

Deployment of a Generative Chatbot Designed with Microsoft Copilot Studio for Public Service Delivery

David Escudero-Mancebo^{1,*†}, Guillermo Vicente-Oliva^{1,*†}

¹University of Valladolid, Paseo de Belén s/n 47002 Valladolid, Spain

Abstract

The emergence of generative artificial intelligence has simplified the design of conversational systems and improved chatbot interactions, leading to their rapid adoption in customer service, including within the public sector. However, there are still few empirical studies that systematically analyze the effectiveness of these systems in real-world service contexts. This study presents the deployment in a real scenario of a chatbot in a citizen service context. To this end, a chatbot was developed to assist with the university admission and enrollment process using the Microsoft Copilot Studio platform, which integrates large-scale language models from OpenAI. The system was designed with a narrow domain and supported by retrieval-augmented generation (RAG) mechanisms to minimize incorrect responses. The results show that the system fulfilled its purpose, enabling us to draw important conclusions about its performance and propose improvements for future campaigns.

Keywords

Generative artificial intelligence, chatbots, public administration, service quality assessment

1. Introduction

Dialogue systems have historically been one of the main areas of application for Natural Language Processing, as they naturally integrate language comprehension and generation tasks into real-world interactive contexts [1]. Within this framework, the recent emergence of generative artificial intelligence has transformed the development of conversational systems, both from a technological standpoint and in terms of their design and deployment [2]. In particular, large-scale language models (LLMs) have facilitated the interpretation of user turns, expanded access to knowledge contained in textual documents—beyond structured databases—and increased the naturalness, expressiveness, and variability of generated responses. As a result, the range of services that can be automated through conversational interfaces has expanded significantly, while the adoption of chatbots has intensified across multiple customer service domains, including public services [3].

Even before the rise of generative AI, the evaluation of dialogue systems and chatbots was already an established field of research, with numerous methodological approaches aimed at measuring dimensions such as correctness, appropriateness, user satisfaction, and task success [4]. However, the incorporation of generative models introduces challenges that hinder the direct transfer of these evaluation frameworks [5]. Among these are the emergence of hallucinations and factual errors [6], as well as new operational constraints associated with computational cost per turn and latency in response generation. At the same time, the rise of generative AI has driven new strategies for automatic evaluation; however, their validity and usefulness remain debatable when applied to conversational systems deployed in real-world service environments. Consequently, a gap persists between the rapid deployment of these technologies and the availability of empirical studies that systematically analyze their actual performance under operational conditions.

This paper presents details on the implementation of a service for assisting future students with doubts concerning the enrollment process. We report on the results of the deployment of the UVaBot

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22–25 September 2026.

*Corresponding authors.

† These authors contributed equally.

✉ escudermancebo.david@uva.es (D. Escudero-Mancebo); guillermo.vicente.oliva@estudiantes.uva.es (G. Vicente-Oliva)

ORCID 0000-0003-0849-8803 (D. Escudero-Mancebo); 0009-0004-9802-9895 (G. Vicente-Oliva)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

service, which was used to assist students with the enrollment process at the University of Valladolid during the 2025–26 academic year. First, the system architecture is described, followed by the service’s efficiency and effectiveness metrics.

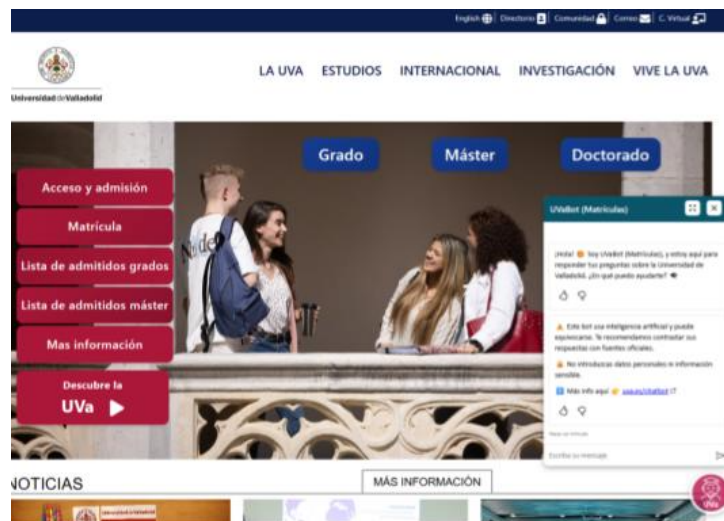


Figure 1: Integration of the service in the web page of the University.

2. System description

The system under review, called UVA Bot, is a chatbot based on generative artificial intelligence designed to assist incoming students with administrative processes related to admission and enrollment at the University of Valladolid. The UVA Bot chatbot was deployed in production during the pre-registration and enrollment periods for the 2025/26 academic year, remaining inactive outside of those campaigns to minimize risks and operational costs. The service was integrated into the institutional portal (<http://www.uva.es>) via a floating conversational widget (bottom left of Figure 1) that appears as an icon with an avatar on the web interface and, when activated, displays a pop-up dialog window where the user can ask questions in natural language.

The system was developed using the Microsoft Copilot Studio platform, which integrates OpenAI’s large-scale language models and retrieval-augmented generation (RAG) capabilities. The chatbot’s architecture is based on a generative orchestration mechanism. In this approach, a language model acts as an orchestrator to analyze the intent of the user’s message and determine the most appropriate topic within the assistant’s domain. Based on the result of this identification, the system can provide a predefined response or generate a response based on retrieving information from a knowledge repository (see Figure 2). The knowledge repository consists of 21 documents with a total of 57,079 BPE tokens (listed in Table 2).

The system was evaluated from two complementary perspectives: the quality of the chatbot’s responses and a set of efficiency and interaction indicators derived from its real use. Activity logs exported from Microsoft Copilot Studio were used as the main data source, although these logs truncate system responses to 512 characters, which prevents reliable assessment of generative outputs. To address this limitation, we built a dedicated evaluation set by regenerating responses to a random sample of real user queries. Each query was automatically replayed through the chatbot interface using a Python script based on Selenium, which submitted the questions and captured the full responses directly from the web client. For multi-turn sessions, only the first substantive user query was retained, excluding greetings and other non-informative openings. Response quality was then assessed on a 1–5 Likert scale measuring overall utility, where lower scores corresponded to incorrect, fabricated, or critically incomplete answers and higher scores to correct, complete, relevant, and clearly formulated

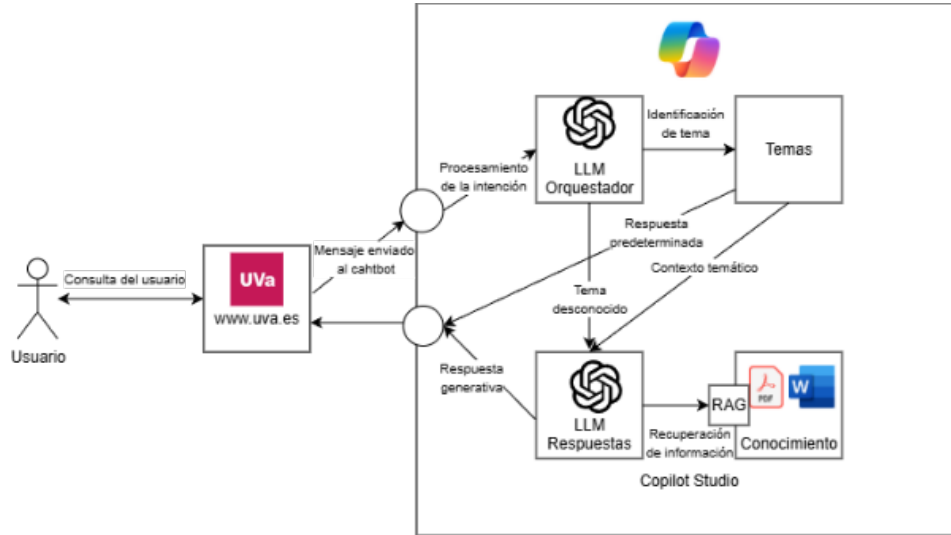


Figure 2: Description of the architecture of the service.

ones. Manual annotation was carried out through a dedicated web interface, which recorded both the assigned score and a short textual justification for each question–answer pair. This human evaluation was performed by staff from the University’s User Support Center (CAU), who also handle real admission and enrollment inquiries. In parallel, an automatic evaluation following the LLM-as-a-judge paradigm was conducted using NotebookLM on the same set of questions and responses. NotebookLM was provided with the same knowledge base files used by the chatbot and instructed to justify its ratings without relying on external sources. The comments produced by both the human evaluators and NotebookLM were subsequently reviewed and categorized by the authors, who also analyzed the consistency between human and automatic judgments. Finally, additional descriptive variables were considered to characterize real interaction with the system, including automatic thematic classification of user queries and approximate response-time measurements obtained through controlled re-execution of consultations.

3. Results

Average system usage during the analyzed period was 19 conversations per day, with a peak of 81 conversations per day. 91.4% of conversations were short (three exchanges or fewer), with the majority of these (68.4%) consisting of a single exchange. The average response time to assist users is 2.5 seconds, with outliers exceeding 10 seconds. The Copilot Studio platform provides an automatic estimate of effective conversation resolution based on the system’s internal heuristics, such as the absence of interpretation errors or the activation of fallback mechanisms. According to this metric, 420 sessions were classified as resolved, while 136 (24.46%) were considered unresolved. Regarding topics, Table 1 shows that, although enrollment-related topics stand out, the system was able to orchestrate questions across the various programmed topics.

Table 1: Distribution of the orchestrated topics.

Category	N	Percentage
Enrolment	135	13,78%
Other queries related to university studies	82	8,37%
University Identity - Access to Services	73	7,45%
Contact Options	71	7,24%
Non-college inquiries	63	6,43%

Category	N	Percentage
Access to university studies	58	5,92%
Academic calendar, class schedules	51	5,20%
PAU, Cut-off marks and places offered	50	5,10%
Cancellation or modification of enrolment	44	4,49%
Information on undergraduate studies	43	4,39%
Greeting	42	4,29%
Enrolment – Bachelor’s Degree	36	3,67%
Access to university studies - Bachelor’s degree	34	3,47%
Pricing Information	32	3,27%
University Identity - Activate or Retrieve Key	32	3,27%
Validation of subjects/credits	29	2,96%
Transfer of file	27	2,76%
Group Information, Group Change	24	2,45%
University identity	15	1,53%
Access to university studies - Master’s degree	8	0,82%
Information on master’s studies	8	0,82%
Final degree project (TFG), Master’s Degree Final Project (TFM)	7	0,71%
External internships	6	0,61%
Enrolment – Master’s Degree	6	0,61%
Erasmus	4	0,41%

Responses using generative AI predominate (71.1%), particularly for inquiries regarding enrollment, which constitute the core functionality of the system, while greetings and inquiries unrelated to the university were primarily handled using predefined responses. Table 2 shows the knowledge sources used both when analyzing questions and when formulating responses.

Both the CAU evaluators and NotebookLM identified a non-negligible number of problematic responses in the selected sample. A closer analysis shows that most of these issues do not correspond to outright false information, but rather to answers that should have been more complete or to cases in which the system should have requested clarification before responding. Nevertheless, genuinely incorrect answers were also observed in around 10% of the sample. These errors often stem from queries that combine more than one topic; for instance, when asked about enrollment dates for Aerospace Engineering, the system provides general enrollment dates without taking into account that this degree is not offered at the University of Valladolid. The main source of error, however, concerns transfer-related queries, where the system fails to recognize that transfer procedures depend on the destination university and instead consistently answers according to the regulations of its own institution. This latter type of error can be mitigated by refining the wording and scope of the corresponding knowledge source on transcript transfer procedures. A data set with the complete list of questions/answers and evaluations records is provided with free access at [7].

Table 2: Use of the different sources of knowledge during the response generation.

Source of knowledge	Answers	Questions
Master’s Degree Self-Enrolment Manual	17,59%	23,57%
Questions and answers for undergraduate students	13,28%	17,79%
Enrolment	8,91%	11,95%
Access and pre-enrolment for master’s degree studies	8,14%	10,91%
Pre-enrolment and admission degree	6,69%	8,96%
Enrolment for master’s degree studies	6,30%	8,44%
Self-enrolment manual 24-25 grade	6,15%	8,25%

Source of knowledge	Answers	Questions
Enrolment for a Master's Degree in Secondary Education	5,72%	7,66%
Access and pre-enrolment for the Master's Degree	5,14%	6,88%
New Admission Enrolment Documentation	3,34%	4,48%
Transfer of file	3,20%	4,29%
Tuition Fee Information	3,00%	4,03%
RESOLUTION BOCYL Num. 93	2,62%	3,51%
Previous student enrolment documentation	2,08%	2,79%
Cut-off marks 2024-2025	1,79%	2,40%
Information PAU	1,74%	2,34%
University Entrance Exam Guide	1,45%	1,95%
New Degree Entry Places 2025-26	1,16%	1,56%
Documentation for transfer students	0,92%	1,23%
Regulations for the recognition and transfer of credits	0,68%	0,91%
Credit recognition and transfer	0,10%	0,13%

4. Discussion

From a development perspective, experience with Copilot Studio shows that incorporating generative AI capabilities simplifies the deployment of conversational services, where initial configuration, topic orchestration, the integration of knowledge sources, and response generation can be accomplished with minimal implementation effort compared to traditional approaches. At the same time, the monitoring mechanisms provided by the platform, while useful for general service tracking, have significant limitations from an analytical standpoint. In our case, the orchestration reports allowed us to verify that the system was used in a manner consistent with the intended domain (Table 2) and that routing (Table 1) to the defined topics was, in general terms, adequate. However, the usage logs available on the platform proved insufficient for a detailed evaluation of system performance. Among the shortcomings observed are the absence of precise metrics on consumption per transaction, especially in terms of tokens; the low granularity regarding the actual use of knowledge units; and the lack of clear documentation on some internal indicators, such as the estimation of conversational success.

From the user's perspective, the results show that the service is used extensively, especially considering that the university enrollment process has long been conducted through a well-established online system accompanied by extensive documentation. In this context, the use of a chatbot as an additional channel for inquiries suggests that there is a genuine demand for more direct and accessible assistance. Analysis of the interactions also indicates that most inquiries fall within the intended scope, although uses outside the system's scope are also observed, highlighting users' tendency to interpret this type of interface as a general entry point to institutional information. Conversations are mostly brief, with a clear predominance of single-turn interactions. This behavior is consistent with the nature of the service, which is geared toward the timely resolution of queries through access to structured information, and with the fact that many responses redirect the user to external resources where further information can be found. One of the most critical issues identified is the system's response time. University officials expressed their satisfaction with the service; no user complaints were received during its implementation, and CAU officials have expressed interest in relaunching it this academic year.

These results also suggest that responses generated by generative AI systems must remain open to expert scrutiny, since apparently plausible answers may still be judged as inadequate, incomplete, or even incorrect in an institutional setting. Although the safeguards incorporated into the system help to constrain its behavior, they do not completely prevent hallucinations or the generation of misleading procedural information. At the same time, the analysis shows that a substantial proportion of these problems can be mitigated through relatively simple interventions, such as adding clarification or

confirmation turns when user intent is ambiguous and refining the wording and scope of the knowledge sources after iterative evaluation and real use.

5. Conclusions

This article presents a real-world evaluation of a public service chatbot based on commercial generative AI technology. The results show that the system is capable of providing accurate and well-formulated responses in most interactions, drawing on the provided knowledge bases and adhering to the control mechanisms and constraints defined in its design. It has also been observed that delays in generating certain responses, combined with the frequent activation of control mechanisms —such as predefined responses to limit the domain or redirects to external resources— can negatively impact users' perception of service quality. It has been found that, despite the precautions taken to ensure the quality of responses —such as restricting the system to institutional knowledge sources and limiting its domain— a significant percentage of interactions receive negative ratings. However, detailed analysis shows that most of these evaluations are not due to hallucinations, but rather to the criteria applied by the evaluators. Both human and automated evaluators tend to penalize predefined or poorly elaborated responses, while human evaluators also introduce nuances based on additional knowledge not explicitly stated in the sources used by the system.

Acknowledgements

This work is part of the Project “Asesoramiento en el uso de sistemas inteligentes para la gobernanza del dato en entornos provinciales: caso de uso en la provincia de Valladolid (SIGODEP-VA)” FECYT FCT-24-21134. We would like to thank the CAU of the University of Valladolid for reviewing the questions and answers and Luis Manuel Montero, hired with the youth employment guarantee program of the Junta de Castilla y León. Also to Susana Álvarez and Juan Pablo de Castro for the initiative and support.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

1. Ni, Jinjie, Tom Young, Vlad Pandelea, Minlie Huang, and Erik Cambria. "Recent advances in deep learning based dialogue systems: A systematic survey." *Artificial Intelligence Review*, vol. 56, no. 4, 2023, pp. 3055-3155.
2. Yi, Zihao, Jinfeng Zhou, Nugraha Priya Utama, and Xian-Ling Mao. "A survey on recent advances in LLM-based multi-turn dialogue systems." *ACM Computing Surveys*, vol. 58, no. 6, 2025, pp. 1-38.
3. Larsen, Anna Grøndahl, and Asbjørn Følstad. "The impact of chatbots on public service provision: A qualitative interview study with citizens and public service providers." *Government Information Quarterly*, vol. 41, no. 2, 2024, p. 101927.
4. Deriu, Jan, Alvaro Rodrigo, Arantxa Otegi, Guillermo Spiegelberg, Morten Juul-Jensen, and Mark Cieliebak. "Survey on evaluation methods for dialogue systems." *Artificial Intelligence Review*, vol. 54, no. 1, 2021, pp. 755-810.
5. Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Wenchao Shangguan, Shuguang Han, Yong Bai,

- Xiaohao Sun, Lijun Wu, Shimin Tao, Wongisig Kim, Hsin-Hsi Chen, and Xing Xie. "A survey on evaluation of large language models." *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, 2024, pp. 1-45.
6. Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Feng Ajiki, Feng-Xuan Han, Tianshi Zhang, Hai-Ying Wang, and Jiazhen S. Zhou. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions." *ACM Transactions on Information Systems*, vol. 43, no. 2, 2025, pp. 1-55.
 7. Escudero-Mancebo, David, González Rodríguez, Melquiades, Vicente-Oliva, Guillermo. "Analysis and Implications of Alignment Between LLM-as-a-Judge and Human Evaluators in a Generative AI Chatbot: A Real-World Study", *Text, Speech and Dialog*, 29th International Conference TSD 2026 Proceedings. *Lecture Notes in Artificial Intelligence*. Vol. 16940, 2026 chap.46