Evaluación empírica de un chatbot generativo en servicio público

Empirical evaluation of a generative chatbot in public service

Nombre Apellidos1,[1] Nombre Apellidos2[2]
[1] Universidad o lugar de trabajo de autor 1
[2] Universidad o lugar de trabajo de autor 2
Información de contacto autores (correo, web…)

Resumen: La irrupción de la inteligencia artificial generativa ha simplificado el diseño de sistemas conversacionales y ha mejorado la interacción en chatbots, lo que ha favorecido su rápida adopción en servicios de atención al usuario, también en el ámbito de la administración pública. Sin embargo, existen aún pocos estudios empíricos que analicen de forma sistemática la eficacia de estos sistemas en contextos reales de servicio. Este trabajo tiene como objetivo evaluar el impacto de un chatbot basado en IA generativa en la percepción de calidad del servicio en un contexto de atención ciudadana. Para ello, se desarrolló un chatbot de asistencia al proceso de acceso y matrícula universitaria utilizando la plataforma Microsoft Copilot Studio, que integra modelos de lenguaje de gran escala de OpenAI. El sistema se diseñó con un dominio acotado y apoyado en mecanismos de generación aumentada por recuperación (RAG) para minimizar respuestas incorrectas. La evaluación de su desempeño se realizó combinando métricas automáticas y valoraciones humanas de la calidad de las respuestas. Los resultados muestran que, aunque el sistema proporciona respuestas correctas en la mayoría de los casos, las respuestas son erróneas en un pequeño porcentaje de interacciones, así como limitaciones relacionadas con la latencia de respuesta y el consumo de tokens. Estos hallazgos permiten identificar oportunidades de mejora y aportan evidencia empírica relevante para el desarrollo y despliegue responsable de chatbots basados en IA generativa en servicios públicos.
Palabras clave: inteligencia artificial generativa; chatbots; administración pública; evaluación de la calidad del servicio.

Abstract: The emergence of generative artificial intelligence has simplified the development of conversational systems and improved interaction in chatbots, fostering their rapid adoption in user support services, including in the public administration sector. However, there are still relatively few empirical studies that systematically evaluate the effectiveness of these systems in real service contexts. This study aims to assess the impact of a generative AI–based chatbot on perceived service quality in a public-facing support scenario. To this end, a chatbot designed to assist students during the university admission and enrollment process was developed using the Microsoft Copilot Studio platform, which integrates large language models provided by OpenAI. The system was designed with a constrained domain and supported by retrieval-augmented generation (RAG) mechanisms to reduce incorrect responses. Its performance was evaluated through a combination of automated assessment methods and human-based ratings of response quality. The results show that although the system provides correct answers in most interactions, hallucinations still occur in a small percentage of cases. Additional limitations related to response latency and token consumption were also observed. These findings provide insights for future system improvements and highlight important considerations for the responsible deployment of generative AI chatbots in public service environments.
Keywords: generative AI; chatbots; public service; service quality evaluation.

1 Introducción

Los sistemas de diálogo han constituido históricamente uno de los principales ámbitos de aplicación del Procesamiento del Lenguaje Natural, al integrar de forma natural problemas de comprensión y generación del lenguaje en contextos interactivos de uso real (Ni et al., 2023). En este marco, la irrupción reciente de la inteligencia artificial generativa se ha introducido en el desarrollo de los sistemas conversacionales, tanto desde el punto de vista tecnológico como desde la perspectiva de su diseño y despliegue (Yi et al., 2025). En particular, los modelos de lenguaje de gran escala (LLMs) han facilitado la interpretación de los turnos de usuario, han ampliado el acceso a conocimiento contenido en documentos textuales —más allá de las bases de datos estructuradas— y han incrementado la naturalidad, expresividad y variabilidad de las respuestas generadas. Como resultado, se ha ampliado significativamente el repertorio de servicios susceptibles de ser automatizados mediante interfaces conversacionales, al tiempo que se ha intensificado la adopción de chatbots en múltiples dominios de atención al usuario, incluidos los servicios públicos (Larsen et al., 2024).

Con anterioridad a la expansión de la IA generativa, la evaluación de sistemas de diálogo y chatbots constituía ya una línea de investigación consolidada, con abundantes propuestas metodológicas orientadas a medir dimensiones como la corrección, la adecuación, la satisfacción del usuario o el éxito de la tarea (Deriu et al., 2021). No obstante, la incorporación de modelos generativos introduce retos que dificultan la transferencia directa de dichos marcos de evaluación (Chang et al., 2024). Entre ellos destacan la aparición de alucinaciones y errores fácticos (Huang et al., 2025), así como nuevas restricciones operativas asociadas al coste computacional por turno y a la latencia en la generación de las respuestas. Paralelamente, el auge de la IA generativa ha impulsado nuevas estrategias de evaluación automática; sin embargo, su validez y utilidad siguen siendo discutibles cuando se aplican a sistemas conversacionales desplegados en entornos reales de servicio. En consecuencia, persiste una brecha entre el rápido despliegue de estas tecnologías y la disponibilidad de estudios empíricos que analicen de forma sistemática su rendimiento efectivo en condiciones operativas.

Este trabajo presenta la evaluación en contexto real de un chatbot basado en IA generativa desplegado en un servicio público universitario; analiza su rendimiento combinando indicadores de uso, tiempos de respuesta y evaluación de la calidad de las respuestas; compara la valoración de un evaluador experto con una evaluación automática basada en LLM, poniendo de relieve las diferencias de criterio entre ambos enfoques; y finalmente, examina los tipos de error observados y sus implicaciones para el despliegue responsable de este tipo de sistemas en contextos administrativos. El resto del artículo se organiza del modo siguiente: la sección 2 revisa el estado del arte; la sección 3 describe el servicio y el procedimiento de evaluación; la sección 4 presenta los resultados; y las secciones 5 y 6 recogen, respectivamente, la discusión y las conclusiones.

2 Estado del arte

La nueva generación de chatbots, aumentados con capacidades de inteligencia artificial generativa, requiere el desarrollo de métodos de evaluación capaces de capturar la naturaleza abierta y no determinista de este tipo de sistemas. Las métricas automáticas tradicionales, como BLEU y ROUGE, han sido ampliamente utilizadas para evaluar sistemas de generación del lenguaje natural comparando

la similitud entre la respuesta generada por el sistema y una respuesta de referencia (Lin, 2004.; Papineni et al., 2002). Sin embargo, estas métricas presentan importantes limitaciones en el contexto de los chatbots modernos, ya que se basan principalmente en la coincidencia léxica entre secuencias de palabras y no capturan adecuadamente la equivalencia semántica ni las preferencias humanas. Como resultado, una respuesta puede ser correcta y apropiada desde el punto de vista comunicativo y, aun así, recibir una baja puntuación por utilizar una formulación diferente a la respuesta de referencia (Abbasian et al., 2024; Deriu et al., 2021; Banerjee et al., 2023; Van der Lee et al., 2019).

Una aproximación alternativa consiste en evaluar el rendimiento del sistema a partir de su impacto en el comportamiento de los usuarios y en los objetivos alcanzados. En lugar de centrarse exclusivamente en la calidad lingüística de las respuestas, estas métricas analizan cómo interactúan los usuarios con el sistema y si este cumple su función dentro del servicio ofrecido (Shawar and Atwell, 2007; Peras, 2018). Entre las métricas más habituales se encuentran las métricas que miden la frecuencia y la intensidad de la interacción de los usuarios con el sistema, incluyendo indicadores como el volumen de conversaciones, el tamaño medio de las interacciones o la tasa de retención de usuarios (Venkatesh et al., 2017; Peras, 2018; Kaiser and Schulze, 2025; Przegalinska et al., 2019). Asimismo, las métricas de éxito de tareas evalúan hasta qué punto el chatbot es capaz de cumplir su propósito, por ejemplo mediante la tasa de resolución de consultas o el porcentaje de interacciones que se resuelven sin intervención humana (Peras, 2018; Kaiser and Schulze, 2025). Finalmente, están las métricas de eficiencia se centran en aspectos operativos del sistema, como el tiempo necesario para resolver una consulta o el número de turnos requeridos para alcanzar el objetivo de la interacción (Venkatesh et al., 2017; Peras, 2018). Aunque estas métricas proporcionan información valiosa sobre el funcionamiento del sistema en un contexto real, su principal limitación es que evalúan principalmente el impacto del chatbot en el comportamiento de los usuarios, y no necesariamente la calidad intrínseca de sus respuestas.

Por esta razón, la evaluación humana continúa considerándose el estándar de referencia para valorar el rendimiento de sistemas conversacionales (Papineni et al., 2002). Los evaluadores humanos son capaces de captar aspectos complejos de la comunicación, como la coherencia, la creatividad, el tono o la veracidad de las respuestas (X. Li et al.,2021). Este tipo de evaluación puede abordarse mediante metodologías cualitativas y cuantitativas. Las metodologías cualitativas buscan comprender la experiencia del usuario y las razones detrás de su percepción del sistema, utilizando herramientas como entrevistas semiestructuradas o encuestas dirigidas a los usuarios (Siddals et al. 2024; Deschenes and McMahon, 2024; Holmes et al., 2019). Por su parte, las metodologías cuantitativas suelen basarse en la evaluación sistemática de las respuestas del chatbot mediante rúbricas estructuradas. El método más habitual consiste en utilizar escalas de Likert en las que evaluadores valoran diferentes dimensiones de la respuesta, como la precisión factual, la completitud de la información, la coherencia lingüística, la relevancia para la consulta del usuario o la seguridad del contenido (Liang and Li, 2021; Finch et al., 2023; Chiu and Chung, 2025; Van der Lee et al., 2019). Para garantizar la fiabilidad de este proceso es necesario definir guías de evaluación claras, formar a los evaluadores y, cuando sea posible, utilizar múltiples revisores para medir el grado de acuerdo entre ellos (Beaver, 2022). No obstante, a pesar de su calidad, la evaluación humana presenta limitaciones importantes, principalmente relacionadas con el tiempo y el coste necesarios para analizar grandes volúmenes

de interacciones, lo que dificulta su aplicación en contextos de desarrollo rápido de sistemas basados en inteligencia artificial (Abeysinghe and Circi, 2024; Chaganty, Mussmann, and Liang, 2018; Clark et al., 2021).

En este contexto ha surgido el paradigma de LLM como juez (Gu et al. 2026; H. Lee 2024), que busca aprovechar las capacidades de los grandes modelos de lenguaje para aproximar la evaluación humana de forma más escalable. Diversos estudios han demostrado que estos modelos pueden alcanzar niveles de acuerdo superiores al 80 % con las preferencias humanas, comparables al grado de concordancia observado entre evaluadores humanos (Zheng et al., 2023; L. Zhu, Wang, and Wang, 2023; T. Zhu et al. 2025). Este enfoque puede implementarse principalmente de dos formas. La primera es la comparación por pares, en la que el modelo recibe una pregunta del usuario junto con dos respuestas alternativas y debe determinar cuál de ellas es mejor. La segunda es la evaluación de una sola respuesta, donde el modelo asigna una puntuación a una respuesta individual siguiendo una escala predefinida. Aunque ambos enfoques han demostrado resultados prometedores, el segundo suele ser más escalable al requerir únicamente una respuesta para cada evaluación (Zheng et al., 2023; H. Li et al., 2024).

A pesar de su potencial, el uso de LLM como evaluadores también presenta diversas limitaciones. Entre ellas destacan ciertos sesgos de evaluación, como la tendencia a favorecer respuestas más largas o a verse influenciados por el orden en el que se presentan las alternativas, así como otros sesgos derivados de los patrones aprendidos durante su entrenamiento (Zheng et al., 2023; Ye et al., 2024). Además, los modelos pueden mostrar dificultades en tareas de razonamiento lógico o matemático, generar evaluaciones incorrectas debido a alucinaciones o verse limitados por la fecha de corte de su conocimiento (Zheng et al., 2023; H. Li et al., 2024; Ye et al., 2024). Para mitigar estos problemas, la literatura propone diversas estrategias, como anonimizar las respuestas evaluadas, alternar el orden de las respuestas en comparaciones por pares, utilizar rúbricas detalladas en las instrucciones del modelo o complementar el proceso con revisiones humanas en casos ambiguos o complejos (Zheng et al., 2023; Kim et al., 2023; Wang et al., 2024).

En el ámbito de la administración pública, los chatbots se están incorporando progresivamente como herramientas para mejorar la atención ciudadana y facilitar el acceso a los servicios digitales (Akar et al, 2024, Arends et al 2024, Cortés-Cediel et al., 2023). Estos sistemas permiten automatizar la resolución de consultas frecuentes, ampliar la disponibilidad de los servicios y reducir la carga operativa de las organizaciones públicas. Sin embargo, la adopción de chatbots basados en inteligencia artificial generativa introduce nuevos desafíos en términos de evaluación, ya que su comportamiento abierto y no determinista dificulta la aplicación de métricas tradicionales de rendimiento. En este contexto, resulta especialmente relevante analizar no solo la precisión técnica del sistema, sino también su impacto en la experiencia y percepción de calidad del servicio por parte de los ciudadanos. Los enfoques de evaluación descritos en la literatura sobre sistemas conversacionales y modelos generativos ofrecen un marco metodológico útil para abordar este problema, aunque su aplicación en contextos de atención ciudadana en el sector público sigue siendo todavía limitada (Aoki, 2020; Chen et al., 2025).

3 Metodología

3.1 Descripción del servicio

El sistema analizado, denominado **animBot**, es un chatbot basado en inteligencia artificial generativa diseñado

para asistir a estudiantes de nuevo ingreso en procesos administrativos relacionados con el acceso y la matrícula en la Universidad de XXXX. El chatbot animBot se desplegó en producción durante los periodos de preinscripción y matrícula de del curso 2025/26, permaneciendo inactivo fuera de esas campañas para minimizar riesgos y costes operativos. El servicio estuvo integrado en el portal institucional (www.uanomim.es) mediante un widget conversacional flotante (parte inferior izquierda de figura 1) que aparece como un icono con avatar en la interfaz web y que, al activarse, despliega una ventana emergente de diálogo donde el usuario puede formular consultas en lenguaje natural.

Figura 1: Integración del servicio en la web de la Universidad.

El sistema fue desarrollado utilizando la plataforma Microsoft Copilot Studio, que integra modelos de lenguaje de gran escala de OpenAI y capacidades de generación aumentada por recuperación (RAG). La arquitectura del chatbot se basa en un mecanismo de orquestación generativa. En este enfoque, un modelo de lenguaje actúa como orquestador para analizar la intención del mensaje del usuario y determinar el tema más adecuado dentro del dominio del asistente. En función del resultado de esta identificación, el sistema puede proporcionar una respuesta predefinida o generar una respuesta basada en la recuperación de información desde un repositorio de conocimiento (ver figura 2). El repositorio de conocimiento está compuesta por 21 documentos con un total de 57079 tokens BPE (listados en tabla 2).

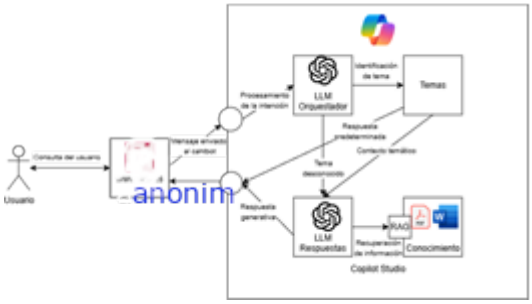

Figura 2: Descripción de la arquitectura del servicio

El dominio del sistema se organiza en ocho temas principales, centrados en procesos de acceso, preinscripción, admisión y matrícula en estudios de grado y máster, así como traslado de expediente. Estos temas utilizan respuestas generativas apoyadas en una base de conocimiento compuesta por páginas institucionales y documentos oficiales (guías, normativa y documentación administrativa). Además, el sistema incorpora doce temas secundarios con respuestas predefinidas que cubren consultas informativas complementarias (por ejemplo, becas, movilidad o trabajos fin de estudios, ver tabla 1).

3.2 Procedimiento de evaluación

La evaluación del sistema se planteó desde una doble perspectiva: por un lado, el análisis de la calidad de las respuestas generadas por el chatbot y, por otro, el estudio de ciertos aspectos de eficiencia e interacción observados durante su uso real. Para ello, se utilizaron los registros de actividad disponibles en la sección de análisis de Microsoft Copilot Studio, que permiten exportar información de uso en formato CSV.

Una limitación relevante de los registros exportados por Copilot Studio es que las respuestas del sistema aparecen truncadas a 512 caracteres. Dado que esta restricción imposibilita una evaluación fiable de las respuestas generativas, se construyó un

conjunto de evaluación específico mediante la regeneración de respuestas a partir de preguntas reales de usuarios. Para ello, se seleccionó aleatoriamente una muestra aleatoria de consultas y se reprodujo cada una de ellas de forma automatizada sobre el chatbot mediante un script en Python Selenium que envía preguntas y recoge las respuestas directamente de la interfaz web del chatbot. En las sesiones con varios turnos se consideró únicamente la primera pregunta sustantiva del usuario, excluyendo saludos u otras aperturas conversacionales no informativas. El script Selenium se emplea no sólo para guardar la pregunta y respuesta sino también para medir el tiempo de respuesta, una información no proporcionada en los logs de Copilot Studio.

La calidad de las respuestas se evaluó mediante una rúbrica con escala Likert de 1 a 5 para valorar la utilidad global de cada respuesta, donde las puntuaciones más bajas corresponden a respuestas erróneas, inventadas o críticamente incompletas, y las más altas a respuestas correctas, completas, pertinentes y claramente formuladas. La anotación manual se realizó sobre las respuestas regeneradas mediante una interfaz web diseñada específicamente para este fin. Para cada par pregunta–respuesta se registró la puntuación asignada y una breve justificación textual. Debe señalarse que esta evaluación fue llevada a cabo por el personal del Centro de Atención al Usuario (CAU) de la propia Universidad que atiende también las dudas de usuarios en el proceso de matrícula.

Junto a la evaluación humana, se realizó una segunda evaluación automática siguiendo el paradigma de LLM-as-a-judge**.** Para ello, se utilizó NotebookLM sobre el mismo conjunto de preguntas y respuestas, solicitando una puntuación y la justificación de la misma. Se proporcionó a NotebookLM los archivos de la base de conocimiento y se pide que responda sin contestar fuentes externas. Los comentarios realizados tanto por el CAU como por Notebook son revisados por los autores y clasificados analizando también la consistencia entre juicios humanos y automáticos.

Finalmente, se analizaron las respuestas clasificadas como erróneas para indagar los motivos y estudiar posibles soluciones. El análisis lleva a una clasificación heurística del feedback en las evaluaciones en categorías temáticas que permite analizar con más detalle las inconsistencias entre juicios humanos y juicios automáticos.

4 Resultado

4.1 Estadísticas de uso

El uso medio del sistema durante el periodo analizado fue de 19 conversaciones al día en el periodo analizado con un pico de 81 conversaciones día. El 91.4% de las conversaciones es corta (tres turnos o menos), siendo la mayoría de éstas (68.4%) de un único turno. El tiempo medio de generación de respuestas para atender a los usuarios es de 2.5 segundos, habiendo outliers por encima de los 10 segundos.

La plataforma Copilot Studio proporciona una estimación automática de resolución efectiva de la conversación basada en medidas internas del sistema, como la ausencia de fallos de interpretación o de activación de mecanismos de fallback. Según esta métrica, 420 sesiones fueron clasificadas como resueltas, mientras que 136 (24,46%) se consideraron no resueltas.

Con respecto a los temas tratados, la tabla 1 muestra que, aunque destaca la temática relativa a la matriculación, el sistema fue capaz de orquestar preguntas entre los distintos temas programados.

	Turnos	
Categoría	**N**	**Porc.**
Matriculación	135	13,78%
Otras consultas relacionadas con los estudios universitarios	82	8,37%
Identidad universitaria - Acceso a servicios	73	7,45%
Opciones de contacto	71	7,24%
Consultas no relacionadas con los estudios universitarios	63	6,43%
Acceso estudios universitario	58	5,92%
Calendario académico, horarios de clases	51	5,20%
PAU, notas de corte y plazas ofertadas	50	5,10%
Cancelación o modificación de matrícula	44	4,49%
Información sobre estudios de grado	43	4,39%
Saludo	42	4,29%
Matriculación – Grado	36	3,67%
Acceso estudios universitarios - Grado	34	3,47%
Información de precios	32	3,27%
Identidad universitaria - Activar o recuperar clave	32	3,27%
Convalidaciones de asignaturas/ créditos	29	2,96%

Traslado de expediente	27	2,76%
Información de grupo, cambio de grupo	24	2,45%
Identidad universitaria	15	1,53%
Acceso estudios universitarios - Máster	8	0,82%
Información sobre estudios de máster	8	0,82%
Trabajo de fin de grado (TFG), Trabajo de fin de máster (TFM)	7	0,71%
Prácticas externas	6	0,61%
Matriculación – Máster	6	0,61%
Erasmus	4	0,41%

Tabla 1: Distribución de temas orquestados

Predominan las respuestas que usan IA generativa (71,1%), especialmente en las consultas sobre matrícula, que constituyen el núcleo funcional del sistema, mientras que los saludos y las consultas ajenas al dominio universitario se resolvieron principalmente con respuestas predefinidas. La tabla 2 muestra las fuentes de conocimiento que se emplean tanto a la hora de analizar las preguntas como a la hora de formular las respuestas.

	Porcentaje de uso	
Fuente de conocimiento	**Respuestas**	**Preguntas**
Manual automatrícula máster	17,59%	23,57%
Preguntas y respuestas estudiantes de Grado	13,28%	17,79%
Matriculación	8,91%	11,95%
Acceso y preinscripción a los estudios de máster	8,14%	10,91%
Preinscripción y admisión grado	6,69%	8,96%
Matriculación estudios de máster	6,30%	8,44%
Manual automatricula 24-25 grado	6,15%	8,25%
Matrícula Máster de secundaria	5,72%	7,66%
Acceso y preinscripción Máster de secundaria	5,14%	6,88%
Documentación matricula nuevo ingreso	3,34%	4,48%
Traslado de expediente	3,20%	4,29%
Información de precios de matricula	3,00%	4,03%
RESOLUCIÓN BOCYL Num. 93	2,62%	3,51%
Documentación de matrícula estudiantes anteriores	2,08%	2,79%
Notas de corte 2024-2025	1,79%	2,40%
Información PAU	1,74%	2,34%
Guía Prueba de Acceso a la Universidad	1,45%	1,95%
Plazas nuevo ingreso grados 2025-26	1,16%	1,56%
Documentación matrícula estudiantes traslado	0,92%	1,23%
Normativa reconocimiento y transferencia de créditos	0,68%	0,91%
Reconocimiento y transferencia de créditos	0,10%	0,13%

Tabla 2 Uso de las fuentes de conocimiento en la generación de respuestas

	Experto N	Experto P	Total
NBLM N	36	17	53 (22%)
NBLM P	57	124	181(77%)
Total	93 (40%)	141 (60%)	234

Tabla 3: Resultados de las evaluaciones de las respuestas.

4.2 Evaluación de calidad

La tabla 3 muestra el número de veces que las respuestas son considerabas válidas, tanto en los juicios expertos como automáticos. También se muestra la contingencia entre las evaluaciones del experto y las evaluaciones automáticas. Se ha optado por una binarización de la escala (Negativo N si la puntuación <= 2 y Positivo P en caso contrario) porque las puntuaciones del CAU (Experto) se concentran en los extremos (1–2 en el 40% de los casos y 5 en el 41%), a la vez que NotebookLM (NBLM) tiende a utilizar también con mayor frecuencia las puntuaciones extremas (1–2 en el 23% y 5 en el 57%), empleando de forma marginal los valores intermedios. El acuerdo observado entre ambos evaluadores es del 68,4%, con un índice kappa de 0,29, lo que corresponde a un nivel de acuerdo bajo-moderado.

Respuesta Experto (N)	**Respuesta NBLM**		
Tipo	**N**	**P**	**Total**
Respuesta genérica (por desconocimiento)	3	6	9
Respuesta insuficiente	18	27	45
Pregunta ambigua	4	15	19
Respuesta incorrecta	11	9	20
Total	**36**	**57**	**93**

Tabla 4: Clasificación de tipo de errores documentados por el CAU (Experto) y juicio otorgado por el LLM

Las tablas 4 y 5 recogen el análisis de las valoraciones negativos de Experto y NBLM respectivamente. Se analiza el feedback reportado en la evaluación y se obtiene la clasificación reportada en las tablas. Para evaluar la diversidad de criterio se reporta en las columnas las puntuaciones otorgadas por el otro evaluador en los diferentes casos. Con respecto a los tipos de fallos encontrados por el CAU (tabla 4) *Respuesta genérica (pregunta fuera de dominio)* son respuestas en las que el chatbot reconoce que no tiene información para responder la pregunta usando una plantilla; *Respuesta insuficiente* significa que el evaluador aporta información adicional y/o sugiere cambios el estilo de la respuesta sin afirmar que la información aportada por el sistema sea falsa; *Pregunta ambigua* significa que el revisor considera que la pregunta está mal formulada y que el sistema debió pedir aclaración antes de contestar, la respuesta podría ser incorrecta según cómo se interprete la pregunta; *Respuesta incorrecta* significa que el chatbot contestó con información falsa a juicio del evaluador. Se constata que en 20 casos de los 234 respuestas analizadas (9%) el sistema reportó respuestas incorrectas y que en más de la mitad de dichos casos fueron también evaluados negativamente por el LLM (última fila de la tabla). El desacuerdo entre Experto y NBLM es elevado en todas las categorías. Además, en el caso de *Pregunta ambigua*, el

LLM sólo encuentra problemas en 4 de los 19 casos reportados por el CAU.

Respuesta NBLM (N)	Respuesta Experto		
Tipo	**N**	**P**	**Total**
Respuesta genérica	12	2	14
Respuesta insuficiente	8	7	15
Respuesta con información externa	1	7	8
Respuesta incorrecta	15	1	16
Total	**26**	**17**	**53**

Tabla 5: Clasificación de tipo de errores documentados por el NotebookLM (NBLM) y juicio otorgado por el CAU

En el caso de las evaluaciones automáticas (tabla 5), *Respuesta genérica* recoge comentarios que evidencian que el chatbot utilizó una plantilla genérica para contestar, principalmente porque la respuesta está fuera del dominio; *Respuesta insuficiente* implica que según el LLM considera que la respuesta puede ser completada; *Respuesta con información externa* recoge respuestas que penalizan que el chatbot esté usando información fuera de la base de conocimiento, no significa que la información reportada sea falsa; *Respuesta incorrecta* de nuevo significa que el sistema reportó información errónea. En total se observa que el sistema identifica 16 errores entre las 234 respuestas analizadas (7%), sólo uno de los cuales no fue considerado como respuesta errónea por el CAU (fila inferior de tabla 5). La consistencia en este caso es elevada a la hora de penalizar las *Respuestas incorrectas* y las *Respuestas genéricas* por humanos. En el caso de las *Respuestas con información externa,* solo en uno de los casos es valorada negativamente por el CAU.

Donde más confusión aparece es en el caso de *Respuestas insuficientes* donde no parece haber concordancia de criterio entre humanos y NBLM: tanto en la tabla 4 como en la tabla 5 se observa que un alto porcentaje de las respuestas penalizadas por uno de los evaluadores es valorada positivamente por el otro.

5 Discusión

5.1 Valoración del uso del servicio

Desde la perspectiva del desarrollo, la experiencia con Copilot Studio pone de manifiesto que la incorporación de capacidades de IA generativa simplifica la puesta en marcha de servicios conversacionales donde la configuración inicial, la orquestación temática, la incorporación de fuentes de conocimiento y la generación de respuestas pueden realizarse con un esfuerzo de implementación bajo en comparación con enfoques tradicionales. Al mismo tiempo, los mecanismos de monitorización proporcionados por la plataforma, aunque útiles para el seguimiento general del servicio, presentan limitaciones relevantes desde el punto de vista analítico. En nuestro caso, los informes de orquestación permitieron comprobar que el sistema se utilizó de forma coherente con el dominio previsto (tabla 2) y que el enrutamiento (tabla 1) hacia los temas definidos resultó, en términos generales, adecuado. Sin embargo, los logs de uso disponibles en la plataforma resultaron insuficientes para una evaluación detallada del rendimiento del sistema. Entre las carencias observadas destacan la ausencia de métricas precisas sobre consumo por transacción, especialmente en términos de tokens, la escasa granularidad sobre el uso efectivo de las unidades de conocimiento y la falta de documentación clara sobre algunos indicadores internos, como la estimación del éxito conversacional o tiempo de respuesta de cada turno de diálogo. A ello se añade que las respuestas almacenadas en los ficheros CSV aparecen truncadas, lo que dificulta su reutilización para tareas de auditoría y evaluación posterior. Aunque las plataformas de desarrollo generativo basadas en IA generativa reducen la barrera de entrada para construir chatbots funcionales, todavía presentan carencias importantes en términos de observabilidad, trazabilidad y soporte a la evaluación rigurosa del servicio.

Desde la perspectiva del usuario, los resultados muestran que el uso del servicio es significativo, especialmente si se tiene en cuenta que el proceso de matrícula universitaria se realiza mediante un sistema online consolidado desde hace años y acompañado de abundante documentación. En este contexto, la utilización de un chatbot como canal adicional de consulta sugiere la existencia de una demanda real de asistencia más directa y accesible. El análisis de las interacciones indica, además, que la mayoría de las consultas se ajustan al dominio previsto, aunque se observan también usos fuera del alcance del sistema, lo

que pone de manifiesto la tendencia de los usuarios a interpretar este tipo de interfaces como puntos de entrada generalistas a la información institucional (ver tabla 1). La estructura de las conversaciones es mayoritariamente breve, con un predominio claro de interacciones de un único turno. Este comportamiento es coherente con la naturaleza del servicio, orientado a la resolución puntual de dudas mediante acceso a información estructurada, y con el hecho de que muchas respuestas redirigen al usuario hacia recursos externos (páginas web o direcciones de contacto) donde ampliar la información. Uno de los aspectos más críticos identificados es el tiempo de respuesta del sistema. Aunque la interfaz proporciona retroalimentación visual durante la generación de la respuesta, la latencia observada en algunos casos puede resultar elevada desde la perspectiva del usuario y favorecer el abandono prematuro de la interacción.

El hecho de ser un estudio sobre un despliegue real ha dificultado la recogida de información sobre el grado de satisfacción de los usuarios. La herramienta da la posibilidad de indicar si la interacción fue satisfactoria o si no lo fue pero esta opción no fue apenas empleada por los usuarios. No se recibieron quejas durante el periodo de uso y los responsables institucionales de la Universidad manifestaron su satisfacción con el servicio, no se recibieron quejas de usuarios durante su implantación y los responsables del CAU se han interesado por volverlo a poner en marcha este nuevo curso.

5.2 Valoración de la calidad de las respuestas

En cuanto a la calidad de las respuestas, la evaluación realizada por el CAU indica que el sistema ofrece un rendimiento global positivo, con un 60% de las respuestas valoradas con puntuaciones iguales o superiores a 3 en la escala utilizada. De las respuestas con valoración menor o igual a 2, el porcentaje de respuestas clasificadas como claramente incorrectas es reducido, situándose por debajo del 10% del total según la tipología de errores identificada. Una buena parte de las respuestas penalizadas no corresponde a errores factuales, sino al uso deliberado de salvaguardas del sistema que evitan responder a consultas fuera de su dominio. Aunque estas respuestas son evaluadas negativamente desde el punto de vista de la utilidad inmediata, reflejan decisiones de diseño orientadas a limitar el riesgo de generación de información incorrecta. En segundo lugar, un número significativo de valoraciones bajas se asocia a respuestas que requieren matización o ampliación. En estos casos, las observaciones del experto resultan especialmente valiosas, ya que identifican mejoras concretas en la referencia a recursos institucionales, como la selección más adecuada de páginas web o direcciones de contacto. Este tipo de error no compromete necesariamente la corrección de la respuesta, pero sí su utilidad práctica, y pone de manifiesto la importancia de la curación fina de las fuentes de conocimiento en sistemas basados en recuperación.

Por otro lado, el análisis de las respuestas consideradas incorrectas muestra que, en la mayoría de los casos, los errores no tienen consecuencias directas para el usuario, al tratarse de respuestas parcialmente adecuadas que no incorporan todos los elementos de la consulta[1]. Sin embargo, se identifican también un número reducido de casos en los que la respuesta puede inducir a error con posibles consecuencias prácticas, especialmente en relación con procedimientos administrativos como el traslado de expediente. En estas situaciones, el sistema propone actuaciones que no se ajustan al procedimiento correcto, lo que pone de relieve el riesgo asociado a la generación de recomendaciones operativas en dominios sensibles. A partir de estos resultados, se plantea como línea de mejora la incorporación de salvaguardas adicionales que limiten la generación de este tipo de respuestas, incluso a costa de reducir la cobertura del sistema, priorizando así la minimización de riesgos sobre la completitud de la información ofrecida.

Un aspecto adicional relevante observado durante la evaluación es la diferente interpretación de las preguntas ambiguas. En varios casos, el CAU penaliza la respuesta del sistema no por su contenido en sí, sino por considerar que la pregunta del usuario no está suficientemente bien formulada. En estas situaciones, la respuesta del chatbot podría

[1] El conjunto de datos con las evaluaciones de las respuestas se hará público en acceso abierto tras la aceptación del artículo.

considerarse adecuada bajo una interpretación concreta de la consulta, pero el experto opta por valorar negativamente la interacción debido a la ambigüedad de partida. Este comportamiento pone de manifiesto la importancia de la gestión de la ambigüedad en sistemas conversacionales, así como la dificultad de establecer criterios de evaluación consistentes cuando la calidad de la respuesta depende en parte de la interpretación de la intención del usuario.

Esta cuestión se relaciona directamente con la baja concordancia observada entre la evaluación humana y la realizada mediante el enfoque LLM-as-a-judge. Incluso tras binarizar la escala de valoración, el nivel de acuerdo entre ambos evaluadores es reducido ($\kappa \approx 0{,}29$), lo que indica diferencias sustanciales en los criterios de evaluación aplicados. En conjunto, estos resultados cuestionan la fiabilidad del uso exclusivo de modelos de lenguaje como evaluadores en contextos aplicados y refuerzan la necesidad de complementar estas técnicas con evaluación humana, especialmente en dominios donde la precisión y la interpretación correcta de la intención del usuario son críticas.

6 Conclusiones

En este artículo se ha presentado la evaluación en contexto real de un agente conversacional de servicio público basado en tecnología comercial de IA generativa. Los resultados muestran que el sistema es capaz de proporcionar respuestas correctas y bien formuladas en la mayoría de las interacciones, apoyándose en las bases de conocimiento proporcionadas y respetando los mecanismos de control y las restricciones definidas en su diseño.

También se ha observado que la latencia en la generación de algunas respuestas, junto con la activación frecuente de mecanismos de control —como respuestas predefinidas para limitar el dominio o redirecciones a recursos externos— puede afectar negativamente a la percepción de calidad del servicio por parte del usuario.

Se ha constatado que, a pesar de las precauciones adoptadas para garantizar la calidad de las respuestas —como la restricción a fuentes de conocimiento institucionales y la limitación del dominio del sistema—, un porcentaje no despreciable de interacciones recibe valoraciones negativas. Sin embargo, el análisis detallado muestra que la mayoría de estas evaluaciones no se debe a alucinaciones, sino a los criterios aplicados por los evaluadores. Tanto los evaluadores humanos como los automáticos tienden a penalizar respuestas predefinidas o poco elaboradas, mientras que los evaluadores humanos, además, introducen matices basados en conocimiento adicional no explicitado en las fuentes utilizadas por el sistema.

Junto a estas valoraciones negativas, que en muchos casos pueden considerarse justificables, se han identificado también respuestas incorrectas que pueden tener consecuencias prácticas para el usuario, especialmente cuando se refieren a procedimientos administrativos. El análisis de estos casos sugiere que su origen se encuentra, principalmente, en dificultades del sistema para interpretar adecuadamente preguntas complejas —por ejemplo, aquellas que combinan múltiples intenciones— o en interpretaciones inadecuadas de la información disponible en la base de conocimiento. Este último aspecto se manifiesta, por ejemplo, en respuestas que aplican procedimientos correctos en un contexto distinto al planteado por el usuario.

Como línea de mejora futura, se plantea la revisión y refinamiento de la información incluida en la base de conocimiento, con el objetivo de evitar interpretaciones ambiguas que puedan dar lugar a valoraciones negativas o respuestas inadecuadas. Asimismo, se considera la incorporación de mecanismos de confirmación de la intención del usuario mediante bucles de clarificación, especialmente en aquellos casos en los que la consulta presenta ambigüedad o múltiples posibles interpretaciones. Más allá de estas mejoras concretas, los resultados de este trabajo ponen de manifiesto la necesidad de abordar el desarrollo y evaluación de chatbots basados en IA generativa desde una perspectiva integrada, que tenga en cuenta no solo la corrección de las respuestas, sino también su adecuación operativa, el riesgo asociado a su uso y las limitaciones de los métodos de evaluación automática en contextos reales de servicio.

Agradecimientos

FCT-KKKKK. CAU y LMM por las revisiones

Bibliografía

Abbasian, Mahyar, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, et al. 2024. "Foundation Metrics for Evaluating Effectiveness of Healthcare Conversations Powered by Generative AI." *NPJ Digital Medicine* 7 (1): 82.

Abeysinghe, Bhashithe, and Ruhan Circi. 2024. "The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches." *arXiv preprint arXiv:2406.03339*.

Akar, S. M., & Çetinkaya, G. (2024). The future of the public sector: era of artificial intelligence and chatbot applications. In *Generating Entrepreneurial Ideas With AI* (pp. 351-374). IGI Global Scientific Publishing.

Aoki, N. (2020). An experimental study of public trust in AI chatbots in the public sector. *Government information quarterly*, *37*(4), 101490.

Arends, M., & Mawela, T. (2024, December). Chatbot adoption in public service delivery. In *2024 International Conference on Computer and Applications (ICCA)* (pp. 1-7). IEEE.

Banerjee, Debarag, Pooja Singh, Arjun Avadhanam, and Saksham Srivastava. 2023. "Benchmarking LLM Powered Chatbots: Methods and Metrics." *arXiv preprint arXiv:2308.04624*.

Beaver, Ian. 2022. "The Success of Conversational AI and the AI Evaluation Challenge It Reveals." *AI Magazine* 43 (1): 139–141.

Chaganty, Arun Tejasvi, Stephen Mussmann, and Percy Liang. 2018. "The Price of Debiasing Automatic Metrics in Natural Language Evaluation." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 643-653).

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, *15*(3), 1-45.

Chen, T., & Gasco-Hernandez, M. (2025). Uncovering the results of AI Chatbot use in the public sector: Evidence from US state governments. *Public Performance & Management Review*, *48*(6), 1331-1356.

Chiu, Edwin Kwan-Yeung, and Tom Wai-Hin Chung. 2025. "Protocol for Human Evaluation of Generative Artificial Intelligence Chatbots in Clinical Consultations." *PLOS One* 20 (3): e0300487.

Clark, Elizabeth, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. "All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 7282-7296).

Cortés-Cediel, M. E., Segura-Tinoco, A., Cantador, I., & Bolívar, M. P. R. (2023). Trends and challenges of e-government chatbots: Advances in exploring open government data and citizen participation content. *Government Information Quarterly*, *40*(4), 101877.

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., & Cieliebak, M. (2021). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, *54*(1), 755-810.

Deschenes, Amy, and Meg McMahon. 2024. "A Survey on Student Use of Generative AI Chatbots for Academic Research." *Evidence based library and information practice*, *19*(2), 2-22.

Finch, Sarah E., James D. Finch, and Jinho D. Choi. 2023. "Don't Forget Your ABC's: Evaluating the State-of-the-Art in Chat-Oriented Dialogue Systems." *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15044-15071).

Gu, Jiawei, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, et al. 2026. "A Survey on LLM-as-a-Judge." *The innovation*

Holmes, Samuel, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael McTear. 2019. “Usability Testing of a Healthcare Chatbot: Can We Use Conventional Methods to Assess Conversational User Interfaces?” *Proceedings of the 31st European conference on cognitive ergonomics* (pp. 207-214).

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1-55.

Kaiser, Maximilian, and Christian Schulze. 2025. “ChatGPT Referrals to E-Commerce Websites: Do LLMs Outperform Traditional Channels?” *Available at SSRN 5585812*

Kim, Seungone, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, et al. 2023. “Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models.” *The Twelfth International Conference on Learning Representations*

Larsen, A. G., & Følstad, A. (2024). The impact of chatbots on public service provision: A qualitative interview study with citizens and public service providers. *Government Information Quarterly*, *41*(2), 101927.

Li, Haitao, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. “LLMs-as-Judges: A Comprehensive Survey on LLM-Based Evaluation Methods.” *arXiv preprint arXiv:2412.05579*.

Li, Qintong, Leyang Cui, Lingpeng Kong, and Wei Bi. 2021. “Exploring the Reliability of Large Language Models as Customized Evaluators for Diverse NLP Tasks.” *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 10325-10344).

Li, Xinmeng, Wansen Wu, Long Qin, and Quanjun Yin. 2021. “How to Evaluate Your Dialogue Models: A Review of Approaches.” *arXiv preprint arXiv:2108.01369*.

Liang, Hongru, and Huaqing Li. 2021. “Towards Standard Criteria for Human Evaluation of Chatbots: A Survey.” *arXiv preprint arXiv:2105.11197*.

Lin, C. Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out Artificial intelligence review*, *56*(4), 3055-3155.

Ni, J., Young, T., Pandelea, V., Xue, F., & Cambria, E. (2023). Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, *56*(4), 3055-3155.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. “BLEU: A Method for Automatic Evaluation of Machine Translation.” *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318)

Peras, Dijana. 2018. “Chatbot Evaluation Metrics.” *Economic and Social Development: Book of Proceedings*, 89-97.

Przegalinska, Aleksandra, Leon Ciechanowski, Anna Stroz, Peter Gloor, and Grzegorz Mazurek. 2019. “In Bot We Trust: A New Methodology of Chatbot Performance Measures.” *Business Horizons* 62 (6): 785–97.

Shawar, Bayan Abu, and Eric Atwell. 2007. “Different Measurement Metrics to Evaluate a Chatbot System.” *Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies* (pp. 89-96).

Siddals, Steven, John Torous, and Astrid Coxon. 2024. “‘It Happened to Be the Perfect Thing’: Experiences of Generative AI Chatbots for Mental Health.” *Npj mental health research*, *3*(1), 48.

Van der Lee, Chris, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. “Best Practices for the Human Evaluation of Automatically Generated Text.” *Proceedings of the 12th international conference on natural language generation* (pp. 355-368).

Venkatesh, Anu, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, 2017 “On Evaluating and Comparing Conversational Agents.” *Amazon Science*

Wang, Peiyi, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. "Large Language Models Are Not Fair Evaluators." *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 9440-9450).

Ye, Jiayi, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, et al. 2024. "Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge." *arXiv preprint arXiv:2410.02736*.

Yi, Z., Ouyang, J., Xu, Z., Liu, Y., Liao, T., Luo, H., & Shen, Y. (2025). A survey on recent advances in llm-based multi-turn dialogue systems. *ACM Computing Surveys*, *58*(6), 1-38.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, et al. 2023. "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.". *Advances in neural information processing systems*, *36*, 46595-46623.

Zhu, Lianghui, Xinggang Wang, and Xinlong Wang. 2023. "JudgeLM: Fine-Tuned Large Language Models Are Scalable Judges." *arXiv preprint arXiv:2310.17631*.

Zhu, Tiffany, Iain Weissburg, Kexun Zhang, and William Yang Wang. 2025. "Human Bias in the Face of AI: Examining Human Judgment Against Text Labeled as AI Generated." *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 25907-25914).