Design and Evaluation of a Conversational Assistant for Supporting Erasmus+ Student-Exchange Volunteers

Efstratios Pantazaras
Computer Engineering & Informatics
University of Patras
Rio, Greece
up1067490@ac.upatras.gr

Andreas Komninos
Computer Engineering & Informatics
University of Patras
Rio, Greece
0000-0001-8331-1645

Stylianos Mystakidis
Museum of Science & Technology
University of Patras
Rio, Greece
0000-0002-9162-8340

Athanasios Christopoulos
Faculty of Science
University of Turku
Turku, Finland
0000-0002-1809-5525

***Abstract*—At the University of Patras, Greece, incoming Erasmus+ students for studies rose by about 172 percent between 2014 and 2025 but the number of local Erasmus+ Student Network (ESN) volunteers that offer crucial assistance and support for incoming students, has remained nearly identical. Local volunteers often cannot answer routine questions whilst, the relevant information is scattered across documents and websites that students rarely consult. In response to this obstacle, this work presents the design and evaluation of a domain-specific conversational assistant, built with the open-source Rasa framework on a documented analysis of real questions, that provides immediate vetted answers. Its aim is not to replace local volunteers, but instead support their work in helping incoming students. The evaluation combined a subjective study of user experience with thirty-four participants and a counterbalanced within-subjects experiment with twenty volunteers that compared answering tasks with and without the assistant using the NASA Task Load Index and objective interaction logs. Participants rated the experience highly with the weighted workload dropping significantly. The reduction was larger for more experienced ESN members. The assistant works as a complement to human support which can help to reduce the effort and frustration of question answering.**

***Index Terms*—Conversational agents, Chatbots, Student support, Erasmus+ mobility, Cognitive workload, Usability evaluation**

I. Introduction

The Erasmus+ programme moves hundreds of thousands of students across Europe each year with the Erasmus+ Student Network (ESN) supporting roughly 350,000 international students annually through more than 500 local sections [1], [2]. However, incoming students face a dense set of practical and bureaucratic tasks on arrival ranging from university services (e.g., enrolment, course selection, identity card) to accommodation and local transport. Most of this information exists on institutional websites and in guides but locating it is time-consuming and often requires fluency in the local language which adds uncertainty for newcomers [2]. At the University of Patras, Greece, the gap between demand and volunteer capacity is visible in the figures. Incoming Erasmus+ students for studies rose from 83 in 2014–2015 to 226 in 2024–2025, an increase of about 172%. Over the same period, the local ESN section (the University of Patras network branch), has remained close to twenty members throughout [3]. The 'Buddy' programme, which pairs a local student with an incoming student, expanded only after 2023 but, even so, the number of Buddies in 2024–2025 (N=152) trailed the number of incoming students [3]. A recurring problem is that many Buddies cannot answer basic questions from incoming students. Attendance at the section's informational and training calls was low across recent years, at 16%, 18%, and 12% of registered Buddies. As such, a large share of the questions, which could be answered from the information sheets already distributed, does not reach ESN volunteers. The pattern is consistent with the principle of least effort and information-seeking behaviour in which people prefer the quickest acceptable source, as fragmented information is harder to use [4]. The consequences are delayed responses, unreliable answers, lower trust in the Buddies, and load shifting onto a small number of more experienced ESN volunteers. Conversational agents are an established response to high-volume, repetitive enquiry; they answer common questions immediately and reduce the load on administrative and volunteer staff [5], [6]. Our study reports the key characteristics and evaluation of a domain-specific conversational assistant developed for the Erasmus+ support context at the University

979-8-3195-2278-8/26/$31.00 © 2026 IEEE

of Patras. In contrast with deployments aiming to replace, or substitute for human assistants, our system is accessible only to local volunteers, as in the context of an Erasmus+ exchange, peer network building and intercultural interaction is a core mechanism for achieving the aims of Higher Education Mobility in the Erasmus+ program. Two features set the study apart from typical chatbot evaluations. The first is the outcome it measures. Most chatbot evaluations ask whether end users find the system usable; the current experiment instead measures the cognitive workload of the volunteers who answer student questions, with and without the assistant available, in a counterbalanced within-subjects design (N = 20), using the NASA Task Load Index and objective interaction logs. A study of user experience (N = 34) records how volunteers and former students experience free interaction with the assistant. The second is that our evaluation demonstrates that the reduction in workload is not uniform, and depends on the volunteer's prior experience and on how the assistant is used.

II. Background and related work

A. Conversational Agents in Student Support

Conversational agents have a long history, from early systems such as ELIZA [7] and PARRY [8] to current assistants that combine natural language understanding, dialogue management, and programmable actions [6], [9], [10]. In Higher Education, they have been widely adopted to answer administrative and frequently asked questions and to ease the burden on staff [5], [11]. In a support setting, the priority is somewhat different. The agent does not need to sustain an open-ended conversation; it needs to return valid, reliable, well-scoped answers from a pre-defined knowledge domain [12]. In this context it can relieve the volunteers and staff who would otherwise field the same questions repeatedly.

B. Development Frameworks

Although significant breakthroughs in LLM technology are driving state-of-the-art solutions in conversational agents, we chose not to deploy an LLM for this project, but instead used the open-source Rasa framework, a machine learning platform for conversational generative artificial intelligence applications such as chatbots and virtual assistants [13]. The selection of the open-source Rasa framework over a commercial, out-of-the-box conversational assistant is driven by critical requirements for deterministic control, data governance, and economic viability tailored to the constraints of the university's Erasmus+ support ecosystem. Unlike generic, cloud-based generative AI agents, which are prone to hallucinations and unverified outputs, Rasa's combination of Intent and Entity Recognition with highly configurable dialogue policies ensures high response predictability. This allows the system to deliver strictly vetted answers regarding sensitive institutional procedures (e.g., enrolment, visa verification, and university identification). Furthermore, as an open-source platform, Rasa supports complete on-premise deployment. This ensures compliance with institutional data privacy regulations (such as GDPR) by preventing student queries and internal university documentation from being exposed to third-party infrastructure. Rasa avoids the issues associated with alternative Large Language Model (LLM) deployments. A self-hosted LLM introduces significant capital expenditure, demanding specialized hardware configurations (such as high-VRAM GPUs) and administrative costs for maintaining the dedicated server. Conversely, relying on cloud-hosted APIs (e.g., Google Gemini, OpenAI GPT, or Anthropic Claude) shifts the financial burden to variable operational expenditure. Because a Retrieval-Augmented Generation (RAG) architecture is required to parse fragmented university documents, every student query necessitates injecting extensive context and system prompts into the context window. Compounding token overhead introduces unpredictable, volume-dependent billing liabilities that are poorly suited for fixed institutional budgets. Rasa circumvents both paradigms; its lightweight, intent-based machine learning pipelines operate efficiently on low-spec, pre-existing university CPU infrastructure, neutralizing both upfront hardware acquisition costs and recurring API fees.

C. Evaluating Conversational Systems

The evaluation of conversational systems draws on several dimensions. Subjective user experience is most often assessed through questionnaires, such as the Chatbot Usability Questionnaire (CUQ) [14], the Bot Usability Scale (BUS-11) [15], and other satisfaction-related instruments developed for such interactions [16]. A systematic mapping of this literature reports that the usability evaluation of chatbots remains fragmented and that it frequently relies on questionnaires alone [17]. Other lines of work have taken different routes; some measure interaction quality at the level of dialogue turns [18], some estimate user satisfaction automatically [19], and some benchmark systems through structured procedures [19]. Composite metrics that combine usability, engagement, and effectiveness have also been proposed for educational chatbots [11]. A secondary study of chatbot usability experiments found that most such studies adopt a within-subjects design with the Natural Language Processing (NLP) aspect being the most frequently identified as needing improvement [20].

D. Workload as an Evaluation Outcome

The workload of support volunteers and staff is a recognised but under-used outcome for support tools. The NASA Task Load Index is a validated instrument that captures perceived workload across mental, physical, and temporal demand, performance, effort, and frustration [21]. It has been applied well beyond its origins in aviation, in healthcare, driving, and Human–Computer Interaction (HCI), because it separates the sources of load and is quick to administer. Treating the supporter's workload as an outcome shifts the question that an evaluation asks of a support chatbot, from whether it answers questions to how it redistributes the effort of answering them. Most student-support chatbot evaluations report usability and satisfaction [17], [20]; fewer pair subjective experience with an experimental, within-subjects measure of the workload

borne by the human supporters and link it to objective usage logs. The present study takes that combined approach in an Erasmus+ support setting.

III. System Design

The assistant was designed from a documented needs analysis so that its scope would reflect the questions incoming students actually ask instead of making anecdotal assumptions about them. To this end, two questionnaires were used to collect the questions that students raise and that volunteers field in practice. One was completed by fifteen ESN members and Buddies and the other by sixteen former incoming Erasmus+ students. The exercise returned about 310 questions, which we grouped into recurring themes including: 1) accommodation, 2) transport, 3) academic procedures, 4) the Learning Agreement, 5) departmental coordinators, and 6) the cost of living in Patras. The themes determined both the scope of the assistant and the content of its answers. On this basis, the functional requirements were specified: intent and entity recognition, multi-step dialogue, structured answers, custom actions for requests that needed additional logic, a fallback mechanism, and web integration. The non-functional requirements concerned usability, mobile compatibility, continuous availability, the avoidance of personal-data collection, interaction in English, and a curated knowledge base that would not update automatically from external sources.

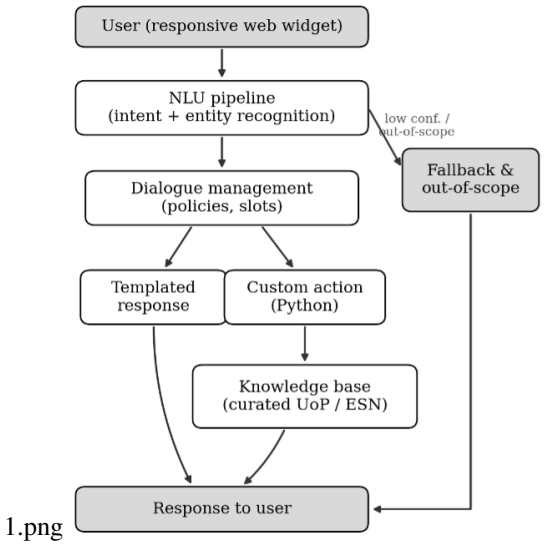
A. Architecture

The assistant was implemented with the open-source Rasa framework which separates language understanding from dialogue management (Fig. 1). An incoming message passes through a natural language understanding pipeline that tokenises the text, extracts features, classifies the user intent, and identifies entities such as a department name or a service. The recognised intent and entities update the dialogue state and dialogue policies decide the next action which may be a templated response, a follow-up question to fill a missing slot, or a custom action. Defining the intents, entities, slots, responses, and actions in the domain kept the assistant's behaviour predictable and inspectable.

The language-understanding model was trained on example phrasings for each intent, drawn from the questions collected in the needs analysis, together with annotated entities; the dialogue behaviour was specified through example conversations and rules. Because the intents, entities, and answers were specified in advance, both the coverage of the assistant and its limits were known before deployment. The rationale behind this design choice is grounded on the fact that a confidently wrong answer is more damaging than a referral to a human.

B. Custom Actions, Knowledge Base

Requests that could not be served by a fixed reply were handled by custom actions written in Python. For example, a request for the Erasmus+ coordinator of a department triggered an action that looked up the contact details in a structured file listing all coordinators at the university, so a single action



1.png

Fig. 1. Architecture of the conversational assistant.

covered every department. The answer content was drawn from official University and ESN sources, then rephrased and organised for direct use, with the wording kept under editorial control, which held them consistent and within scope.

C. Fallback and Out-of-Scope Handling

Inputs the assistant could not confidently interpret were handled in two stages. A two-tier fallback first asked the user to rephrase the question and, if understanding still failed, listed the topics the assistant can help with so the user could redirect the conversation. Separately, dedicated out-of-scope intents recognised common requests outside the assistant's remit and pointed the user to the appropriate ESN contact channel.

D. Deployment and Data Handling

The assistant was embedded in a responsive web widget usable on both desktop and mobile devices, matching how students and volunteers typically look for information. No personal data were stored. For the evaluation, a per-participant access-code controlled entry and linked each participant's questionnaire responses to their interactions which were subsequently logged with the message text, the timestamps, the response latency, and the page. The logs supported both the analysis of use and a check on the experimental manipulation.

IV. Evaluation Methodology

The assistant was evaluated in two phases. The first, assessed the subjective experience of using the assistant under free interaction. The second was a controlled, within-subjects experiment, that compared answering student questions with and without the assistant, with perceived workload as the primary outcome, supported by objective interaction logs.

A. Phase 1: Subjective Evaluation

Thirty-four participants took part, drawn from three groups relevant to the application, Local Buddies, ESN members, and

former incoming Erasmus+ students. Each used the assistant freely, asking as many questions as they wished, and then completed a twenty-item questionnaire on a five-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree) shown when they closed the chat. Only fully completed questionnaires could be submitted so there were no partial responses. The instrument was a composite assembled from published items, usability items from the BUS-11 [15] and the CUQ [14], with information-quality, dialogue, trust, usefulness, satisfaction, and continuance items from a study of AI chatbots [16]. One item on conversational naturalness was negatively worded and reverse-scored. The instrument is a non-validated composite and thus, its results are reported descriptively.

B. Phase 2: Experimental Design

The second phase used a within-subjects design with two conditions: answering questions 1) without and 2) with the assistant. To limit order effects, two equivalent versions of an online form presented the conditions in opposite orders and each version was completed by ten participants (groups of ten)—one group was Local Buddies and the other was volunteers who were both ESN members and Buddies.

1) Tasks: In each condition, participants answered fifteen open-ended questions on topics within the assistant's scope such as contact details, procedures, transport, infrastructure, and daily-life matters. The fifteen questions were organised in three equally distributed difficulty tiers (low, medium, and high), graded by the estimated effort to locate a correct answer and the depth of knowledge required, so that each condition spanned the same range of difficulty. The two question sets were matched on topic and difficulty but were not identical, which prevented participants from simply repeating a previous answer and kept the conditions comparable. For instance, one condition asked for the "Erasmus+ coordinator of one department" and the other for "the coordinator of a different department". Representative items ranged from a single contact lookup at the low tier (e.g., "the email of a department's Erasmus+ coordinator"), through a transport question at the medium tier (e.g., "which bus connects the city centre to the campus"), to a multi-step procedure at the high tier (e.g., "how to apply for a place in university accommodation").

2) Measures: After each condition, participants completed the NASA Task Load Index across its six dimensions (mental, physical, and temporal demand, performance, effort, and frustration) and a weighted overall workload score was computed by combining the dimension ratings with the weights obtained from the instrument's pairwise comparisons [21]. At the end of the form, five Likert items recorded the participant's overall judgement of the assistant covering the quality of support, comfort when helping students, the reduction of ESN members' workload, the completeness of answers relative to those typically given by Buddies, and help with questions that would otherwise require additional searching. A unique identifier linked each participant's responses to their logged interactions which were examined retrospectively to confirm that no assistant use occurred in the without-assistant condition.

TABLE I
PHASE 1 USER-EXPERIENCE RATINGS

User-experience item	**Agreement** ($N = 34$)
Ease of use	99%
Clarity of responses	99%
Timeliness of responses	99%
Innovative way to provide the service	97%
Ease of finding the assistant	95%
Willingness to recommend	95%
Accuracy of information	94%
Sufficiency of information	94%
Reliability of information	94%
Time saving	94%
Useful answer format	94%
Trust and privacy	89–91%
Understands the exact meaning	80%
Not too robotic (reverse-scored)	69%
Overall composite	92%

3) Analysis: Analyses were carried out in SPSS Statistics. A two-way ANOVA was used with assistant condition (without, with) as the within-subjects factor, user type (Local Buddy; ESN member and Local Buddy) as the between-subjects factor, and the weighted NASA-TLX score as the dependent variable. Approximately normal distributions were supported by visual inspection and by skewness and kurtosis values within conventional thresholds [22]. Likewise, the variances were comparable across the two groups in both conditions. The significance threshold for the analyses was set at .05 with the effects reported with partial η^2.

C. Ethics and Data Handling

Participation in both phases was voluntary and based on informed consent. Participants were eligible to drop out from the study at any point. No personal or identifying data were collected by the assistant except for the encoded per-participant access codes which were used to link the questionnaire responses to interaction logs (message text, timestamps, response latency, and the page) without revealing identities.

V. RESULTS

A. Phase 1: User Experience

Ratings across the twenty items were high and tightly clustered with a composite agreement of 92% (M = 4.58 on the five-point scale; Table I). The assistant scored highest on the mechanical qualities of the interaction, ease of use, clarity, and speed, with the information-quality and trust items a little below them. The two lowest ratings are the informative part. The assistant's ability to grasp the exact meaning of a question (80%) and the judgement that it seemed too robotic (69%) sit well below the rest. The identified weak-points (i.e., semantic understanding and conversational naturalness) are essentially what an intent-based design constrains by construction.

B. Phase 2: Perceived Workload

Table II reports the weighted NASA-TLX scores by group and condition. Both groups reported lower workload with the assistant available as also seen from the pooled mean

TABLE II
WEIGHTED NASA-TLX SCORES BY USER TYPE AND CONDITION

User type	N	Without M (SD)	With M (SD)	Δ
ESN member and Local Buddy	10	42.00 (9.06)	22.47 (8.83)	19.53
Local Buddy	10	52.80 (12.62)	40.47 (7.10)	12.33

TABLE III
TWO-WAY ANOVA ON THE WEIGHTED NASA-TLX SCORE

Effect	$F(1, 18)$	p	η_p^2
Condition (within)	88.88	$< .001$	.83
User type (between)	13.27	.002	.42
Condition $\times$ user type	4.54	.047	.20

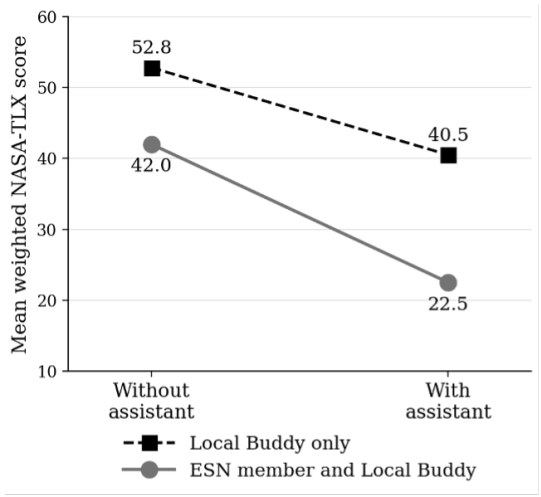
which fell from 47.40 to 31.47. The two-way ANOVA (Table III) showed a large, significant main effect of condition, $F(1, 18) = 88.88, p < .001$, partial $\eta^2 = .83$, indicating substantially lower workload when the assistant could be used. The main effect of user type was also significant, $F(1, 18) = 13.27, p = .002$, partial $\eta^2 = .42$; Local Buddies reported higher workload than participants who were also ESN members, in both conditions. The condition-by-user-type interaction was significant as well, $F(1, 18) = 4.54, p = .047$, partial $\eta^2 = .20$, with a larger reduction for the ESN-member-and-Buddy group (19.5 points) than for the Local-Buddy group (12.3 points) (Fig. 2). The interaction sits near the .05 threshold so it is treated as preliminary.

Separating the six dimensions shows where the reduction originated. Every dimension declined, although unevenly. Effort fell by far the most; frustration, perceived performance, and mental demand fell moderately; and physical and temporal demand barely moved. The pattern is coherent; having vetted answers to hand reduced the effort of composing a reply and the frustration of failing to locate one. It did little to the duration of the task, which still had to be read, answered, and submitted. In this sense, the reduction is best understood as a redistribution of effort as opposed to a saving of time.

C. Objective Use & Closing Perception Items

Across the 20 participants, 253 interactions with the assistant were logged (M = 12.65). Use was slightly higher and more uniform among Local Buddies (M = 13.3, SD = 1.34) than among the ESN-member-and-Buddy group (M = 12.0, SD = 2.54). The mean response time was 2499.77ms and the median 2400ms, with a maximum of 5868ms (about 2.5 seconds in typical use). The logs confirmed absence of manipulation, as no assistant interactions were recorded for any participant in the without-assistant condition. Their slightly higher use is consistent with higher reported workload.

The five closing items drew agreement with responses in the two highest categories and none in disagreement (Table IV). The clearest endorsements were that the assistant improves the quality of support to incoming students and that it lightens the



2.png

Fig. 2. Mean weighted NASA-TLX score by condition and user type.

TABLE IV
PHASE 2 CLOSING PERCEPTION ITEMS ($N = 20$)

Closing perception item	Agreement
Improves the quality of support	99%
Can reduce ESN members' workload	96%
Helps with otherwise-effortful questions	90%
More complete than typical Buddy answers	88%
Would make me more comfortable supporting students	87%

load of answering routine questions. Agreement was a weaker, though still positive, on the two items that set the assistant against the volunteers themselves (if its answers are more complete than a Buddy's and if it would make the respondent more comfortable in the role). The ordering fits volunteers crediting the assistant with routine questions but reserving judgement on how it compares with their own help.

D. Exploratory Follow-up on the Interaction

To interpret the interaction, the lead researcher contacted thirteen participants after the experiment, eight from the ESN-member-and-Buddy group and five Local Buddies. The follow-up was not planned in advance and is thus, an exploratory way to contextualise the quantitative results. Local Buddies reported that the assistant often supplied information that was genuinely new to them, such as the procedure for applying for university accommodation, so using it involved understanding and evaluating new content, which kept part of their workload even with the assistant available. The more experienced participants were already familiar with most of the content and used the assistant mainly for quick confirmation and ready-phrased answers, which reduced the effort of composing a reply. These findings point to a difference in use (learning versus confirmation) which warrants confirmatory testing.

VI. DISCUSSION AND LIMITATIONS

The assistant received high ratings for usability and information quality, with its availability significantly reducing volunteer workload. These results extend traditional chatbot

evaluations, which predominantly report usability and satisfaction [11], [17], toward a measured operational effect on human workload. Experienced volunteers obtained a larger workload reduction than less experienced peers. Follow-up accounts indicate this stems from usage patterns rather than baseline domain knowledge: experienced volunteers used the tool to quickly confirm familiar information, whereas less experienced volunteers were actively learning while processing responses, retaining some cognitive load [4]. These interaction dynamics warrant confirmation in a larger study.

The findings confirm that the assistant should complement human support rather than replace it. A controlled, intent-based architecture provides guaranteed data privacy, predictable scope, and reliability over raw generative models. However, lower scores in conversational naturalness and semantic understanding mark where controlled hybrid generative components could be integrated. Practically, local ESN chapters can use the assistant to absorb routine questions, free volunteer time for complex cases, and improve Buddy onboarding. The observed workload reduction aligns with reduced cognitive load from chatbot task support [23], addressing an operational outcome rarely evaluated [17], [20]. Key limitations include a small single-site sample ($N = 20$), short-term controlled testing, a non-validated Phase 1 composite scale, and a lack of dialogue-resolution logging.

VII. Conclusion and Future Work

A domain-specific conversational assistant for Erasmus+ student support was designed and evaluated across two phases. Participants rated the system highly on usability, clarity, speed, and reliability. Experimental availability significantly reduced volunteer workload, with the effect being stronger for experienced volunteers using it for information confirmation than for newer volunteers using it as a learning resource.

Future work will focus on extending the knowledge base using fallback logs, integrating hybrid LLM components for enhanced natural language understanding while preserving control, adding multilingual capabilities, and connecting the platform to live institutional sources. Longer deployments in real-world settings will determine whether workload reductions are sustained over time and enable dialogue-level resolution tracking. Finally, a confirmatory multi-section study will further examine the interaction between volunteer experience and workload reduction, with potential adaptations for outgoing student mobility and other peer-support institutions facing repetitive inquiries.

References

[1] "ESN Annual Report 2024/2025," https://issuu.com/esnint/docs/esn_annual_report_2024_2025, Jul. 2025.

[2] "ESNsurvey XV: Making Quality Mobility a Reality for All," https://issuu.com/esnint/docs/xv_esnsurvey_final_single_pages, Nov. 2024.

[3] "Erasmus+ at the University of Patras – Statistics," https://www.upatras.gr/international-relations/erasmus/programma-erasmus/erasmus-sto-panepistimio-patron/.

[4] G. K. Zipf, "Human behavior and the principle of least effort." *Journal of Clinical Psychology*, vol. 6, no. 3, pp. 306–306, 1950.

[5] A. Følstad and P. B. Brandtzæg, "Chatbots and the new world of HCI," *Interactions*, vol. 24, no. 4, pp. 38–42, Jun. 2017.

[6] M. McTear, *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. Springer Nature, May 2022.

[7] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1966.

[8] K. M. Colby, S. Weber, and F. D. Hilf, "Artificial Paranoia," *Artificial Intelligence*, vol. 2, no. 1, pp. 1–25, Mar. 1971.

[9] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Machine Learning with Applications*, vol. 2, p. 100006, Dec. 2020.

[10] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd ed., 2026.

[11] A. Alabbas and K. Alomar, "A Weighted Composite Metric for Evaluating User Experience in Educational Chatbots: Balancing Usability, Engagement, and Effectiveness," *Future Internet*, vol. 17, no. 2, p. 64, Feb. 2025.

[12] D. Tolis, S. Mystakidis, I. Hatzilygeroudis, and K. Siozopoulos, "Generative AI Agents for Instructional Co-design: A Sequential Agent-Based Approach Using a Low-Code/No-Code Platform," in *Generative Systems and Intelligent Tutoring Systems*, S. Graf and A. Markos, Eds. Cham: Springer Nature Switzerland, 2026, pp. 301–306.

[13] T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol, "Rasa: Open Source Language Understanding and Dialogue Management," Dec. 2017.

[14] S. Holmes, A. Moorhead, R. Bond, H. Zheng, V. Coates, and M. Mctear, "Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?" in *Proceedings of the 31st European Conference on Cognitive Ergonomics*, ser. ECCE '19. New York, NY, USA: Association for Computing Machinery, Sep. 2019, pp. 207–214.

[15] S. Borsci and M. Schmettow, "Re-examining the chatBot Usability Scale (BUS-11) to assess user experience with customer relationship management chatbots," *Personal and Ubiquitous Computing*, vol. 28, no. 6, pp. 1033–1044, Dec. 2024.

[16] C. G. Møller, K. E. Ang, M. de Lourdes Bongiovanni, M. S. Khalid, and J. Wu, "Metrics of Success: Evaluating User Satisfaction in AI Chatbots," in *Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence*, ser. ICAAI '24. New York, NY, USA: Association for Computing Machinery, Mar. 2025, pp. 168–173.

[17] L. V. de Carvalho, P. H. D. Valle, B. R. Leifheit, L. E. R. Cabrejos, W. Nakamura, G. C. Guerino, R. D. S. Garcia, and W. Silva, "What Do We Know About Usability Evaluation for Chatbots?: A Systematic Mapping Study," in *Proceedings of the 20th Brazilian Symposium on Information Systems*, ser. SBSI '24. New York, NY, USA: Association for Computing Machinery, May 2024, pp. 1–11.

[18] P. K. Bodigutla, L. Wang, K. Ridgeway, J. Levy, S. Joshi, A. Geramifard, and S. Matsoukas, "Domain-Independent turn-level Dialogue Quality Evaluation via User Satisfaction Estimation," Aug. 2019.

[19] Y.-C. Lin, J. Neville, J. W. Stokes, L. Yang, T. Safavi, M. Wan, S. Counts, S. Suri, R. Andersen, X. Xu, D. Gupta, S. K. Jauhar, X. Song, G. Buscher, S. Tiwary, B. Hecht, and J. Teevan, "Interpretable User Satisfaction Estimation for Conversational Systems with Large Language Models," Jun. 2024.

[20] R. Ren, M. Zapata, J. W. Castro, O. Dieste, and S. T. Acuña, "Experimentation for Chatbot Usability Evaluation: A Secondary Study," *IEEE Access*, vol. 10, pp. 12 430–12 464, 2022.

[21] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in *Human Mental Workload*, ser. Advances in Psychology, P. A. Hancock and N. Meshkati, Eds. North-Holland, Jan. 1988, vol. 52, pp. 139–183.

[22] H.-Y. Kim, "Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis," *Restorative Dentistry & Endodontics*, vol. 38, no. 1, pp. 52–54, Feb. 2013.

[23] J. Schmidhuber, S. Schlögl, and C. Ploder, "Cognitive Load and Productivity Implications in Human-Chatbot Interaction," in *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, Sep. 2021, pp. 1–6.