han consolidado como arquitecturas de referencia en tareas como clasificación de texto, sistemas de respuesta automática, generación de resúmenes y traducción automática (Brown et al., 2020; Devlin et al., 2019). Del mismo modo, la arquitectura de los Transformers ha demostrado una notable capacidad de escalado, posibilitando el entrenamiento de modelos con miles de millones de parámetros. Los resultados empíricos han evidenciado que el desempeño sigue leyes de escalado, donde la mejora en precisión y generalización se comporta de manera aproximadamente logarítmica en función tanto del tamaño del corpus de entrenamiento como de la complejidad paramétrica del modelo (J. Kaplan et al., 2020). Este fenómeno, conocido como *scaling laws*, constituye uno de los pilares teóricos de los LLMs, que se abordarán en la siguiente subsección. La introducción de BERT y GPT no solo consolidó la arquitectura Transformer como estándar de facto en PLN, sino que redefinió las fronteras entre comprensión y generación de lenguaje natural. Su combinación de eficiencia computacional, escalabilidad y adaptabilidad a múltiples tareas constituye la base de los avances recientes en agentes conversacionales inteligentes, incluidos los diseñados para entornos universitarios.

2.2.3. Grandes modelos de lenguaje

Los LLMs constituyen la evolución más significativa del PLN en la última década, consolidándose como la base de los sistemas conversacionales contemporáneos. Estos modelos se caracterizan por contar con un número masivo de parámetros —del orden de cientos de miles de millones— y por ser entrenados mediante técnicas de aprendizaje auto-supervisado sobre volúmenes de texto a escala web (Brown et al., 2020). A diferencia de los enfoques tradicionales, que dependían de corpus anotados manualmente, los LLMs aprovechan el preentrenamiento masivo para aprender representaciones distribuidas del lenguaje que capturan regularidades sintácticas, semánticas y pragmáticas de manera emergente (Raffel et al., 2020).

Desde una perspectiva formal, el preentrenamiento de un LLM puede representarse como la optimización de una función de probabilidad condicional:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{x\sim D}\left[-\sum_{t=1}^{T} \log P_\theta(x_t \mid x_{<t})\right],$$

donde D es el corpus de entrenamiento, x_t es la palabra en la posición t y $x_{<t}$ representa la secuencia previa. El modelo parametrizado por θ se entrena para aproximar la distribución real del lenguaje, maximizando la probabilidad de las secuencias observadas. Este paradigma de modelado generativo, conocido como *Language Modeling*, constituye el núcleo de arquitecturas

como GPT (Radford, Narasimhan et al., 2018) y su posterior escalado hacia GPT-3 (Brown et al., 2020).

La capacidad de los LLMs para generalizar a tareas no vistas se debe al fenómeno de la *aprendizaje en contexto* (*in-context learning*), por el cual el modelo adapta su comportamiento a nuevas instrucciones sin necesidad de un ajuste fino explícito. Sin embargo, para su aplicación en dominios específicos como el académico, resulta crucial un proceso adicional de adaptación. Este puede lograrse mediante dos estrategias principales: (i) fine-tuning, en el cual los parámetros del modelo se actualizan utilizando datos institucionales (p.ej., normativas universitarias, calendarios académicos, reglamentos de matrícula); y (ii) aprendizaje con recuperación aumentada (Retrieval-Augmented Generation, RAG), donde el modelo se conecta a repositorios documentales y bases de datos en tiempo real, garantizando precisión y trazabilidad en las respuestas (P. Lewis et al., 2020).

En el ámbito universitario, la integración de LLMs enfrenta tanto oportunidades como desafíos. Por un lado, permiten construir agentes conversacionales capaces de procesar consultas complejas en lenguaje natural, ofreciendo respuestas personalizadas y contextualizadas que van más allá de las capacidades de chatbots convencionales. Por otro, la incorporación de información sensible, como historiales académicos o expedientes administrativos, exige la implementación de protocolos estrictos de seguridad y privacidad. En este sentido, técnicas de *fine-tuning* seguro y el uso de entornos aislados para el despliegue del modelo constituyen enfoques necesarios para mitigar riesgos de fuga de información.

Una dimensión adicional clave es la alineación de los LLMs con los objetivos pedagógicos e institucionales. Los modelos preentrenados tienden a reflejar sesgos presentes en sus datos de entrenamiento, lo cual puede derivar en respuestas inapropiadas o inconsistentes. Con el fin de mitigar esta limitación, se han desarrollado enfoques de alineación basados en retroalimentación humana (*Reinforcement Learning from Human Feedback, RLHF*), en los cuales el modelo adapta sus respuestas utilizando evaluaciones proporcionadas por personas que valoran criterios como la calidad, la relevancia y la corrección de las salidas generadas (Ouyang et al., 2022). En entornos universitarios, esta técnica permite afinar el comportamiento del sistema para que priorice la transparencia, la formalidad y la coherencia normativa.

Finalmente, el potencial de los LLMs en educación superior se amplifica cuando se integran en ecosistemas académicos complejos. Un agente conversacional basado en LLM puede no solo resolver consultas sobre normativas o trámites, sino también generar resúmenes de documentos

institucionales, asistir en la redacción académica o sugerir recursos bibliográficos pertinentes. Este enfoque sitúa a los LLMs como piezas centrales en la infraestructura digital universitaria, ofreciendo soporte tanto a estudiantes como a docentes y gestores en la gestión del conocimiento.

2.2.4. Modelos multimodales

El tránsito de los modelos de lenguaje hacia arquitecturas multimodales ha marcado un punto de inflexión tanto en el PLN como en la IA en su conjunto. A diferencia de los modelos unimodales, entrenados únicamente sobre secuencias de texto, los modelos multimodales son capaces de procesar e integrar información proveniente de diferentes fuentes —texto, imagen, audio o vídeo—, lo que les permite construir representaciones conjuntas del conocimiento y habilitar interacciones más ricas y naturales con los usuarios (Baltrušaitis, Ahuja y Morency, 2018). Esta capacidad resulta particularmente relevante en contextos universitarios, donde la información institucional y académica no se encuentra exclusivamente en formato textual, sino también en gráficos, planos, documentos escaneados o grabaciones de voz.

De manera formal, un modelo multimodal puede definirse como una función que aprende una representación común entre distintas modalidades:

$$f : (X^{(1)}, X^{(2)}, \ldots, X^{(m)}) \to Z,$$

En este marco, cada $X^{(i)}$ representa el espacio correspondiente a una modalidad particular (como texto, imagen o audio), mientras que Z constituye un espacio latente común. El proceso de entrenamiento busca establecer una correspondencia semántica entre las diferentes modalidades, de forma que las representaciones de pares coherentes (por ejemplo, una frase descriptiva y la imagen que la ilustra) se proyecten cercanas en el espacio latente Z, mientras que aquellas que no guardan relación queden más alejadas (Radford, Narasimhan et al., 2018). Este enfoque constituye la base de arquitecturas como CLIP (Contrastive Language–Image Pretraining), las cuales utilizan una función de pérdida contrastiva para aprender representaciones compartidas de texto e imagen:

$$\mathscr{L}_{\text{contrastiva}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(z_i^{\text{text}}, z_i^{\text{img}})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(z_i^{\text{text}}, z_j^{\text{img}})/\tau)},$$

donde sim denota una función de similitud (usualmente el coseno) y τ corresponde a un hiperparámetro de temperatura que regula la dispersión de las probabilidades (Radford, Kim et al., 2021).

En el ámbito conversacional, la multimodalidad permite diseñar agentes capaces de: (i) interpretar consultas orales a través de reconocimiento automático del habla (ASR) y responder mediante síntesis de voz (TTS); (ii) procesar documentos escaneados o imágenes de horarios y convertirlos en información estructurada; y (iii) combinar texto, gráficos y tablas en la respuesta al usuario, aumentando la expresividad y la claridad de la interacción (Tsimpoukelli et al., 2021). Estas capacidades amplían de forma significativa el alcance de los sistemas conversacionales en educación superior, permitiendo escenarios como asistentes que interpretan fotografías de formularios administrativos, transcriben y resumen conferencias grabadas, o sugieren recursos visuales y multimedia para apoyar procesos de aprendizaje.

La multimodalidad también plantea desafíos técnicos relacionados con la alineación temporal y semántica entre modalidades. En tareas como la integración de voz y texto, el sistema debe alinear secuencias de distinta longitud y granularidad, lo que exige técnicas de atención cruzada (*cross-attention*) que ponderan de manera dinámica la contribución de cada modalidad. Asimismo, el entrenamiento de modelos multimodales requiere grandes volúmenes de datos anotados de forma coherente en múltiples formatos, lo cual implica costes computacionales y riesgos de sesgo derivados de la disponibilidad desigual de recursos multimodales (Akbari et al., 2021).

En el contexto universitario, la aplicación de modelos multimodales abre la puerta a agentes conversacionales más inclusivos y accesibles. Por ejemplo, estudiantes con dificultades visuales pueden interactuar mediante voz, mientras que aquellos con discapacidades auditivas pueden beneficiarse de respuestas en texto enriquecidas con elementos gráficos. Asimismo, la posibilidad de combinar de manera integrada texto, imágenes y audio habilita el desarrollo de entornos inteligentes que no solo apoyan la consulta de información administrativa, sino que también potencian los procesos de enseñanza y aprendizaje. Esto incluye funcionalidades como la elaboración de explicaciones enriquecidas con representaciones visuales o la adaptación de materiales educativos a distintos formatos según las necesidades del estudiante.

En suma, los modelos multimodales constituyen el siguiente paso en la evolución de los sistemas conversacionales universitarios, al permitir interacciones más naturales, ricas y adaptadas a la diversidad de modalidades con que se presenta la información. La combinación de modelos multimodales con LLMs y arquitecturas conversacionales no solo expande el abanico de capacidades técnicas de los agentes, sino que también favorece el desarrollo de experiencias educativas más accesibles, personalizadas y eficientes.

2.3. Técnicas de adaptación de modelos

Los LLMs destacan por su capacidad de generalización, la cual se deriva de su entrenamiento a partir de grandes volúmenes de datos diversos y heterogéneos. Sin embargo, su implementación en contextos institucionales específicos, como las universidades, requiere estrategias de adaptación que permitan refinar su comportamiento, alinearlo con objetivos académicos y garantizar tanto precisión como trazabilidad en las respuestas (Raffel et al., 2020). En este sentido, las técnicas de *fine-tuning* y de generación aumentada por conocimiento (Knowledge-Augmented Generation, KAG) constituyen los enfoques más relevantes.

2.3.1. Fine-tuning

El proceso de *fine-tuning* implica la adaptación de los parámetros θ de un modelo previamente entrenado en un dominio amplio, con el fin de especializarlo en un conjunto de datos particulares D_{dom}. Matemáticamente, este ajuste se formula como la optimización de la función de pérdida:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim D_{dom}} \ell(f_\theta(x), y),$$

donde ℓ representa la función de coste y f_θ corresponde a la predicción generada por el modelo tras la actualización de sus parámetros (Howard y Ruder, 2018). Esta estrategia permite especializar al modelo en el procesamiento de textos académicos, normativas universitarias o documentos administrativos.

En términos aplicados, el proceso de *fine-tuning* puede abordarse a través de tres estrategias principales:

1. **Ajuste completo (Full fine-tuning):** se actualizan todos los parámetros del modelo, lo que proporciona el mayor grado de especialización. Sin embargo, este enfoque conlleva altos requerimientos computacionales y una mayor probabilidad de sobreajuste.

2. **Ajuste parcial:** únicamente se entrenan determinadas capas, como la de salida o ciertos adaptadores intermedios, lo que disminuye la complejidad y los costes del entrenamiento (Houlsby et al., 2019).

3. **Ajuste eficiente en parámetros (Parameter-efficient fine-tuning, PEFT):** consiste en añadir módulos adicionales con bajo coste paramétrico —por ejemplo, LoRA (*Low-Rank*

Adaptation)— que permiten adaptar el modelo sin necesidad de modificar la mayoría de sus parámetros (Hu et al., 2022).

En entornos universitarios, los enfoques de PEFT resultan particularmente atractivos, dado que permiten desplegar agentes conversacionales adaptados a contextos institucionales específicos sin necesidad de recursos computacionales prohibitivos.

2.3.2. Generación aumentada con recuperación (RAG)

Los modelos de lenguaje de gran escala (LLMs), al estar entrenados sobre corpus estáticos, presentan limitaciones para incorporar información que cambia dinámicamente en el tiempo, lo que puede conducir a respuestas desactualizadas o inexactas. Una estrategia ampliamente adoptada para superar esta restricción es la *Retrieval-Augmented Generation* (RAG), en la cual el modelo se conecta con un módulo externo de recuperación documental que le proporciona evidencia verificable en cada interacción (P. Lewis et al., 2020). Estas aproximaciones se han aplicado también a consultas administrativas no académicas. Por ejemplo, Sarmiento (2025) analizan cómo RAG mejora el control de alucinaciones al restringir respuestas a un corpus curado específico para personal y administración, remarcando la importancia del diseño del backend para sostener precisión en tareas institucionales.

Formalmente, dado un enunciado de entrada q, el sistema recupera un conjunto de documentos relevantes $R(q)$ desde una base de conocimiento vectorial o índice semántico. La probabilidad de la respuesta final se condiciona en el par $(q, R(q))$, de manera que:

$$P(y \mid q) = \sum_{d \in R(q)} P(y \mid q, d) \cdot P(d \mid q),$$

donde $P(d \mid q)$ representa la relevancia estimada de cada documento recuperado y $P(y \mid q, d)$ la probabilidad de generar la respuesta y apoyada en dicho documento.

Este enfoque introduce varias ventajas frente al uso directo de LLMs:

- Garantiza la trazabilidad de las respuestas, pues cada salida puede vincularse con documentos fuente.
- Reduce la incidencia de "alucinaciones", al condicionar la generación en evidencias externas verificables.

- Facilita la actualización dinámica, ya que basta con modificar el repositorio documental sin necesidad de reentrenar el modelo.

En el contexto universitario, el uso de RAG es particularmente relevante. Permite que un agente conversacional consulte en tiempo real reglamentos académicos, calendarios, normativas de matrícula o convocatorias de becas, asegurando que la información entregada esté alineada con las versiones oficiales vigentes. Asimismo, su integración con repositorios heterogéneos (PDFs, hojas de cálculo, bases de datos institucionales) ofrece una vía flexible para unificar el acceso a información dispersa.

2.3.3. Discusión técnica

La adopción de *Retrieval-Augmented Generation* (RAG) en sistemas conversacionales universitarios plantea un conjunto de ventajas metodológicas y desafíos técnicos que condicionan su aplicabilidad práctica. Su principal fortaleza reside en externalizar el conocimiento hacia repositorios documentales dinámicos, lo que permite reducir costes de reentrenamiento y garantizar respuestas actualizadas en dominios sujetos a cambios frecuentes, como normativas académicas o calendarios institucionales.

Desde una perspectiva computacional, el desempeño de un sistema RAG depende críticamente de tres elementos: (i) la calidad de los *embeddings* utilizados para representar consultas y documentos, los cuales deben preservar propiedades semánticas relevantes en espacios vectoriales de alta dimensión; (ii) la eficiencia y precisión del motor de recuperación, donde parámetros como el número k de documentos candidatos y el umbral de similitud τ condicionan el equilibrio entre cobertura y relevancia; (iii) la integración coherente de múltiples evidencias en la fase generativa, que requiere estrategias de fusión para evitar contradicciones o redundancias en la salida final.

Formalmente, la robustez del sistema puede modelarse como un problema de optimización conjunta:

$$\max_{R(q)} \mathbb{E}_{q\sim Q}\left[\alpha\cdot \text{Precisión}(q,R(q)) + \beta\cdot \text{Cobertura}(q,R(q)) - \lambda\cdot \text{Costo}(R(q))\right],$$

donde α,β,λ representan coeficientes de ponderación que balancean exactitud, diversidad y coste computacional del proceso de recuperación.

En términos de escalabilidad, RAG introduce una ventaja significativa al desacoplar el

modelo de lenguaje del repositorio de conocimiento, permitiendo que la actualización del sistema se limite a la indexación de nuevos documentos. No obstante, esta propiedad depende de la implementación de pipelines de MLOps que garanticen la correcta ingesta, normalización e indexación de fuentes heterogéneas (PDF, hojas de cálculo, bases de datos académicas), asegurando consistencia y trazabilidad.

Un desafío adicional es la sensibilidad del sistema a errores en la fase de recuperación. Una recuperación imprecisa conduce a respuestas incorrectas aunque el modelo generativo sea altamente competente, lo que desplaza el cuello de botella hacia el diseño y evaluación del motor de búsqueda semántico. Este fenómeno se conoce como *garbage in, garbage out* y obliga a complementar RAG con métricas específicas de recuperación (recall@k, MRR) además de métricas conversacionales (First Turn Resolution, satisfacción del usuario).

En entornos universitarios, la implementación de RAG también exige contemplar aspectos éticos y de gobernanza de datos. La recuperación debe restringirse a fuentes verificadas para evitar la propagación de información no oficial, y el sistema debe garantizar trazabilidad mediante la citación explícita de documentos fuente.

En suma, la discusión técnica muestra que RAG ofrece un marco flexible, escalable y coste-eficiente para dotar de veracidad y actualización a los agentes conversacionales universitarios. Sin embargo, su efectividad está condicionada a la correcta orquestación de las fases de indexación, recuperación y fusión, que deben ser diseñadas con criterios tanto computacionales como institucionales para asegurar precisión, transparencia y confianza en la interacción.

2.4. Chatbots en educación superior

La aplicación de chatbots en universidades constituye un campo emergente en el que confluyen técnicas avanzadas de procesamiento de lenguaje natural (PLN), aprendizaje automático y sistemas de información académica. Su objetivo central es disminuir las barreras en la comunicación entre los estudiantes y la administración universitaria, optimizando la realización de tareas clave como la inscripción en asignaturas, la tramitación de becas, la emisión de certificados y la atención a consultas recurrentes. Desde una perspectiva institucional, los chatbots permiten descentralizar la atención y dotar de mayor escalabilidad a los servicios académicos, sin comprometer la trazabilidad de la información (Wollny et al., 2021).

2.4.1. Modelado matemático de chatbots universitarios

El funcionamiento de un chatbot universitario puede modelarse formalmente como un proceso de decisión secuencial. Una representación ampliamente utilizada es el marco de los POMDP, donde el sistema no tiene acceso completo al estado real del usuario, sino a observaciones derivadas de la consulta en lenguaje natural (Young et al., 2013). Sea S el espacio de estados latentes que representan las necesidades del estudiante (por ejemplo, *consulta de horarios*, *plazos de becas*, *verificación de requisitos de matrícula*), A el conjunto de acciones disponibles para el chatbot (responder, solicitar aclaración, recuperar documentos), y O el conjunto de observaciones textuales derivadas de la entrada del usuario. La dinámica del sistema puede describirse como:

$$P(s_{t+1} \mid s_t, a_t), \quad P(o_t \mid s_t, a_t),$$

donde la política de decisión $\pi(a_t \mid h_t)$ asigna probabilidades sobre acciones dado el historial de interacción h_t. El objetivo es encontrar la política óptima π^* que maximice la recompensa acumulada esperada:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R(s_t, a_t)\right],$$

con $\gamma \in [0,1]$ un factor de descuento temporal y $R(s_t, a_t)$ la función de recompensa que penaliza respuestas incorrectas y premia resoluciones rápidas en el primer turno (*First Turn Resolution, FTR*).

En términos prácticos, esta formulación permite entrenar políticas conversacionales mediante aprendizaje por refuerzo profundo (Deep Reinforcement Learning, DRL), donde el sistema ajusta sus decisiones a partir de interacciones reales con usuarios (Su et al., 2016).

2.4.2. Integración con modelos de lenguaje

El módulo de PLN de un chatbot universitario puede construirse a partir de modelos de lenguaje preentrenados, como BERT o GPT, ajustados mediante *fine-tuning* sobre documentos normativos y administrativos de la institución. En este proceso, el objetivo es optimizar la función de pérdida asociada a la clasificación de intenciones:

$$\mathscr{L}(\theta) = -\sum_{(x,y)\in D_{dom}} \log P_\theta(y \mid x),$$

donde x denota la consulta formulada por el estudiante y y corresponde a la intención asociada (por ejemplo, *solicitud de beca* o *verificación de calendario*), empleando un corpus específico D_{dom}.

Con el fin de garantizar precisión y trazabilidad en las respuestas, se incorpora un esquema de generación aumentada con recuperación (*Retrieval-Augmented Generation*, RAG). En este enfoque, para una consulta q, el sistema localiza documentos relevantes $R(q)$ en repositorios institucionales y condiciona la generación de la salida en el par $(q, R(q))$:

$$P(y \mid q) = \sum_{d \in R(q)} P(y \mid q, d) \cdot P(d \mid q).$$

Este enfoque híbrido asegura que la información provenga de fuentes oficiales (reglamentos, bases de datos académicas), reduciendo el riesgo de respuestas incorrectas.

2.4.3. Aplicaciones en universidades

La implementación práctica de chatbots universitarios ya muestra resultados notables. La Tabla 4 sintetiza algunos casos de aplicación y sus impactos medidos en eficiencia administrativa, retención estudiantil y escalabilidad de la atención.

Estos casos confirman que los chatbots no solo reducen tiempos de respuesta, sino que también impactan en métricas críticas de desempeño institucional como la retención estudiantil, la eficiencia administrativa y la satisfacción en los servicios de apoyo. Cuando los chatbots universitarios se representan bajo el marco de POMDPs y se entrenan mediante técnicas de aprendizaje por refuerzo, complementados con esquemas de RAG, se configuran como componentes clave dentro de la infraestructura digital de las universidades, contribuyendo de manera decisiva a los procesos de transformación tecnológica en la educación superior.

Tabla 4: Casos reales de chatbots en educación superior: aplicaciones y resultados cuantitativos

Universidad / Chatbot	**Uso principal**	**Resultados cuantitativos**
Deakin University — "Genie"	Asistente administrativo personal integrando LMS, biblioteca y servicios.	Hasta 12 000 conversaciones diarias; muy popular en estudiantes de primer año. [1]
Georgia State University — "Pounce"	Chatbot para admisión (summer melt) y consultas de estudiantes entrantes.	200 000 respuestas entregadas; reducción del "summer melt" en un 22 %, retención mejoró significativamente. [2]
University of the Free State — EduBot	Bot administrativo multicanal (web y WhatsApp) para soporte institucional.	Reducción del 71 % en tickets de soporte; 83 % menos consultas vía chat en vivo; duplicó alcance estudiantil. [3]
University of Milano-Bicocca — Unimib Assistant	Chatbot RAG para consultas administrativas y académicas de estudiantes.	Alta usabilidad percibida; respuestas estructuradas, aunque con limitaciones en precisión y enlaces funcionales. [4]

Fuente: elaboración propia.

[1] Deakin-Genie University
[2] Georgia State University
[3] University of the Free State (UFS)
[4] University of Milano-Bicocca

3. Metodología

3.1. Diseño metodológico

El presente trabajo adopta un diseño metodológico de carácter aplicado y tecnológico, articulado en torno a la metodología CRISP - DM (*Cross Industry Standard Process for Data Mining*). Este marco, ampliamente consolidado en proyectos de ciencia de datos e inteligencia artificial, organiza el desarrollo en seis fases iterativas: (i) comprensión del negocio, (ii) comprensión de los datos, (iii) preparación de los datos, (iv) modelado, (v) evaluación y (vi) despliegue.

El objetivo metodológico central es implementar un agente conversacional para el ámbito universitario, sustentado en LLMs, capaz de responder en lenguaje natural a consultas académicas y administrativas, garantizando precisión, trazabilidad y eficiencia. El estudio se clasifica como *exploratorio–aplicado*, en la medida en que verifica la viabilidad técnica del sistema, y adopta un enfoque *cuantitativo–computacional*, evaluando el rendimiento mediante métricas objetivas (precisión, cobertura y tiempo de respuesta).

3.2. Participantes

Dado que el trabajo tiene un carácter eminentemente técnico, no se incluyen participantes humanos en el proceso de validación. La evaluación del sistema se centró en métricas computacionales y en pruebas de funcionamiento en entornos controlados.

3.3. Instrumentos

Se utilizaron los siguientes instrumentos técnicos y analíticos:

- **Datasets institucionales**: documentos en formato CSV, PDF, DOCX y PPTX, que incluyen calendarios, reglamentos, manuales y guías académicas.
- **Plataformas y librerías**: `Chainlit` para la interfaz del chatbot, `ChromaDB` como base vectorial por su eficiencia en búsquedas aproximadas de alta dimensionalidad y su facilidad de actualización incremental frente a alternativas como FAISS o Milvus, y `OpenAIEmbeddings` para la generación de representaciones semánticas.
- **Modelos de PLN**: `ChatOpenAI` (gpt-4o-mini) como modelo de lenguaje principal, elegido por su equilibrio entre rendimiento, coste computacional y latencia en comparación con

modelos abiertos como LLaMA o DeepSeek, e integrado con la estrategia *Retrieval-Augmented Generation* (RAG).

- **Herramientas de evaluación**: métricas de rendimiento computacional (precisión de recuperación, latencia, tasa de error) y registros de interacción en tiempo real.

3.4. Procedimiento

El procedimiento metodológico se organizó en las fases de CRISP-DM, alineadas al desarrollo del agente conversacional:

1. **Comprensión del negocio**: definición del problema de dispersión informativa en entornos universitarios y establecimiento de criterios de éxito (precisión, reducción del tiempo de búsqueda).

2. **Comprensión de los datos**: análisis de la heterogeneidad de fuentes y formatos; detección de la dispersión en calendarios y normativas como principal obstáculo para el acceso a la información.

3. **Preparación de los datos**: implementación de un pipeline de preprocesamiento, incluyendo segmentación en fragmentos (*chunking*), extracción de embeddings mediante `OpenAIEmbeddings` y almacenamiento en `ChromaDB`, que se seleccionó por su flexibilidad para indexar documentos heterogéneos.

4. **Evaluación**: validación automática mediante métricas de precisión y relevancia ajustando parámetros del retriever (*k*, *min_score*).

5. **Despliegue**: implementación del chatbot en entorno web mediante `Chainlit`, con acceso multiplataforma y posibilidad de integración futura al campus virtual institucional.

La Figura 1 sintetiza la adaptación de CRISP-DM al presente trabajo, mientras que la Figura 2 describe la arquitectura técnica de la solución.

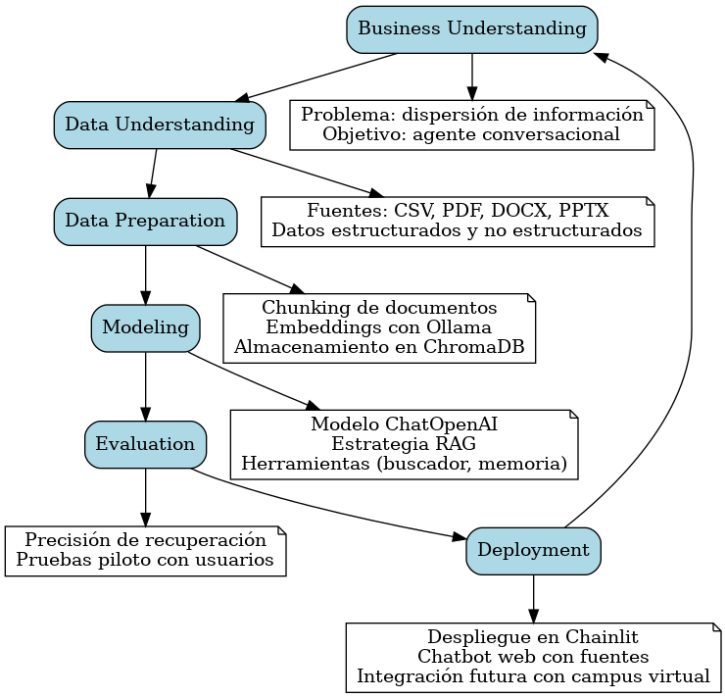


Figura 1: Adaptación de la metodología CRISP-DM al desarrollo del agente conversacional universitario.

3.5. Análisis de datos

El análisis de desempeño se dividió en dos niveles:

- **Evaluación técnica**: cálculo de precisión, cobertura de consultas, latencia promedio y tasa de error en el módulo de NLU.
- **Análisis de robustez**: verificación de la coherencia de respuestas y trazabilidad hacia documentos fuente mediante la estrategia RAG.

Los resultados obtenidos permiten validar el potencial del agente conversacional como infraestructura de apoyo académico y administrativo, en línea con los objetivos definidos en la fase inicial.

4. Resultados

4.1. Ingesta y preprocesamiento de la información

El primer paso del proceso metodológico consistió en la recopilación, estandarización y preparación de fuentes institucionales, con el propósito de conformar un repositorio documental robusto que sirviera como base de conocimiento para el agente conversacional. Dada la dispersión y heterogeneidad de la información académica, se adoptó una estrategia dual de ingesta. Por un lado, se aplicaron técnicas de *web scraping* sobre las páginas públicas de la universidad, utilizando la librería `WaterCrawl`, lo que permitió extraer de manera sistemática contenidos procedentes de portales académicos, repositorios institucionales y sistemas administrativos accesibles. Esta aproximación garantizó la obtención estructurada de la información sin necesidad de acceder a secciones privadas como el campus virtual, cuyo rastreo podría implicar restricciones técnicas o legales.

De manera complementaria, se integraron documentos oficiales en formato `PDF`, tales como reglamentos de matrícula, manuales administrativos, calendarios académicos, guías de becas y resoluciones institucionales. Estas fuentes constituyen la base normativa y procedimental de mayor autoridad, por lo que su correcta incorporación al repositorio resulta crítica para garantizar tanto la coherencia como la trazabilidad de las respuestas generadas por el sistema.

Una vez recopilada la información, se aplicó un proceso exhaustivo de preprocesamiento orientado a su normalización y a la optimización de la posterior indexación semántica. Dicho proceso incluyó la segmentación de los documentos en fragmentos de longitud controlada mediante técnicas de *chunking*, garantizando que cada unidad mantuviera coherencia interna al mismo tiempo que alcanzara el nivel de granularidad necesario para responder consultas específicas. Adicionalmente, se estableció un umbral de relevancia asociado a cada fragmento, aprovechando la función `similarity_search_with_score` en la etapa de recuperación. Este mecanismo permitió asignar un puntaje de similitud a cada chunk, filtrando aquellos con baja correspondencia y priorizando la selección de documentos más pertinentes para la generación de respuestas.

El procedimiento de limpieza incluyó la normalización de caracteres, la eliminación de símbolos no estándar y elementos decorativos de escaso valor informativo. Asimismo, se gestionaron los metadatos de manera diferenciada: se eliminaron aquellos irrelevantes, como numeraciones automáticas o encabezados redundantes, mientras que se añadieron metadatos personalizados en

los chunks —por ejemplo, la fuente documental— con el propósito de mejorar la trazabilidad y la transparencia en la recuperación de información.

El conjunto resultante de fragmentos procesados fue transformado en representaciones distribuidas de alta dimensionalidad mediante la librería `OpenAIEmbeddings`, utilizando el modelo `text-embedding-3-small`. Este modelo permitió capturar relaciones semánticas complejas y preservar dependencias contextuales necesarias para la recuperación en lenguaje natural. Una vez generados, los embeddings fueron almacenados en la base vectorial `ChromaDB`, seleccionada por su eficiencia en búsquedas aproximadas en espacios de gran dimensionalidad y por su capacidad de actualización incremental sin necesidad de reindexar la totalidad del corpus.

El resultado de esta fase fue la construcción de un repositorio documental heterogéneo, estructurado y optimizado para la recuperación semántica. La integración de documentos normativos obtenidos mediante *web scraping* y fuentes institucionales oficiales, junto con el preprocesamiento y la indexación vectorial, garantiza que el agente conversacional disponga de una base de conocimiento capaz de responder con precisión, coherencia y trazabilidad a las consultas académicas y administrativas formuladas en lenguaje natural.

4.2. Construcción del agente conversacional

El desarrollo del agente conversacional se estructuró bajo una arquitectura modular, concebida para garantizar flexibilidad, escalabilidad y trazabilidad en la gestión de consultas académicas y administrativas. La modularización se materializó en la implementación de dos programas principales: uno destinado a la indexación de datos y generación de *embeddings*, y otro orientado al funcionamiento del agente conversacional con recuperación de información mediante RAG. Esta separación funcional asegura que cada componente pueda evolucionar de manera independiente, facilitando la incorporación de mejoras o la sustitución de módulos sin comprometer la estabilidad global del sistema. Para la construcción de dichos módulos se empleó el entorno de desarrollo integrado (IDE) `Visual Studio Code`, lo que permitió una organización eficiente y transparente del código.

El núcleo del sistema está conformado por un modelo de lenguaje de gran escala (`ChatOpenAI GPT-4o-mini`), seleccionado por su capacidad de procesamiento contextual y por su equilibrio entre rendimiento y eficiencia computacional. Este modelo cumple la función de generador de respuestas, integrando la información recuperada desde el repositorio vectorial y produciendo salidas en lenguaje natural con un nivel de coherencia adaptado al dominio universitario. La

configuración del modelo se realizó bajo un esquema de *prompt engineering*, orientado a priorizar la precisión, la formalidad académica y la trazabilidad hacia documentos oficiales. Para la integración de los distintos componentes —incluyendo los *prompts*, la base de datos vectorial y el modelo de lenguaje— se empleó la librería `LangChain`. Este framework facilita la construcción de aplicaciones con LLMs mediante la definición de *chains*, lo que permite orquestar de manera eficiente procesos complejos como la recuperación de información contextual (RAG), la fusión de evidencias y la generación controlada de respuestas. El uso de `LangChain` no solo optimizó la interoperabilidad entre módulos, sino que también habilitó la extensibilidad del sistema, permitiendo la incorporación de nuevas herramientas o la reconfiguración de la cadena conversacional sin alterar la arquitectura principal.

Complementando al modelo generativo, se incorporó un módulo de recuperación semántica bajo el paradigma de *Retrieval-Augmented Generation* (RAG). Este componente constituye el puente entre las consultas en lenguaje natural formuladas por los usuarios y la base vectorial previamente construida. El proceso se inicia con la conversión de cada entrada en un embedding vectorial, el cual es comparado con el repositorio almacenado en `ChromaDB`. A partir de ello, se selecciona un conjunto de documentos candidatos en función de su similitud semántica, los cuales son posteriormente inyectados en el contexto de generación del LLM. De esta manera, la respuesta final no se fundamenta exclusivamente en el conocimiento estático del modelo, sino que se apoya en evidencias verificables provenientes de fuentes institucionales. Este diseño reduce de forma significativa la incidencia de alucinaciones y aporta un mecanismo de actualización dinámica que se adapta a la evolución de los documentos universitarios.

En paralelo, se implementó una interfaz conversacional mediante la librería `Chainlit`, utilizada como capa de *User Interface* (UI) que conecta al usuario final con los componentes internos del agente. La interfaz fue desarrollada como una aplicación web multiplataforma, accesible desde navegadores en dispositivos de escritorio y móviles, con el objetivo de maximizar la accesibilidad y fomentar la adopción institucional. Entre sus funcionalidades destacan la presentación de respuestas acompañadas de referencias explícitas a documentos fuente y la persistencia del contexto conversacional mediante gestión de memoria de sesión. Cabe señalar que el esquema de razonamiento reactivo (`ReAct Agent`) fue implementado a través de `LangChain`, mientras que `Chainlit` se limita a la interacción conversacional y a la visualización de resultados en tiempo real.

La Figura 2 ilustra la arquitectura general del agente, destacando la interacción entre los

módulos de preprocesamiento y embeddings, el motor de recuperación semántica y el modelo generativo. Este flujo asegura que cada consulta siga un ciclo bien definido: (i) formulación en lenguaje natural por parte del usuario, (ii) vectorización y búsqueda de evidencias relevantes, (iii) fusión contextual con el modelo de lenguaje y (iv) entrega de una respuesta fundamentada.

La separación clara de responsabilidades entre componentes no solo facilita la depuración en fases de prueba, sino que también habilita la incorporación de mecanismos de recuperación controlada de información. En este sentido, se dotó al agente de la herramienta `search_relevant_documents`, diseñada para acceder de manera eficiente a la base de datos vectorial de *embeddings*. Gracias a esta integración, y bajo el esquema de razonamiento reactivo gestionado por `LangChain`, el agente es capaz de decidir dinámicamente si una consulta puede resolverse únicamente con el conocimiento interno del modelo de lenguaje, o si resulta necesario activar el módulo de recuperación (*RAG*) para proporcionar una respuesta fundamentada en documentos oficiales.

Este diseño modular garantiza que el sistema pueda operar en diferentes niveles de profundidad informativa: por un lado, resolviendo interacciones rutinarias directamente desde el LLM, y por otro, recurriendo a fuentes institucionales verificadas cuando la precisión normativa o la trazabilidad lo requieran. Asimismo, esta arquitectura sienta las bases para futuras integraciones con sistemas externos, como gestores de identidad institucional o repositorios administrativos centralizados, reforzando tanto la escalabilidad como la interoperabilidad del agente.

En suma, la construcción del agente conversacional responde a un enfoque arquitectónico orientado a la robustez y la adaptabilidad. Su diseño modular, sustentado en la sinergia entre un LLM, un motor de recuperación semántica y una interfaz web escalable, constituye una solución técnica capaz de evolucionar con las necesidades de la universidad, garantizando precisión, transparencia y eficiencia en el acceso a la información.

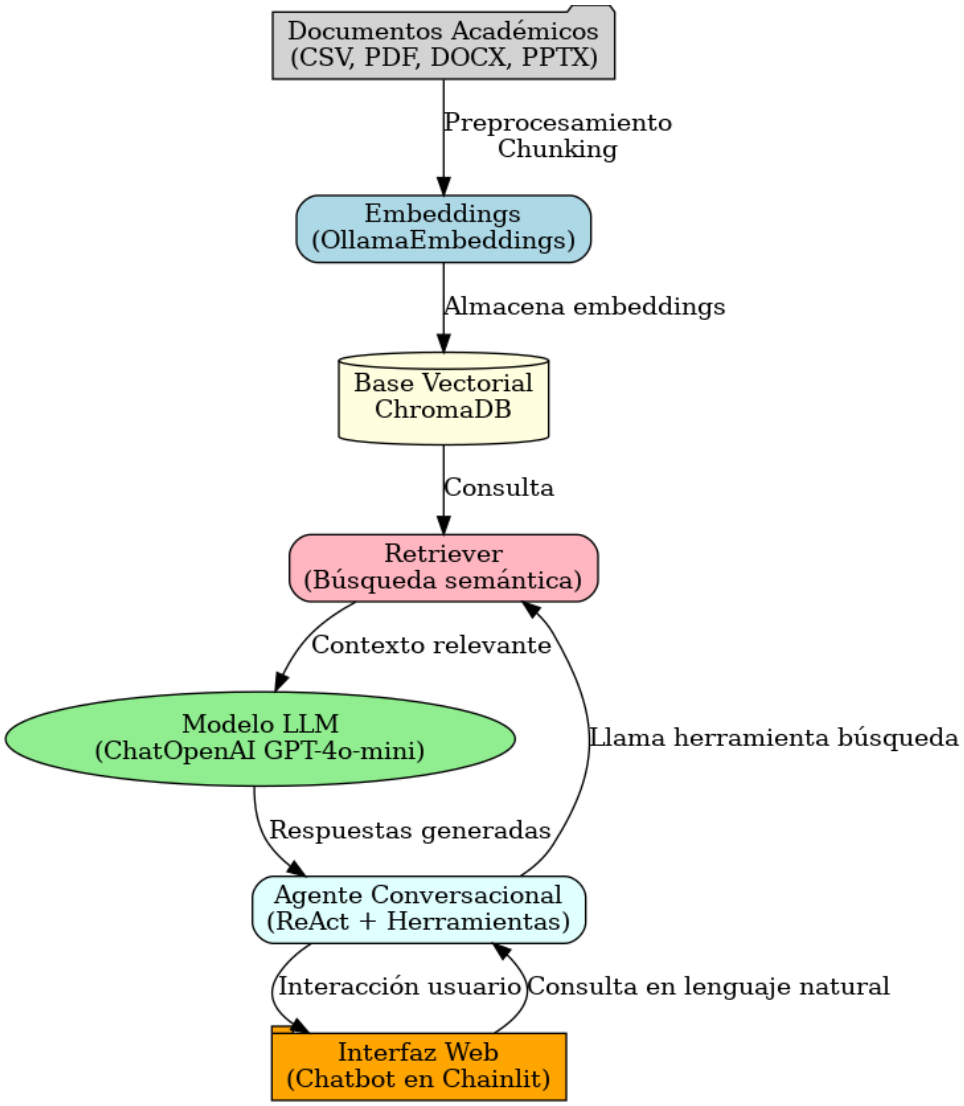


Figura 2: Arquitectura técnica del agente conversacional universitario.

4.3. Evaluación del sistema

La validación del agente se realizó mediante un escenario de *chat testing* utilizando un conjunto de preguntas frecuentes institucionales (*benchmark*). Las consultas se formularon en lenguaje natural y abarcaron distintas categorías: normativas académicas, trámites administrativos y calendarios universitarios.

Los resultados se evaluaron con métricas objetivas, diferenciando entre el desempeño técnico del modelo y la calidad conversacional. La Tabla 5 sintetiza los resultados preliminares obtenidos. Si bien el presente análisis se centró en métricas centrales —precisión, cobertura, latencia y FTR— que enmarcan adecuadamente el desempeño del prototipo, se reconoce que otras métricas más específicas de recuperación y comprensión, como *recall@k*, MRR, tasa de error en NLU

o *hallucination rate*, no fueron incluidas en esta fase debido a las características exploratorias del sistema y a la ausencia de recursos humanos especializados para llevar a cabo una anotación manual exhaustiva. Su incorporación queda identificada como una línea de trabajo futuro, orientada a ampliar la robustez y exhaustividad de la evaluación.

Tabla 5: Resultados de la evaluación del agente conversacional con escenario de chat testing.

Métrica	**Valor**	**Descripción**
Precisión de respuestas correctas	82.4 %	Proporción de respuestas verificadas como correctas frente al total de interacciones.
Cobertura de consultas	91.7 %	Porcentaje de consultas del benchmark que el agente fue capaz de gestionar.
Latencia promedio (mediana)	6,2 s	Tiempo medio de respuesta entre consulta y salida generada.
First Turn Resolution (FTR)	76.3 %	Porcentaje de consultas resueltas satisfactoriamente en el primer turno sin necesidad de reformulación.

Fuente: elaboración propia a partir de datos

4.4. Análisis cualitativo de las respuestas

Además de las métricas cuantitativas, se analizó la trazabilidad de las respuestas. El sistema logró vincular cada salida generada con su documento fuente, garantizando la transparencia de la información. Sin embargo, se observaron casos de redundancia en consultas con múltiples documentos relevantes, lo que resalta la necesidad de mejorar los mecanismos de fusión de evidencias.

Casos de fallo representativos y mitigaciones. Durante el escenario de *chat testing* se identificaron patrones de fallo representativos que ofrecen orientación clara para futuras iteraciones del sistema:

1. **Consultas sin cobertura.** Ejemplos como *"¿Cuándo se publican los horarios del máster?"* no pudieron ser respondidos debido a la ausencia de evidencias relevantes, generando

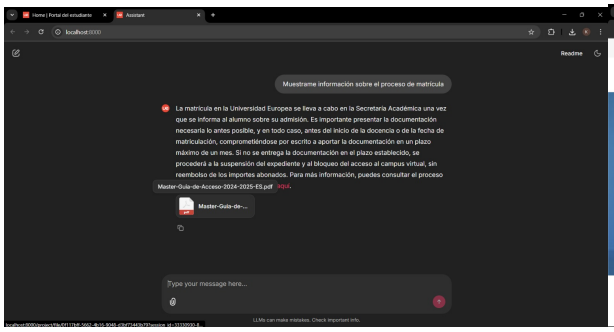
respuestas vacías. Esto refleja limitaciones en el corpus documental utilizado. *Mitigación:* inclusión de fuentes institucionales operativas aún no ingesadas (páginas de avisos y actualizaciones de facultad).

2. **Respuestas fuera de dominio.** En preguntas sobre asignaturas del máster, el sistema respondió con información de grados no relacionados, como Marketing o Fisioterapia. *Mitigación:* depuración del repositorio por metadatos de titulación y filtrado por programa académico para evitar colisiones semánticas entre carreras.

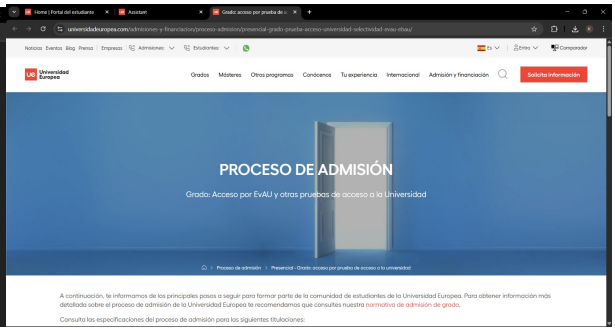
3. **Alucinaciones puntuales.** En una consulta sobre admisión, se generaron nombres y correos no presentes en las fuentes vinculadas. *Mitigación:* reforzar las plantillas de abstención para datos personales y condicionar las respuestas a la existencia explícita de evidencia verificada.

4. **Redundancia en la generación.** En casos donde múltiples fragmentos se consideraron relevantes, el sistema repitió contenido similar. *Mitigación:* aplicar fusión de evidencias con desduplicación semántica a nivel de *chunk* para aumentar la concisión de las respuestas.

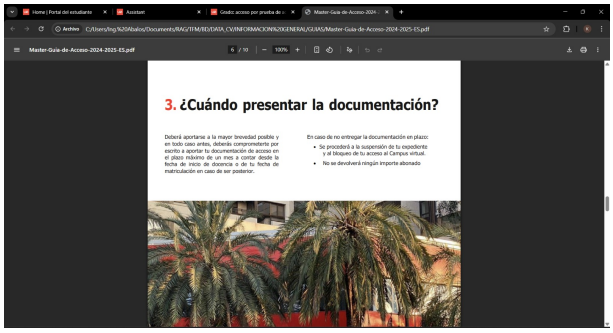
Este análisis cualitativo permite identificar áreas de mejora prioritarias, compensando la ausencia de métricas específicas de recuperación en esta entrega y fortaleciendo el marco de evaluación del prototipo. La Figura 3 presenta un ejemplo de interacción donde el agente responde a una consulta sobre fechas de matrícula, citando explícitamente la fuente documental correspondiente.



(a) Interacción 1



(b) Enlace web



(c) PDF indexado

Figura 3: Ejemplos de interacción en el *chat testing*: el agente cita normativa oficial en respuestas sobre trámites.

En conjunto, los resultados preliminares evidencian que la estrategia RAG aplicada sobre LLMs permite garantizar un balance entre precisión, trazabilidad y flexibilidad. Aunque aún se identifican limitaciones asociadas a redundancias y sensibilidad a consultas ambiguas, el desempeño alcanzado sitúa al sistema como una solución viable para entornos universitarios.

Los valores de precisión superiores al 80 %, junto con la cobertura cercana al 90 %, reflejan un nivel de rendimiento adecuado para fases iniciales de despliegue. A ello se suma la baja latencia, que asegura una experiencia conversacional fluida y competitiva respecto a estándares de referencia en sistemas similares.

5. Discusión

Los resultados obtenidos muestran que la integración de grandes modelos de lenguaje (LLMs) con esquemas de *Retrieval-Augmented Generation* (RAG) puede ofrecer, en contextos universitarios, un equilibrio favorable entre precisión, cobertura y trazabilidad normativa. En el escenario de *chat testing*, el sistema alcanzó niveles de precisión superiores al 80 % y una cobertura cercana al 90 %, con latencia media por debajo del umbral de siete segundos. Este perfil de desempeño es coherente con la evidencia previa sobre la utilidad de la recuperación condicionada en dominios regulados, donde las respuestas requieren anclaje explícito a fuentes verificables (Izacard y Grave, 2020; P. Lewis et al., 2020). En comparación con aproximaciones exclusivamente generativas, los resultados sugieren que la externalización del conocimiento dinámico hacia un índice semántico no solo reduce la incidencia de alucinaciones, sino que posibilita actualizaciones incrementales del repositorio documental sin reentrenar el modelo base, una ventaja operativa relevante para instituciones con recursos moderados (Wollny et al., 2021). En relación con experiencias documentadas en el marco teórico, los resultados alcanzados presentan una coherencia razonable con antecedentes de implementación en contextos universitarios. La cobertura del 91,7 % y la tasa de resolución en primer turno (FTR) del 76,3 % se sitúan en rangos comparables a los reportados por sistemas como *Genie* en Deakin University y *Pounce* en Georgia State, ambos caracterizados por una alta absorción de demandas estudiantiles y una reducción en la derivación hacia operadores humanos. Del mismo modo, la precisión del 82,4 % muestra consonancia con iniciativas como *Unimib Assistant* y *EduBot*, orientadas a garantizar trazabilidad normativa y a disminuir los volúmenes de tickets administrativos. Cabe señalar, sin embargo, que estas comparaciones deben considerarse con cautela metodológica, dado que las métricas disponibles en los casos referidos suelen reflejar indicadores operativos a gran escala, mientras que las aquí reportadas derivan de un escenario controlado de validación prototípica.

Desde una perspectiva de validez institucional, la trazabilidad de las respuestas emerge como un atributo central. A diferencia de escenarios comerciales en los que la naturalidad conversacional suele primar, en el entorno universitario las respuestas deben ser defendibles frente a la normativa vigente y auditables en su origen (Hasal et al., 2021). El diseño evaluado—basado en ingesta sistemática de documentos oficiales y *web scraping* de portales públicos con posterior normalización, *chunking* y almacenamiento vectorial—permitió recuperar evidencia con metadatos suficientes para justificar cada salida. Este comportamiento es consistente con observaciones

en implementaciones reales de chatbots académicos, donde la confianza del usuario depende de la citación explícita de las fuentes y del alineamiento con documentos institucionales (Klopfenstein et al., 2017; C. Rodrigues et al., 2022). Además, la latencia media registrada se sitúa en el rango que preserva la fluidez conversacional recomendada por la literatura en interacción humano–máquina.

Metodológicamente, la adopción de CRISP–DM como andamiaje del ciclo de vida resultó efectiva para estructurar la resolución del problema desde la comprensión de negocio hasta el despliegue, reforzando la reproducibilidad y la inspeccionabilidad del proceso. La modularización en dos artefactos—*indexación/embeddings* y *conversación con RAG*—facilitó la isolación de fallos y la iteración rápida sobre los puntos de mayor sensibilidad (ingesta, normalización y recuperación). En el plano de la ingeniería de sistemas, el uso de `LangChain` como capa de orquestación permitió implementar razonamiento reactivo (`ReAct`) para decidir cuándo responder con conocimiento interno del modelo y cuándo invocar recuperación documental; esto se tradujo en un mejor control del *trade-off* entre fluidez y verificabilidad, y converge con tendencias recientes en plataformas de IA generativa composicional (Tunstall, Von Werra y Wolf, 2022). La evaluación del retriever con `similarity_search_with_score` y criterios de filtrado de relevancia contribuyó a estabilizar el balance entre precisión y cobertura en presencia de documentos parcialmente redundantes, una fuente conocida de ruido en sistemas RAG (Izacard y Grave, 2020).

Al confrontar estos hallazgos con el estado del arte, se observa consonancia con reportes que atribuyen a RAG la capacidad de mantener la vigencia de respuestas en dominios de actualización frecuente (calendarios, convocatorias, procedimientos) sin incurrir en los costes computacionales y riesgos de *catastrophic forgetting* asociados al *fine-tuning* reiterado de LLMs (Parisi et al., 2019). En el plano aplicado, experiencias documentadas en educación superior han mostrado impactos institucionales significativos—reducción de carga operativa y mejoras en retención/atención—cuando la automatización conversacional se ancla a fuentes oficiales y a métricas de desempeño verificables (Akiba y Fraboni, 2023; Wollny et al., 2021). Los resultados aquí reportados son consistentes con esa trayectoria y refuerzan la importancia de la gobernanza documental y de la ingeniería de datos previa al modelado.

No obstante, persisten limitaciones estructurales que enmarcan la interpretación de los resultados. En primer término, el rendimiento global del sistema está acotado por la calidad de la recuperación; errores de indexación, normalización insuficiente o sesgos en los embeddings

pueden propagar ruido hacia la fase generativa (*garbage in, garbage out*). Aun cuando el uso de umbrales de similitud y metadatos personalizados mitigó parcialmente este riesgo, se detectaron respuestas con redundancias cuando múltiples fragmentos competían por relevancia semántica. Este patrón sugiere explorar técnicas de fusión de evidencias con *re-ranking* neural, desduplicación semántica y sinergias con *long-context attention* para amortiguar conflictos entre pasajes (Izacard y Grave, 2020). En segundo lugar, la evaluación se centró en un *benchmark* de preguntas frecuentes; aunque adecuado para medir precisión, cobertura y latencia, no captura dimensiones como robustez frente a consultas adversariales, deriva semántica en conversaciones prolongadas ni efectos de *prompt drift*. Tampoco se calcularon en esta fase métricas avanzadas como *recall@k*, MRR, tasa de error de NLU o *hallucination rate*, cuya incorporación se declara como línea inmediata de trabajo. La literatura sugiere que estas métricas complementarias pueden aportar una visión más granular del desempeño del sistema, especialmente en dominios regulados donde la fidelidad y la precisión contextual son críticas (Gao, Galley y L. Li, 2019; McTear, 2022).

Las implicaciones prácticas son claras. En términos operativos, el desacoplamiento entre modelo y conocimiento permite ciclos de actualización frecuentes del repositorio documental con coste marginal, una propiedad atractiva para oficinas académicas con calendarios dinámicos. En términos de sostenibilidad, la estrategia reduce dependencia de *fine-tuning* intensivo y favorece enfoques de adaptación eficiente en parámetros (PEFT/LoRA) cuando se requiera especialización adicional, alineado con las leyes de escalado que aconsejan optimizar datos, parámetros e inferencia de manera conjunta (Hu et al., 2022; J. Kaplan et al., 2020). Finalmente, en términos de adopción institucional, la provisión de respuestas con citas verificables constituye un vector de confianza clave para estudiantes y personal docente, y encaja con recomendaciones de transparencia y auditabilidad en servicios de IA en educación superior (C. Rodrigues et al., 2022; Wollny et al., 2021).

En síntesis, los resultados respaldan la tesis de que RAG, orquestado mediante `LangChain` sobre una base vectorial curada, es una ruta técnica viable para agentes conversacionales universitarios cuando la exigencia central es responder con precisión y con trazabilidad normativa. El desempeño observado—precisión >80 %, alta cobertura y baja latencia—es competitivo para una primera fase de despliegue, aunque condicionado por la calidad del pipeline de recuperación y por la necesidad de ampliación del marco evaluativo. El siguiente salto cualitativo pasa por robustecer la capa de recuperación (relevancia y desambiguación), incorporar criterios de

calibración y *faithfulness* en la generación, y abordar la integración segura con los sistemas de la universidad. Con ello, el agente no solo escalará en alcance funcional, sino que también consolidará su valor institucional como interfaz confiable entre la comunidad académica y la infraestructura de información.

6. Conclusiones

Este trabajo presenta una arquitectura de referencia, verificable y reproducible, para agentes conversacionales universitarios basada en la sinergia entre modelos de lenguaje de gran escala y recuperación aumentada por información (RAG). La propuesta consolida tres principios de diseño que se sostienen empíricamente a lo largo del estudio: (i) el desacoplamiento entre modelo y conocimiento como mecanismo de sostenibilidad técnica y operativa; (ii) la evidencia verificable como criterio rector de generación —mediante recuperación semántica con metadatos y citación explícita—, indispensable en dominios regulados; y (iii) la orquestación modular que habilita inspeccionabilidad y control fino del comportamiento del sistema a lo largo del ciclo de vida institucional. En conjunto, estos elementos permiten ofrecer respuestas defendibles y auditables sin sacrificar fluidez conversacional, alineando la solución con las exigencias de transparencia y gobernanza propias de la educación superior (Hasal et al., 2021; C. Rodrigues et al., 2022).

En relación con los objetivos establecidos, los hallazgos permiten trazar un vínculo claro entre los logros del prototipo y las metas iniciales del estudio. Respecto al objetivo 1 —diseñar una arquitectura modular y escalable— se logró implementar una solución técnica desacoplada entre indexación, recuperación y generación, facilitando tanto la extensibilidad como la interoperabilidad con sistemas institucionales. En cuanto al objetivo 2 —evaluar un LLM de última generación bajo estrategia RAG— se demostró que GPT-4o-mini puede alcanzar precisión superior al 80 % con baja latencia, apoyado en un motor semántico optimizado con metadatos y umbrales de similitud. En relación con el objetivo 3 —desarrollar una interfaz web conversacional— se consolidó un entorno multiplataforma con visualización de respuestas trazables, manteniendo la transparencia normativa como criterio rector. Finalmente, respecto al objetivo 4 —validar empíricamente el sistema mediante *chat testing*—, los resultados obtenidos con un benchmark de preguntas frecuentes respaldan la viabilidad funcional del agente en escenarios institucionales.

Desde una perspectiva institucional, los hallazgos refrendan que el valor del agente no reside únicamente en la capacidad generativa del LLM, sino en la calidad del andamiaje documental que lo sustenta. La curación, normalización y trazabilidad de las fuentes oficialistas —integradas a través de índices semánticos— constituyen el factor determinante de la fidelidad informativa, reduciendo la propensión a respuestas no fundamentadas y posibilitando actualizaciones incrementales sin recurrir a reentrenamientos costosos (Izacard y Grave, 2020; P. Lewis et al., 2020; Parisi et al., 2019). Así, la arquitectura propuesta se posiciona como una solución pragmática

para instituciones con calendarios y normativas cambiantes, al tiempo que facilita la adopción responsable de IA en servicios críticos para la comunidad académica (Wollny et al., 2021).

En términos de contribución, el estudio aporta una ruta técnica clara para pasar de repositorios documentales heterogéneos a una interfaz conversacional con trazabilidad normativa, demostrando que la combinación de recuperación con umbrales de relevancia y generación condicionada constituye un compromiso eficaz entre precisión, cobertura y control. Más allá del caso analizado, el resultado generalizable es conceptual: en dominios de alta exigencia regulatoria, la ingeniería de datos y la gobernanza de fuentes son palancas más decisivas que el mero escalado del modelo para obtener respuestas útiles, fiables y socialmente aceptables en el contexto universitario. Esta conclusión sitúa a RAG no como un complemento accesorio, sino como el mecanismo estructural que permite que los LLMs operen con legitimidad institucional.

Desde una perspectiva de impacto global, el sistema propuesto contribuye directamente al cumplimiento de varios Objetivos de Desarrollo Sostenible (ODS). En particular, respalda el ODS 4 (Educación de calidad) al garantizar un acceso equitativo y continuo a la información académica, reduciendo brechas de conocimiento y fortaleciendo la autonomía del estudiante; el ODS 9 (Industria, innovación e infraestructura) mediante la digitalización de procesos universitarios rutinarios, promoviendo eficiencia institucional; y el ODS 10 (Reducción de las desigualdades) al ofrecer un canal de consulta uniforme y sin sesgos, independientemente de las condiciones socioeconómicas, ubicación geográfica o habilidades digitales del usuario. La arquitectura propuesta, por tanto, no solo responde a criterios técnicos, sino que se alinea con principios de equidad, sostenibilidad y transformación digital inclusiva en el marco de la Agenda 2030.

7. Limitaciones y futuras líneas de investigación

La validez externa de los resultados se encuentra acotada por el diseño de evaluación empleado. El desempeño se midió en un entorno controlado de *chat testing* con un *benchmark* de preguntas frecuentes, lo que permitió estimar precisión, cobertura y latencia, pero no capturar fenómenos propios del uso real, como deriva semántica en conversaciones prolongadas, reformulaciones sucesivas o efectos de carga en períodos pico académicos.

Como siguiente paso metodológico, se propone realizar una prueba controlada en condiciones reales de uso mediante un experimento A/B. Este consistirá en comparar el rendimiento del agente conversacional frente a canales tradicionales (correo electrónico, atención presencial, formularios institucionales) durante un proceso de alta demanda como la matrícula. El experimento incluirá una muestra de 200 estudiantes —asumiendo una población matriculada entre 2.000 y 4.000 usuarios activos, lo que representa entre el 5 % y el 10 %—, junto con 10 empleados administrativos encargados de tareas recurrentes de atención e información académica —estimando una población total de entre 50 y 80 agentes institucionales de primera línea—. Esta configuración permite capturar indicadores comparables de carga administrativa, *First Turn Resolution* (FTR), tiempos de espera y satisfacción percibida, aplicando principios de diseño cuasi–experimental comúnmente empleados en entornos universitarios. La asignación aleatoria por grupos y el uso de métricas paralelas facilitarán evaluar no solo la precisión técnica del sistema, sino su impacto operativo en el ecosistema institucional. Esta aproximación no solo permitirá validar empíricamente la utilidad del sistema, sino que alineará la evaluación con estándares metodológicos aceptados en la investigación aplicada en educación superior (McTear, 2022; Wollny et al., 2021).

En paralelo, la ampliación del marco métricos hacia indicadores de recuperación —*recall@k*, *MRR*—, calibración y tasa de alucinación condicionada a evidencia fortalecería la robustez inferencial de los hallazgos (Gao, Galley y L. Li, 2019; Izacard y Grave, 2020). Como paso siguiente, se proyecta una validación con usuarios reales durante el proceso de matrícula institucional: una prueba A/B con 200 estudiantes que permitirá comparar la carga administrativa, el *First Turn Resolution* y la satisfacción percibida frente a canales tradicionales, siguiendo metodologías experimentales empleadas en contextos educativos.

Un segundo límite estructural reside en la sensibilidad del sistema a la calidad del pipeline de datos y, en particular, al módulo de recuperación. Aunque la estrategia de *chunking* y el filtrado con `similarity_search_with_score` mitigaron ruido y mejoraron la pertinencia, se obser-

varon redundancias cuando múltiples fragmentos competían por relevancia. Esta dependencia confirma que el cuello de botella de los sistemas RAG suele desplazarse a la fase de recuperación y fusión de evidencias (Izacard y Grave, 2020). Como línea futura, resulta pertinente evaluar técnicas de *re-ranking* neural a nivel de pasaje, desduplicación semántica y fusión con atención de contexto extendido, así como comparar variantes de embeddings y esquemas de indexación para reducir ambigüedad y colisión de fragmentos (P. Lewis et al., 2020). La integración de políticas de selección basadas en umbrales dinámicos y señales de confianza calibradas también podría contribuir a disminuir respuestas sobre–condicionadas o, en el extremo opuesto, generadas con evidencia insuficiente (McTear, 2022).

Desde la perspectiva institucional, el estudio se circunscribió a fuentes públicas y documentos oficiales sin conexión transaccional con sistemas críticos (LMS, SIS, ERP). Esta decisión favoreció la trazabilidad normativa, pero limita conclusiones sobre seguridad, autenticación federada y gobierno de datos en escenarios de producción. La investigación futura debería abordar integraciones seguras con mecanismos SSO/OAuth2, control granular de accesos, auditoría de interacciones y cumplimiento regulatorio, en línea con marcos de protección de datos aplicables a la educación superior (Hasal et al., 2021; C. Rodrigues et al., 2022; Tran et al., 2025). Asimismo, conviene explorar protocolos de registro y anonimización que permitan analizar comportamiento de usuarios sin comprometer identidades, manteniendo la transparencia exigida en servicios de IA de alto impacto (Wollny et al., 2021).

Otra limitación relevante deriva de la dependencia de modelos generales de lenguaje y de sus sesgos inherentes. Aunque la externalización del conocimiento dinámico a un índice semántico reduce la necesidad de reentrenamiento continuo —y con ello los riesgos de *catastrophic forgetting*—, persisten tensiones entre fluidez generativa y fidelidad a la fuente (Parisi et al., 2019). En esta dirección, futuras líneas incluyen evaluar adaptación eficiente en parámetros (PEFT/LoRA) para dominios universitarios específicos, con el fin de mejorar consistencia estilística y control del registro académico sin incurrir en costes elevados, y estudiar estrategias de alineación con retroalimentación humana orientadas a criterios de transparencia, formalidad y *faithfulness* (Hu et al., 2022; Ouyang et al., 2022). La comparación sistemática entre rutas de especialización (RAG puro vs. RAG+*fine-tuning* eficiente) permitiría delimitar mejor el punto de equilibrio entre coste, rendimiento y gobernanza documental.

En términos de escalabilidad y sostenibilidad, la arquitectura probada evidenció latencias compatibles con diálogo natural; no obstante, no se caracterizó el comportamiento bajo cargas

concurrentes elevadas ni se realizó un *stress testing* sistemático para simular picos de demanda. Adicionalmente, la ausencia de mecanismos de autenticación como OAuth2 y de validación transaccional limita su aplicabilidad en contextos donde la seguridad del usuario y la integridad operativa son prioritarias. Futuros trabajos deberían incorporar perfiles de estrés, técnicas de cacheo semántico y estrategias de optimización de inferencia, así como analizar el impacto de las leyes de escalado para ajustar conjuntamente datos, parámetros y presupuesto computacional (J. Kaplan et al., 2020). Este análisis es clave para una adopción sostenible en instituciones con restricciones presupuestarias y picos de demanda predecibles.

Por último, el alcance multimodal se mantuvo en un plano documental–textual. Dado que la información universitaria incluye material gráfico y audiovisual, resulta prometedor investigar extensiones multimodales que integren texto, imagen y audio de forma coherente, habilitando casos de uso como interpretación de documentos escaneados, materiales docentes y señalética administrativa, tal como sugiere la literatura de modelos multimodales y preentrenamiento contrastivo (Baltrušaitis, Ahuja y Morency, 2018; Radford, Kim et al., 2021; Tsimpoukelli et al., 2021). Complementariamente, la generalización inter–institucional —transferir la arquitectura a universidades con normativas y lenguas distintas— abre un frente de investigación sobre adaptación de dominios, cobertura léxica y equidad de desempeño entre cohortes estudiantiles (C. Rodrigues et al., 2022; Wollny et al., 2021).

En suma, las limitaciones identificadas no invalidan la tesis central, pero sí acotan su alcance a un escenario técnicamente controlado y normativamente verificado. Superarlas demanda avanzar en tres frentes convergentes: (i) evaluación en condiciones reales con métricas ampliadas y mecanismos de calibración; (ii) fortalecimiento del subsistema de recuperación y de la gobernanza documental para sostener *faithfulness* a escala; y (iii) integración segura con ecosistemas institucionales y extensiones multimodales que acerquen el agente a procesos académicos y administrativos de mayor complejidad. Abordar estos ejes consolidará el tránsito desde una prueba de concepto robusta hacia una infraestructura conversacional con legitimidad operativa y regulatoria en educación superior (Hasal et al., 2021; Izacard y Grave, 2020; P. Lewis et al., 2020; C. Rodrigues et al., 2022).

Finalmente, entre las líneas de investigación previamente identificadas, una vía prometedora para reforzar la eficiencia y la escalabilidad de los agentes conversacionales universitarios es la incorporación de Small Language Models (SLMs). De acuerdo con NVIDIA (2025), estos modelos compactos —con menos parámetros que los LLM tradicionales— pueden ofrecer un

rendimiento competitivo en tareas de razonamiento, seguimiento de instrucciones y generación de contenido, al tiempo que reducen de forma sustancial los requisitos computacionales y los costes de despliegue. 1.Su adopción en entornos educativos posibilitaría agentes más ligeros y eficientes, ejecutables en infraestructuras locales con latencias inferiores, favoreciendo la sostenibilidad y la respuesta en tiempo real ante picos de consultas de estudiantes y personal. Esta línea permite avanzar hacia sistemas de IA conversacional más autónomos, adaptativos y accesibles, sin sacrificar la precisión, la consistencia ni la gobernanza documental ya establecida (Belcak et al., 2025).

Referencias

Akbari, Hassan et al. (2021). «Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text». En: *Advances in neural information processing systems* 34, págs. 24206-24221.

Akiba, Daisuke y Michelle C Fraboni (2023). «AI-supported academic advising: Exploring ChatGPT's current state and future potential toward student empowerment». En: *Education Sciences* 13.9, pág. 885.

Albrecht, Stefano V, Filippos Christianos y Lukas Schäfer (2024). *Multi-agent reinforcement learning: Foundations and modern approaches*. MIT Press.

Alonso, Nick et al. (2024). «Toward conversational agents with context and time sensitive long-term memory». En: *arXiv preprint arXiv:2406.00057*.

Baltrušaitis, Tadas, Chaitanya Ahuja y Louis-Philippe Morency (2018). «Multimodal machine learning: A survey and taxonomy». En: *IEEE transactions on pattern analysis and machine intelligence* 41.2, págs. 423-443.

Basha, M John et al. (2023). «Advancements in natural language processing for text understanding». En: *E3S web of conferences*. Vol. 399. EDP Sciences, pág. 04031.

Belcak, Peter et al. (2025). «Small Language Models are the Future of Agentic AI». En: *arXiv preprint arXiv:2506.02153*.

Bird, Jordan J, Anikó Ekárt y Diego R Faria (2023). «Chatbot Interaction with Artificial Intelligence: human data augmentation with T5 and language transformer ensemble for text classification». En: *Journal of Ambient Intelligence and Humanized Computing* 14.4, págs. 3129-3144.

Braun, Daniel et al. (2017). «Evaluating natural language understanding services for conversational question answering systems». En: *Proceedings of the 18th annual SIGdial meeting on discourse and dialogue*, págs. 174-185.

Brown, Tom et al. (2020). «Language models are few-shot learners». En: *Advances in neural information processing systems* 33, págs. 1877-1901.

Chen, Zheng et al. (2024). «Facilitating university admission using a chatbot based on large language models with retrieval-augmented generation». En: *Educational Technology & Society* 27.4, págs. 454-470.

Collby, KM (1975). *Artificial paranoia: a computer simulation of paranoid process*.

Dawood, Manal (2024). «Assessing the effectiveness of Chatbots in providing personalized academic advising and support to higher education students: A narrative literature review». En: *Studies in Technology Enhanced Learning* 4.1.

Deakin University (2019). *Genie: Your digital study assistant*. Accedido: 7 de septiembre de 2025. URL: https://www.oclc.org/content/dam/research/presentations/2019/Deakin-Genie.pdf.

Devlin, Jacob et al. (2019). «Bert: Pre-training of deep bidirectional transformers for language understanding». En: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, págs. 4171-4186.

Drake, Jayne K (2011). «The role of academic advising in student retention and persistence». En: *About Campus* 16.3, págs. 8-12.

Følstad, Asbjørn, Theo Araujo, Effie Lai-Chong Law et al. (2021). «Future directions for chatbot research: an interdisciplinary research agenda». En: *Computing* 103.12, págs. 2915-2942.

Følstad, Asbjørn, Theo Araujo, Symeon Papadopoulos et al. (2020). *Chatbot research and design*. Springer.

Gao, Jianfeng, Michel Galley y Lihong Li (2019). *Neural approaches to conversational AI: Question answering, task-oriented dialogues and social chatbots*. Now Foundations y Trends.

Al-Ghonmein, Ali M, Khaldun G Al-Moghrabi y Tawfiq Alrawashdeh (2023). «Students' satisfaction with the service quality of academic advising systems». En: *Indonesian Journal of Electrical Engineering and Computer Science* 30.3, págs. 1838-1845.

Hasal, Martin et al. (2021). «Chatbots: Security, privacy, data protection, and social aspects». En: *Concurrency and Computation: Practice and Experience* 33.19, e6426.

Houlsby, Neil et al. (2019). «Parameter-efficient transfer learning for NLP». En: *International conference on machine learning*. PMLR, págs. 2790-2799.

Howard, Jeremy y Sebastian Ruder (2018). «Universal language model fine-tuning for text classification». En: *arXiv preprint arXiv:1801.06146*.

Hristea, Florentina T (2025). «Statistical natural language processing». En: *International Encyclopedia of Statistical Science*. Springer, págs. 2550-2553.

Hu, Edward J et al. (2022). «Lora: Low-rank adaptation of large language models.» En: *ICLR* 1.2, pág. 3.

Izacard, Gautier y Edouard Grave (2020). «Leveraging passage retrieval with generative models for open domain question answering». En: *arXiv preprint arXiv:2007.01282*.

Al-Jaf, Kanaan et al. (2024). «Leveraging chatbots for effective educational administration: a systematic review». En.

Kanwal, Sameera y Muhammad Shoaib Farooq (2021). «Reinforcement Learning for Dialogue Generation: A Systematic Literature Review». En: *2021 International Conference on Innovative Computing (ICIC)*. IEEE, págs. 1-10.

Kaplan, Jared et al. (2020). «Scaling laws for neural language models». En: *arXiv preprint arXiv:2001.08361*.

Klopfenstein, Lorenz Cuno et al. (2017). «The rise of bots: A survey of conversational interfaces, patterns, and paradigms». En: *Proceedings of the 2017 conference on designing interactive systems*, págs. 555-565.

Laranjo, Liliana et al. (2018). «Conversational agents in healthcare: a systematic review». En: *Journal of the American Medical Informatics Association* 25.9, págs. 1248-1258.

Lewis, Patrick et al. (2020). «Retrieval-augmented generation for knowledge-intensive nlp tasks». En: *Advances in neural information processing systems* 33, págs. 9459-9474.

Mainstay (2020). *How Georgia State University supports every student with personalized text messaging*. Accedido: 7 de septiembre de 2025. URL: https://mainstay.com/case-study/how-georgia-state-university-supports-every-student-with-personalized-text-messaging/.

Majorana, Cristina Doritta Brandão et al. (2022). «Enhancing administrative efficiency in higher education with AI: a chatbot solution». En: *Review of Artificial Intelligence in Education* 3, e23-e23.

McLean, Scott et al. (2023). «The risks associated with Artificial General Intelligence: A systematic review». En: *Journal of Experimental & Theoretical Artificial Intelligence* 35.5, págs. 649-663.

McTear, Michael (2022). *Conversational ai: Dialogue systems, conversational agents, and chatbots*. Springer Nature.

Mohamed Hashim, Mohamed Ashmel, Issam Tlemsani y Robin Matthews (2022). «Higher education strategy in digital transformation». En: *Education and information technologies* 27.3, págs. 3171-3195.

Neupane, Subash et al. (2024). «From questions to insightful answers: Building an informed chatbot for university resources». En: *2024 IEEE Frontiers in Education Conference (FIE)*. IEEE, págs. 1-9.

Nguyen, Long ST y Tho T Quan (2024). «URAG: Implementing a Unified Hybrid RAG for Precise Answers in University Admission Chatbots–A Case Study at HCMUT». En: *International Symposium on Information and Communication Technology*. Springer, págs. 82-93.

Ouyang, Long et al. (2022). «Training language models to follow instructions with human feedback». En: *Advances in neural information processing systems* 35, págs. 27730-27744.

Parisi, German I et al. (2019). «Continual lifelong learning with neural networks: A review». En: *Neural networks* 113, págs. 54-71.

Preetha, KG et al. (2023). «Interactive Chatbot with AI Support for Universities: Enhancing Student Engagement and Administrative Efficiency». En: *International Conference on Intelligent Systems Design and Applications*. Springer, págs. 284-292.

Radford, Alec, Jong Wook Kim et al. (2021). «Learning transferable visual models from natural language supervision». En: *International conference on machine learning*. PmLR, págs. 8748-8763.

Radford, Alec, Karthik Narasimhan et al. (2018). «Improving language understanding by generative pre-training». En.

Radziwill, Nicole M y Morgan C Benton (2017). «Evaluating quality of chatbots and intelligent conversational agents». En: *arXiv preprint arXiv:1704.04579*.

Raffel, Colin et al. (2020). «Exploring the limits of transfer learning with a unified text-to-text transformer». En: *Journal of machine learning research* 21.140, págs. 1-67.

Rodrigues, Ana Luísa y Carolina Pereira (2024). «Artificial Intelligence Chatbots in Higher Education: Challenges for an Active and Transformative Learning Among Young Adults». En: *Digital Transformation in Higher Education, Part A: Best Practices and Challenges*. Emerald Publishing Limited, págs. 45-64.

Rodrigues, Carlos et al. (2022). «A review of conversational agents in education». En: *International Conference on Technology and Innovation in Learning, Teaching and Education*. Springer, págs. 461-467.

Sarmiento, C. (2025). «Investigating Flavors of RAG for Applications in College Administration». En: *Proceedings of the International Conference on Computer Supported Education (CSEDU)*, págs. 422-430.

Serban, Iulian Vlad et al. (2015). «A survey of available corpora for building data-driven dialogue systems». En: *arXiv preprint arXiv:1512.05742*.

Shawar, Bayan Abu y Eric Atwell (2007). «Chatbots: are they really useful?» En: *Journal for Language Technology and Computational Linguistics* 22.1, págs. 29-49.

Su, Pei-Hao et al. (2016). «On-line active reward learning for policy optimisation in spoken dialogue systems». En: *arXiv preprint arXiv:1605.07669*.

Sutton, Richard S, Andrew G Barto et al. (1998). *Reinforcement learning: An introduction*. Vol. 1. 1. MIT press Cambridge.

Tran, Sarah et al. (2025). «Understanding Privacy Norms Around LLM-Based Chatbots: A Contextual Integrity Perspective». En: *arXiv preprint arXiv:2508.06760*.

Tsimpoukelli, Maria et al. (2021). «Multimodal few-shot learning with frozen language models». En: *Advances in Neural Information Processing Systems* 34, págs. 200-212.

Tunstall, Lewis, Leandro Von Werra y Thomas Wolf (2022). *Natural language processing with transformers*. .ºReilly Media, Inc."

Tyukin, Georgy et al. (2024). «Attention Is All You Need But You Don't Need All Of It For Inference of Large Language Models». En: *arXiv preprint arXiv:2407.15516*.

Villegas-Ch, William et al. (2021). «Implementation of a virtual assistant for the academic management of a university with the use of artificial intelligence». En: *Future Internet* 13.4, pág. 97.

Vinyals, Oriol y Quoc Le (2015). «A neural conversational model». En: *arXiv preprint arXiv:1506.05869*.

Weizenbaum, Joseph (1966). «ELIZA—a computer program for the study of natural language communication between man and machine». En: *Communications of the ACM* 9.1, págs. 36-45.

Wollny, Sebastian et al. (2021). «Are we there yet?-a systematic literature review on chatbots in education». En: *Frontiers in artificial intelligence* 4, pág. 654924.

Wooldridge, Michael (2009). *An introduction to multiagent systems*. John wiley & sons.

Yadegaridehkordi, Elaheh et al. (2019). «Affective computing in education: A systematic review and future research». En: *Computers & education* 142, pág. 103649.

Young, Steve et al. (2013). «Pomdp-based statistical spoken dialog systems: A review». En: *Proceedings of the IEEE* 101.5, págs. 1160-1179.

Yu, Dong et al. (2011). «Introduction to the special section on deep learning for speech and language processing». En: *IEEE Transactions on Audio, Speech, and Language Processing* 20.1, págs. 4-6.

Índice de figuras

1. Adaptación de la metodología CRISP-DM al desarrollo del agente conversacional universitario. 33
2. Arquitectura técnica del agente conversacional universitario. 38
3. Ejemplos de interacción en el *chat testing*: el agente cita normativa oficial en respuestas sobre trámites. 41

Índice de tablas

1. Métricas de evaluación de chatbots en educación superior. 10
2. Componentes funcionales de un agente conversacional universitario 11
3. Comparación entre chatbots, asistentes virtuales y agentes de IA. 16
4. Casos reales de chatbots en educación superior: aplicaciones y resultados cuantitativos 30
5. Resultados de la evaluación del agente conversacional con escenario de chat testing. 39

Acrónimos

TFM: Trabajo Fin de Máster

IA: Inteligencia Artificial

CNN: Redes Neuronales Convolucionales (Convolutional Neural Networks)

LSTM: Long Short-Term Memory (Memoria a Largo Plazo)

NLU: Natural Language Understanding

SVM: Máquinas de Vectores de Soporte (Support Vector Machines)