

From RAG Metrics to Quality of Experience and Trust: An Extended TAM Study of a University Chatbot

NELA DURANEC¹, and EMIL DUMIC² (Member, IEEE)

¹Department of Multimedia, University North, 104. brigade 3, 42000 Varaždin, Croatia (e-mail: neladuranec@gmail.com)

²Department of Electrical Engineering, University North, 104. brigade 3, 42000 Varaždin, Croatia (e-mail: edumic@unin.hr)

Corresponding author: Emil Dumić (e-mail: edumic@unin.hr).

This research was conducted within the project “Application of 3D LiDAR Point Clouds for Measuring the Efficiency of Solar Energy Systems (3D-SOLAR)” (IP-UNIN-TEH-2025-9), funded by the European Union – NextGenerationEU.

ABSTRACT Retrieval-Augmented Generation (RAG) systems are increasingly used in conversational AI applications; however, the relationship between automated RAG evaluation metrics and users’ subjective perceptions remains insufficiently understood. This study investigates technology acceptance of a RAG-based university chatbot by extending the Technology Acceptance Model (TAM) with Quality of Experience (QOE) and trust (TR), while examining whether Retrieval-Augmented Generation Assessment (RAGAS) metrics can predict users’ subjective evaluations. The proposed framework extends the traditional technology acceptance perspective by jointly considering perceived interaction quality and trust in AI-generated responses. The model was evaluated using questionnaire responses from 90 participants who interacted with a RAG-based chatbot integrated into the UNINwalk university application. Subjective responses were analyzed using reliability, factor, correlation, and regression analyses, while quantitative evaluation employed four RAGAS metrics: faithfulness, answer relevance, context precision, and a derived overall score. The results demonstrated that QOE and TR were significant predictors of behavioral intention (BI). Answer relevance exhibited the strongest correlations with QOE and TR. Regression models further showed that answer relevance achieved root mean square error improvements of up to 6.3% over a baseline predictor in predicting perceived QOE and TR, whereas no meaningful predictive performance was observed for BI. These findings suggest that answer relevance may provide useful complementary information about perceived QOE and TR during the evaluation and development of educational AI chatbots.

INDEX TERMS Retrieval-Augmented Generation (RAG), Retrieval-Augmented Generation Assessment (RAGAS), Quality of Experience (QOE), Trust, Technology Acceptance Model (TAM), AI Chatbots, Higher Education

I. INTRODUCTION

RECENT advances in large language models (LLMs) have accelerated the adoption of conversational AI systems across numerous application domains, including higher education. Among these systems, Retrieval-Augmented Generation (RAG) [1] has emerged as an effective approach for reducing hallucinations by grounding generated responses in retrieved external knowledge. As a result, RAG-based chatbots are increasingly being deployed to provide students with information about academic programs, university services, campus facilities, and administrative procedures. While evaluation frameworks such as Retrieval-Augmented Generation Assessment (RAGAS) [2] enable automated assessment of retrieval and generation quality, it remains unclear to what extent

these metrics correspond to users’ subjective perceptions of chatbot quality and trust.

Technology acceptance has traditionally been explained using the Technology Acceptance Model (TAM) [3], which proposes that perceived ease of use (PEOU) and perceived usefulness (PU) influence users’ behavioral intention to use (BI) a system, ultimately leading to actual system use (AU). The Unified Theory of Acceptance and Use of Technology (UTAUT) [4] further extends this perspective by incorporating additional individual and contextual determinants of technology adoption. Recent studies have continued to extend technology acceptance frameworks by incorporating context-specific factors when examining the adoption of emerging AI-enabled technologies [5]. Although both models have

been successfully applied to numerous information systems, AI-based conversational systems introduce additional user-centered factors, particularly trust (TR) [6], [7] in generated responses and Quality of Experience (QOE) [8], which have been recognized as important determinants of users' perceptions but are not explicitly represented in the original TAM framework.

In this study, the original TAM is extended by incorporating three external constructs, self-efficacy (SE), system accessibility (SA), and earlier experience (EE), together with QOE and TR, while omitting attitude toward using (ATT) and actual system use (AU). The omission of ATT is consistent with later extensions of TAM, particularly TAM2 [9], whereas AU was omitted because the system had not yet been deployed for sustained regular use. The proposed model examines how SE, SA, and EE influence PEOU and PU, how these constructs contribute to QOE, how QOE influences TR, and how QOE and TR explain users' BI. The incremental contribution of QOE and TR beyond PU to explaining BI is further assessed using hierarchical regression. In addition to evaluating technology acceptance, this work investigates whether automated RAGAS metrics can serve as indicators of users' perceived QOE, TR, and BI.

To evaluate the proposed model, a RAG-based university chatbot was integrated into the UNINwalk web application developed for University North. The chatbot employed a hybrid retrieval architecture, while its performance was assessed using four RAGAS metrics: faithfulness, answer relevance, context precision, and a derived overall score. Subjective user evaluations were collected from 90 participants using a questionnaire based on the proposed TAM extension, enabling direct comparison between automated RAG evaluation metrics and users' perceptions. Although validated using the UNINwalk application, the proposed model is applicable to the evaluation of domain-specific RAG-based conversational systems. It provides a unified framework that integrates an extended TAM with RAG evaluation metrics to investigate their relationships with subjective user perceptions.

A similar objective-subjective evaluation approach has been widely adopted in multimedia applications, particularly in image and video quality assessment, where objective quality metrics are routinely compared with human perception [10]–[13]. Inspired by this approach, the present study investigates whether automated RAG evaluation metrics are associated with users' subjective perceptions of conversational AI.

The main contributions of this work are as follows:

- An extended TAM for RAG-based university chatbots that incorporates QOE and TR while replacing the original ATT and AU constructs. The proposed framework is evaluated through five groups of hypotheses covering external factors, technology acceptance, user experience, trust, and behavioral intention.
- A joint evaluation framework for a RAG-based university chatbot combining subjective questionnaire responses and automated RAGAS metrics collected from

90 participants interacting with the UNINwalk application.

- An analysis of the relationship between automated RAG evaluation metrics and users' subjective perceptions, including correlation analysis and nested cross-validation regression models to investigate the predictive capability of RAGAS metrics for estimating QOE and TR.

The remainder of this paper is organized as follows. Section II reviews related work on university chatbots, RAG evaluation, and Technology Acceptance Models. Section III presents the UNINwalk application and its RAG-based chatbot architecture. Section IV introduces the proposed extended TAM and describes the survey methodology. Section V presents the relationships between RAG evaluation metrics and subjective user perceptions through correlation and regression analyses. Section VI discusses the findings, and Section VII concludes the paper and outlines directions for future work.

II. RELATED WORK

Previous research relevant to this study spans three complementary areas. First, studies on educational chatbots and university support services provide the application context and identify key factors influencing user acceptance. Second, research on RAG metrics and QOE assessment provides the basis for investigating relationships between automated evaluation metrics and users' subjective perceptions. Finally, studies extending TAM with trust (TR) provide the theoretical foundation for the proposed research model.

A. CHATBOTS IN HIGHER EDUCATION AND UNIVERSITY SUPPORT SERVICES

Research on educational chatbots has expanded from conversational tutoring toward broader support for learning, administration, and student services. Review studies show that chatbots are used for course assistance, homework support, personalized learning, skill development, and educator support, while also highlighting limitations related to reliability, accuracy, ethical use, and evaluation consistency [14]–[16]. Recent reviews of generative AI chatbots in higher education indicate that the field remains conceptually fragmented, with diverse assumptions about learning theories and mixed views on the institutional role of AI systems [17]. Studies on student motivation report that chatbot interaction is positively associated with engagement, interest, and intrinsic motivation, although long-term effects remain insufficiently examined [18].

A specific part of this literature addresses chatbots for university support services. Reviews show that such systems are used for student administration, recruitment, enrollment, teaching support, and FAQ-like services, while also noting that closed-domain datasets are difficult to obtain and curate [19]. Domain-specific systems therefore rely on institutional data and tailored architectures. Hybrid approaches combine rule-based components, machine learning, and large language models to answer questions about academic programs, ad-

missions, and student life [20]. RAG-based chatbots extend this approach by retrieving information from institutional resources and generating answers about departments, programs, facilities, and student services [21]. Similar designs have been applied in learning management systems, supporting self-regulated learning and help-seeking using course-specific content [22].

Empirical work also examines chatbot use in online and course-based learning. Chatbots support goal setting, social presence, help-seeking, self-evaluation, and listening practice, with evaluations based on behavioral engagement, perceived usefulness, perceived ease of use, and qualitative feedback [23], [24]. Course-assistant chatbots evaluated using TAM show that perceived usefulness influences intention to use, while continuous availability and ease of use are valued by students [25]. Recent studies further show that behavioral intention and actual use of generative AI in higher education are shaped by technology acceptance factors, digital literacy, and broader behavioral and institutional factors [26], [27]. Broader studies on AI use in higher education report benefits such as personalized learning and increased engagement but also concerns, including over-reliance, reduced critical thinking, privacy risks, and academic dishonesty [28], [29].

The smart-campus perspective extends chatbot use beyond classroom learning. AI-based architectures utilize student data to support active learning and personalized recommendations [30]. Chatbot services based on the Internet of Educational Things (IoET) integrate environmental data and acceptance models to support health, safety, and continued usage [31].

Overall, prior work indicates that university chatbots should be evaluated not only as technical systems but also as student-facing services influenced by usefulness, ease of use, trust, perceived experience, and institutional context.

B. FROM AUTOMATED RAG EVALUATION TO PERCEIVED QOE

RAG-based chatbots integrate retrieval and generation to provide domain-specific, context-grounded answers, reducing hallucinations in educational applications [32]. However, their evaluation requires assessing both retrieval and answer quality.

The RAGAS framework introduces reference-free metrics for evaluating RAG pipelines [2]. In this work, faithfulness, answer relevance, and context precision were selected to capture complementary aspects of response grounding, answer relevance, and retrieval quality, respectively. Three RAGAS metrics and one derived score are considered:

- Faithfulness (RAGAS1), defined as the alignment of the answer with the retrieved context;
- Answer relevance (RAGAS2), measuring how well the answer addresses the query;
- Context precision (RAGAS3), representing the proportion of relevant retrieved information;
- Overall score (RAGAS4), a derived measure calculated in this study as the unweighted arithmetic mean of faith-

fulness, answer relevance, and context precision.

The overall score is therefore calculated as follows:

$$\text{RAGAS4} = \frac{\text{RAGAS1} + \text{RAGAS2} + \text{RAGAS3}}{3}. \quad (1)$$

Higher overall score values indicate better performance across the three evaluated dimensions. Context Recall was not included because its calculation requires ground-truth reference answers, which were not available for the student queries in this study.

Existing studies show that while such metrics are useful, they do not fully capture human perceptions of answer quality. Some metrics correlate with human judgments, but limitations remain, especially in domain-specific settings and under varying retrieval conditions [33], [34]. In educational applications, RAG-based chatbots are typically evaluated using a combination of automated evaluation metrics and user-centered measures such as usability, satisfaction, and latency analysis [21], [22], [35].

In conversational AI, QOE provides a subjective evaluation framework for chatbot interaction, including perceived usefulness, clarity, and satisfaction [8]. User experience research further emphasizes that users evaluate interactive systems not only based on functionality and performance but also through subjective experiential aspects of interaction [36]. Since system performance does not directly translate into perceived quality, mapping RAGAS metrics to QOE dimensions enables system-level evaluation to be linked with user perception. While previous studies have primarily examined relationships between automated RAG metrics and human assessments of answer quality, their relationships with broader user-centered constructs such as QOE remain largely unexplored.

C. TRUST AND EXTENDED TECHNOLOGY ACCEPTANCE FACTORS IN AI CHATBOT ADOPTION

TR has been identified as an important extension of TAM in digital environments. BI is influenced not only by PU and PEOU, but also by TR, particularly in contexts with limited human interaction [37]. This is directly relevant for AI chatbots, where users must assess the reliability and credibility of generated responses.

Recent studies extend TAM and UTAUT frameworks with constructs tailored to AI systems. In chatbot adoption, attitude and BI predict usage, while TR, anthropomorphism, and design-related factors influence attitude formation [38]. Other extensions show that PU remains a key determinant of attitude, while PEOU may be less influential for experienced users [39]. Ethical considerations and TR also act as direct adoption factors. TR also positively influences PU, both directly and through AI-EdTech readiness [40].

These findings support modeling TR as a core construct in AI chatbot adoption. In university settings, TR is closely related to the perceived correctness and reliability of chatbot responses, supporting its integration into extended TAM alongside PU, PEOU, QOE, and BI. This also suggests

a potential connection between automated RAG evaluation metrics and user trust. However, empirical evidence linking RAG evaluation metrics to QOE, TR, and BI remains limited, particularly in educational chatbot settings.

III. UNINWALK APPLICATION DESIGN

UNINwalk is a web-based information and navigation assistant designed for students of University North. The application integrates three core modules: a campus navigation system with interactive floor plans, a searchable professor directory, and a chatbot. Together, these modules aim to reduce the friction students experience when orienting themselves on campus, finding a professor's contact information, or answering general questions about university services. Because the application targets first-time and infrequent users, most notably incoming students, its design was guided by two goals: minimizing the number of steps between a question and its answer, and making the system's behavior transparent enough to be trusted. The latter goal directly motivates the user-facing evaluation mechanism described in Section III-E and the analysis of its relationship to perceived trust presented in Section V.

A. SYSTEM ARCHITECTURE

The application follows a client-server architecture. The backend is implemented using FastAPI and exposes a RESTful API serving three static JSON datasets: professor information, room information including pixel coordinates for each room in the two campus buildings, and institutional knowledge covering rooms, parking, and general university information. The backend is containerized with Docker and deployed on a university server alongside a static build of the frontend.

The frontend is implemented as a Single Page Application (SPA) built with Vue 3, with application state managed by Pinia and client-side routing handled by Vue Router. Rather than adopting a third-party UI framework, the interface is built from custom Vue single-file components styled with scoped SCSS modules; this gives full control over the visual language, interaction patterns, and accessibility of every component while ensuring a consistent interface across all three modules. Two additional design choices improve accessibility and usability: the interface supports both Croatian and English through the `i18n` library and provides light and dark visual themes, with automatic fallback to the operating system preference. Both the language and theme choices persist across sessions in the browser's `localStorage`, so a returning student is never asked to reconfigure the application.

The application consists of four main routes: a home view that hosts the chatbot and an interactive campus overview, separate floor plan views for each university building, and a professor directory.

B. CAMPUS NAVIGATION

The navigation module allows students to locate rooms, facilities, and points of interest across the university's buildings. The endpoint `/unin-data` returns the complete room dataset, which the frontend uses to render the floor plans. Each room entry in the dataset includes a building identifier, a room name and type (e.g., lecture hall, computer lab, student registry, or library), and pixel-level coordinates (x , y) and dimensions.

The floor plans are rendered as Scalable Vector Graphics (SVGs) with a fixed `viewBox` coordinate system rather than as static raster images. This choice makes every room a first-class interactive element: each room is drawn as a labeled, filled rectangle, and selecting a room opens a modal overlay with room details and the list of professors assigned to it. The user can thus move directly from a location on the map to the people associated with it, without switching modules or re-entering a search. The vector representation also scales without loss of legibility on mobile screens, where most on-campus lookups are expected to occur.

The home view additionally contains a campus-level SVG that visualizes all buildings from an isometric perspective. When the user grants geolocation permission, the application reads the device position via the Geolocation API and maps it to SVG coordinates using an affine transformation computed from a set of known GPS-to-SVG reference points, placing an animated marker on the map. This gives newcomers an immediate answer to the most basic orientation question "where am I?", before they formulate an explicit query.

Beyond lecture rooms, the spatial dataset includes the full range of destinations students may need to locate, including toilets, both campus parking lots, and key university services across both campuses, such as student affairs offices, the library, IT support, the finance department, and the EU funds office. This information is surfaced both through the navigation interface and through the chatbot, so the same underlying data answers both a visual "show me" interaction and a conversational "tell me" interaction.

C. PROFESSOR DIRECTORY

The professor directory provides students with contact and office information for all academic staff members. Each entry includes the professor's full name, academic title, institutional email address, phone number, a link to their official university profile page, and a profile image. The directory supports real-time search with multi-word, case-insensitive matching, so results narrow with every keystroke and no explicit search action is required.

Where available, entries also include an office room number and the corresponding building identifier, enabling the frontend to link directly from a professor's profile to their office location on the floor plan. This cross-linking mirrors the map-to-professor path described in the previous subsection in the opposite direction: whether a student starts from a person or from a place, the other is always one interaction away. The directory is exposed through two endpoints: one for listing all

professors and another for retrieving a specific professor by ID.

D. CHATBOT

The chatbot is implemented as a RAG system using the LangChain framework. This approach enables domain-specific answers about professors, rooms, and services based on institutional data, without requiring a custom-trained model.

At application startup, the system indexes two data sources: the professor dataset and the institutional knowledge dataset. The institutional data are structured as individual professor and room records and converted into separate LangChain documents with metadata identifying their type. Because each record represents a compact semantic unit, no additional text splitting or overlapping chunking is applied. The indexing strategy also accounts for the frequent omission of Croatian diacritical characters in student queries, particularly on mobile keyboards. Professor records are therefore represented using multiple name variants, including the original and natural name order and ASCII-normalized forms.

Retrieval uses a hybrid strategy combining a Facebook AI Similarity Search (FAISS) semantic retriever, based on the sentence-transformers/all-MiniLM-L6-v2 embedding model, with an Okapi Best Matching 25 (BM25) keyword retriever. The FAISS retriever uses maximal marginal relevance (MMR) with Euclidean distance, returning eight documents from an initial pool of 20 candidates ($k = 8$, $\text{fetch}_k = 20$, $\lambda = 0.7$). The BM25 retriever also returns eight documents. The two retrievers are combined using equal weights (0.5/0.5), after which duplicate professor records are removed before the retrieved context is provided to the language model. Consequently, the final number of documents supplied as context is query-dependent rather than fixed. The language model used for response generation is Llama 3.3-70B-Versatile, served through Groq's inference API, with the temperature set to 0. No explicit maximum output length is set by the application.

The English-language system prompt restricts the chatbot to university-related topics and instructs it to respond in the language of the user's question (Croatian or English). It also integrates responses with the navigation module: room numbers are rendered as interactive links that trigger in-app routing, while email addresses, phone numbers, and external URLs are formatted as their respective HTML link types.

Two further design choices target perceived responsiveness and continuity. The chatbot delivers its responses token by token using Server-Sent Events (SSEs), allowing the frontend to display the answer progressively as it is generated. Chat history persists across sessions in the browser's `localStorage` to preserve the conversation in the user interface. However, the stored chat history is not incorporated into retrieval or response generation; both operate solely on the current user query.

E. RESPONSE EVALUATION AND TRANSPARENCY

A distinguishing design decision of UNINwalk is that the quality of chatbot responses is evaluated automatically and the resulting scores are shown to the end user. The evaluation is based on the RAGAS dimensions described in Section II and is implemented as a custom LLM-as-a-judge pipeline using the same Llama 3.3-70B-Versatile model. The judge instructions are written in Croatian to match the application's dominant language context, while user queries and chatbot responses are evaluated in their original language. Faithfulness, answer relevance, and context precision are evaluated separately for each chatbot response. After each response, the scores for each metric are averaged across all responses in the conversation up to that point, producing three aggregated conversation-level values. The overall score is calculated as the arithmetic mean of these three values.

The four evaluation scores are displayed in a collapsible evaluation panel, which is collapsed by default to avoid distracting users who simply want an answer, but can be expanded by users wishing to assess the quality of the chatbot interaction. The rationale for this mechanism is to operationalize transparency as a candidate trust-building feature: rather than asking students to take the assistant's answers in good faith, the interface exposes the system's own assessment of response quality. The final conversation-level scores, reflecting all responses generated during a respondent's interaction, were subsequently used to examine their relationships with the respondent's subjective evaluations of QOE, TR, and BI in Section V. However, whether individual users expanded the evaluation panel was not recorded.

IV. EXTENDED TAM USING QOE AND TR

A. RESEARCH MODEL AND HYPOTHESES

The proposed research model extends the original TAM [3] by incorporating three external constructs, the QOE construct, and a chatbot-specific TR construct. In contrast to the original TAM, ATT and AU were not included as separate constructs. The omission of ATT is consistent with later extensions of TAM, particularly TAM2, in which behavioral intention is modeled without ATT as a separate mediating construct [9]. QOE was introduced as an additional experience-oriented construct reflecting users' perceived quality of interaction with the chatbot and is not considered conceptually equivalent to ATT. AU was omitted because UNINwalk had not yet been deployed for sustained regular use at the time of the study; participants interacted with the application within the study setting, making BI a more appropriate acceptance outcome than actual system use. Consequently, the proposed model should be interpreted as explaining students' perceived interaction quality, trust, and intention to use the chatbot following their initial experience with the system, rather than their sustained post-adoption behavior.

The core TAM constructs include the PEOU construct, which is defined as the degree to which a user believes that interacting with the application would be effortless, and the PU construct, referring to the extent to which the application

is perceived as beneficial for accomplishing academic tasks. The QOE construct represents users' overall evaluation of their interaction with the application, including satisfaction and fulfillment of expectations. The TR construct captures users' confidence in the reliability, competence, and credibility of the chatbot-generated responses. The BI construct represents users' intention to continue using the proposed application in the future.

Three external constructs were included to capture individual and system-related factors that may influence technology acceptance. Self-efficacy (SE) represents users' confidence in their ability to learn and use similar applications independently. System accessibility (SA) reflects the perceived accessibility and technical availability of the proposed application, while earlier experience (EE) captures respondents' prior experience with chatbots and AI-based conversational systems.

In this study, the term application refers to the UNINwalk web application together with its integrated chatbot functionality. Consequently, PEOU, PU, QOE, and BI are used to evaluate perceptions of the overall application. TR represents trust in the chatbot component of the application. SE, SA, and EE are modeled as external factors, with EE specifically referring to prior chatbot experience, while SE and SA represent user characteristics. The resulting research model is illustrated in Figure 1.

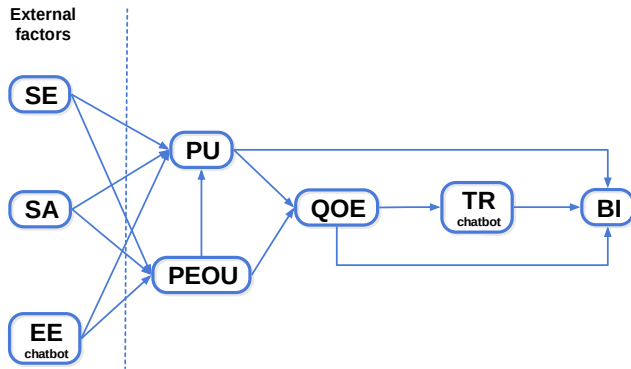


FIGURE 1. Proposed research model extending the Technology Acceptance Model (TAM). SE = Self-Efficacy, SA = System Accessibility, EE = Earlier Experience (with chatbot), PEOU = Perceived Ease of Use, PU = Perceived Usefulness, QOE = Quality of Experience, TR = Trust (in chatbot), BI = Behavioral Intention.

Prior research on LLM-driven educational chatbots has identified SE, SA, and prior chatbot experience as external variables that may influence PEOU and PU [22]. In addition, the established TAM relationship between PEOU and PU was retained [3], [9]. Accordingly, SE, SA, and EE were expected to positively influence PEOU, while SE, SA, EE, and PEOU were expected to positively influence PU. These relationships were examined both separately and jointly. Therefore, the following hypotheses were formulated:

- H1a: SE positively influences PEOU.
- H1b: SA positively influences PEOU.
- H1c: EE positively influences PEOU.

- H1d: PEOU positively influences PU.
- H1e: SE positively influences PU.
- H1f: SA positively influences PU.
- H1g: EE positively influences PU.

User experience research distinguishes the overall experience of interacting with a system from its pragmatic qualities, including usability and utility, while recognizing that these qualities can contribute to a positive user experience [36]. Consistent with this perspective, recent chatbot research has shown that both perceived ease of use and perceived usefulness positively influence users' chatbot experience [41]. Accordingly, PEOU and PU were expected to contribute positively to users' overall QOE. The second research question examined their individual and joint contributions to QOE. The following hypotheses were proposed:

- H2a: PEOU positively influences QOE.
- H2b: PU positively influences QOE.

Previous research on AI chatbot use has identified customer experience as an antecedent of trust in the chatbot [42]. Since QOE in the present study captures users' overall evaluation of their interaction with UNINwalk, a more positive QOE was expected to contribute to greater TR in its chatbot component. Therefore, the following hypothesis was formulated:

- H3: QOE positively influences TR.

Positive evaluations of chatbot interactions and trust in the system have been identified as important antecedents of users' intention to continue using chatbot services. In particular, previous research on AI chatbots in higher education has shown that both user satisfaction and perceived trust positively influence continuance intention [43], while trust has also been linked to reuse intention in the context of generative AI chatbots [42]. In the proposed model, BI was considered the primary outcome variable. The fourth research question examined whether QOE and TR separately and jointly influence students' BI regarding the proposed application:

- H4a: QOE positively influences BI.
- H4b: TR positively influences BI.

Finally, consistent with TAM, PU was expected to positively influence BI [3], [9]. A hierarchical regression analysis was then performed to determine whether QOE and TR provide additional explanatory power beyond PU when predicting BI:

- H5a: PU positively influences BI.
- H5b: QOE explains additional variance in BI beyond PU.
- H5c: TR explains additional variance in BI beyond PU and QOE.

B. SURVEY INSTRUMENT

The survey instrument was designed based on established constructs from the Technology Acceptance Model (TAM) and related extensions identified in the prior literature. The questionnaire consisted of 21 items grouped into five primary

constructs (PEOU, PU, QOE, TR, and BI) and three external-variable constructs (SE, SA, and EE). It was administered bilingually in English and Croatian to university students after they had interacted with the proposed chatbot application. The questionnaire items are presented in Table 1. Responses were recorded using a five-point Likert scale: 1 (strongly disagree), 2 (disagree), 3 (neither agree nor disagree), 4 (agree), and 5 (strongly agree).

The PEOU construct was measured using three items adapted from [3]. The items assessed ease of learning, clarity of interaction, and ease of navigation within the proposed application. The PU construct was measured using three items adapted from [3]. These items evaluated whether the application could improve information retrieval, efficiency, and support users' academic needs. The QOE construct was measured using three items based on established QOE and the user experience literature [36], [44], [45]. The items captured overall user experience, expectation fulfillment, and perceived positivity of interaction with the application. The TR construct was measured using three items adapted from prior trust and technology acceptance studies [37], [46], [47]. The construct evaluated trust in the chatbot's information, perceived reliability, and perceived competence. The BI construct was measured using three items adapted from prior technology acceptance studies [4], [47], [48]. The items assessed future usage intentions, frequency of intended use, and willingness to rely on recommendations generated by the application.

The SE external construct was measured using two items adapted from Compeau and Higgins [49]. The construct assessed respondents' confidence in learning to use similar applications independently and their perceived ability to solve problems encountered during usage. The SA external construct was measured using two items based on the information systems success literature [50]. The construct evaluated the perceived accessibility of the proposed application and the reliability of its operation. The EE external construct was measured using two items adapted from prior technology acceptance research [51]. The construct assessed respondents' prior experience with chatbots and the frequency with which they use AI-based conversational systems.

The questionnaire items were selected from established instruments and prior studies based on their conceptual correspondence with the constructs examined in the proposed model. Where necessary, the wording was adapted to the context of the UNINwalk application and AI chatbot interaction while preserving the intended meaning of the original items. A formal content validity assessment, such as expert review or a content validity index, was not conducted prior to data collection.

C. SAMPLE CHARACTERISTICS AND DESCRIPTIVE STATISTICS

Participants were recruited by e-mail through courses within the Multimedia, Electrical Engineering, Mechatronics, and Civil Engineering study programs. The study was conducted

remotely and without supervision. The e-mail invitation included a link to the questionnaire, which provided a link to the UNINwalk application at the beginning of the survey. Participants were instructed to interact with the chatbot before proceeding with the questionnaire. They were free to formulate their own questions rather than complete standardized tasks, and no minimum interaction duration or number of interactions was prescribed. Individual chatbot use was not independently verified by the researchers. Participation in the study was voluntary. Before starting the questionnaire, participants were informed about the purpose of the study. Completion of the questionnaire was considered to indicate informed consent. A total of 96 survey responses were collected.

Prior to statistical analysis, six responses were excluded due to incomplete RAGAS evaluations, resulting in a final sample of 90 respondents used for further analysis. Participant characteristics are summarized in Table 2. The sample included students from different study levels and disciplines; however, as participants were recruited from a single university, the sample should not be considered representative of the broader university student population.

Table 3 summarizes the descriptive statistics for all questionnaire items. Overall, the highest mean values were observed for PEOU, SE, and SA (all greater than 4), indicating generally positive perceptions of the evaluated system. In contrast, BI items showed comparatively lower mean values and greater variability, suggesting more heterogeneous user intentions regarding future system usage. The final respondent-level RAGAS values described in Section III-E were subsequently used in analyses examining their relationships with subjective user evaluations.

D. EXPLORATORY FACTOR ANALYSIS

Exploratory factor analysis (EFA) was conducted using unweighted least squares (ULS) extraction with oblimin rotation. The sample included 90 participants for 15 questionnaire items, corresponding to a participant-to-item ratio of 6:1. Sampling adequacy for EFA was assessed using the Kaiser–Meyer–Olkin (KMO) measure [52]. The overall KMO value was 0.855, indicating meritorious sampling adequacy. Bartlett's test of sphericity was statistically significant ($\chi^2 = 966.55$, $p < 0.001$), further confirming the suitability of the correlation matrix for factor analysis. Construct-level reliability was good, with Cronbach's α values ranging from 0.760 to 0.922, as summarized in Table 4.

Parallel analysis suggested a two-factor solution, whereas the eigenvalue-greater-than-one criterion indicated three factors, as shown in Figure 2. The eigenvalues for the first five factors were 7.408, 1.708, 1.341, 0.962, and 0.793.

Given the differing indications from the factor-retention criteria and the theoretically specified five-construct framework, two-, three-, and five-factor solutions were examined and compared using the same ULS extraction method and oblimin rotation, as shown in Tables 5 to 7. The two-factor solution explained 51.67% of the total variance but provided

TABLE 1. Questionnaire items used in the study.

Construct	Item	Questionnaire item
PEOU	PEOU1	Learning to use the proposed application would be easy for me.
	PEOU2	My interaction with the proposed application would be clear and understandable.
	PEOU3	I would find the proposed application easy to navigate.
PU	PU1	Using the proposed application would enable me to find university-related information more quickly.
	PU2	Using the proposed application would improve my efficiency when searching for information.
	PU3	I would find the proposed application useful for my academic needs.
QOE	QOE1	I am delighted with my overall experience using the proposed application.
	QOE2	The proposed application fulfilled my expectations.
	QOE3	Using the proposed application provided a positive user experience.
TR	TR1	I trust the information provided by the chatbot within the application.
	TR2	I believe the chatbot is reliable and unbiased.
	TR3	I believe the chatbot is competent in providing the information I need.
BI	BI1	I intend to use the proposed application in the future.
	BI2	I plan to use the proposed application frequently.
	BI3	I plan to make decisions based on the recommendations provided by the proposed application.
SE	SE1	I am confident that I could learn to use similar applications on my own.
	SE2	I am capable of solving most problems encountered when using such applications.
SA	SA1	I can easily access the proposed application using the devices I normally use.
	SA2	The proposed application generally operates reliably without interruptions.
EE	EE1	I have prior experience using chatbots.
	EE2	I frequently use AI-based chatbots to obtain information.

TABLE 2. Participant Characteristics.

Characteristic	Category / Statistic	Value
Age (years)	Mean $\pm$ SD	22.72 $\pm$ 3.59
	Range	20–46
Gender	Male	59 (65.56%)
	Female	31 (34.44%)
Study level	Undergraduate	39 (43.33%)
	Graduate	51 (56.67%)
Study discipline	Multimedia	51 (56.67%)
	Electrical Engineering	21 (23.33%)
	Mechatronics	16 (17.78%)
	Civil Engineering	2 (2.22%)

limited construct differentiation, with PU, QOE, TR, and BI predominantly loading on the same factor, while PEOU formed a separate factor. The three-factor solution explained 54.24% of the total variance and provided greater differentiation; however, PU and BI predominantly loaded on the same factor, while QOE exhibited substantial cross-loadings and did not form a distinct factor. The five-factor solution explained 61.29% of the total variance, with the five factors corresponding to TR, QOE, BI, PU, and PEOU.

Based on the comparison of the alternative factor solutions, the five-factor solution was retained because it provided greater construct differentiation and interpretability, while also being consistent with the proposed theoretical framework. As shown in Table 7, all items exhibited acceptable loadings on their expected factors, with no substantial cross-

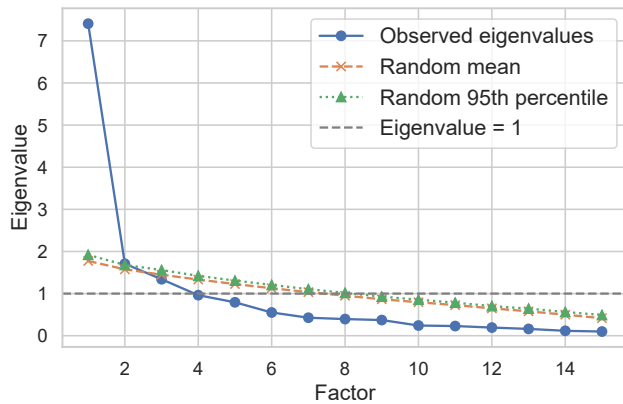
TABLE 3. Descriptive statistics of survey items.

Item	Mean	Standard Deviation
PEOU1	4.522	0.691
PEOU2	4.389	0.817
PEOU3	4.544	0.690
PU1	4.233	0.937
PU2	4.122	0.910
PU3	3.856	1.107
QOE1	3.667	1.039
QOE2	3.833	1.052
QOE3	4.056	0.952
TR1	3.856	1.023
TR2	3.856	0.943
TR3	3.811	1.059
BI1	3.033	1.146
BI2	2.589	1.150
BI3	3.011	1.268
SE1	4.544	0.673
SE2	4.244	0.852
SA1	4.656	0.523
SA2	4.411	0.748
EE1	4.356	0.916
EE2	3.811	1.208
RAGAS1	0.559	0.331
RAGAS2	0.648	0.263
RAGAS3	0.426	0.316
RAGAS4	0.544	0.257

loadings observed. The retained factor structure, together with the internal consistency results, supported the suitability of the survey instrument for subsequent analyses.

TABLE 4. Reliability and factorability assessment.

Construct	Cronbach α	KMO	Bartlett χ^2	p-value
PEOU	0.760	0.659	70.06	< 0.001
PU	0.866	0.723	137.13	< 0.001
QOE	0.922	0.751	200.69	< 0.001
TR	0.900	0.740	170.92	< 0.001
BI	0.870	0.721	138.23	< 0.001
Overall	–	0.855	966.55	< 0.001

**FIGURE 2.** Scree plot with parallel analysis for the proposed TAM-based model.

E. REGRESSION ANALYSIS

Regression analyses were conducted to test the proposed hypotheses. Simple regression was used for direct relationships, multiple regression was used to examine the joint effect of QOE and TR on BI, and hierarchical regression was used to evaluate whether QOE and TR explained additional variance in BI beyond PU. The reported statistics include standardized regression coefficients (β), corresponding t -statistics, significance levels (p), coefficients of determination (R^2), overall model F -tests for multiple regression models, and changes in explained variance (ΔR^2) for hierarchical models.

To assess the adequacy of the available sample for the regression analyses, a power analysis was conducted using G*Power 3.1.9.7 [53] for the largest multiple regression model, which included four predictors. Assuming a medium effect size of Cohen's $f^2 = 0.15$ [54], corresponding to $R^2 \approx 0.13$, $\alpha = 0.05$, and statistical power of 0.80, the required sample size was 85 participants. The final sample of $N = 90$ therefore exceeded the required sample size for detecting a medium-sized overall regression effect. Given the available sample size and the complexity of the proposed framework, a regression-based approach was adopted to evaluate the hypothesized relationships, while integrated structural equation modeling (SEM) is reserved for future validation using a larger independent sample.

All joint multiple regression models were statistically significant. Variance inflation factors (VIFs) ranged from 1.112 to 2.320, indicating no evidence of problematic multicollinearity among the predictors. For hierarchical regres-

TABLE 5. EFA using ULS extraction with oblimin rotation and two factors. Loadings below 0.30 are omitted for clarity.

Item	F1	F2	h^2
PEOU1		0.624	0.410
PEOU2		0.731	0.537
PEOU3		0.683	0.495
PU1	0.593		0.412
PU2	0.588		0.376
PU3	0.610		0.381
QOE1	0.620		0.469
QOE2	0.688		0.545
QOE3	0.630	0.302	0.488
TR1	0.742		0.560
TR2	0.758		0.574
TR3	0.770		0.594
BI1	0.819		0.690
BI2	0.730		0.549
BI3	0.807		0.672
Variance (%)	39.643	12.027	

TABLE 6. EFA using ULS extraction with oblimin rotation and three factors. Loadings below 0.30 are omitted for clarity.

Item	F1	F2	F3	h^2
PEOU1			0.613	0.394
PEOU2			0.727	0.532
PEOU3			0.681	0.481
PU1	0.724			0.588
PU2	0.730			0.562
PU3	0.778			0.618
QOE1	0.336	0.329	0.327	0.328
QOE2		0.481	0.321	0.400
QOE3		0.502	0.361	0.412
TR1		0.945		0.903
TR2		0.824		0.682
TR3		0.758		0.585
BI1	0.772			0.645
BI2	0.734			0.570
BI3	0.564	0.318		0.436
Variance (%)	22.497	19.303	12.444	

sion models, hypothesis outcomes are reported only for the incremental tests (H5b and H5c), whereas the remaining rows present regression coefficients for predictors included in the final model. Detailed hypothesis-testing results are summarized in Table 8.

V. QOE, TR, AND BI ESTIMATION USING RAGAS METRICS

The final stage of the analysis examined the relationship between automated RAGAS evaluation metrics and users' perceived QOE, TR, and BI. Since RAGAS metrics were calculated individually for each respondent following chatbot interaction, they could be directly compared with the corresponding QOE, TR, and BI evaluations collected through the questionnaire. The analysis was conducted in two stages. First, correlation analysis was used to identify associations between individual RAGAS metrics and QOE, TR, and BI dimensions. Second, regression models were developed to

TABLE 7. EFA using ULS extraction with oblimin rotation and five factors. Loadings below 0.30 are omitted for clarity.

Item	F1	F2	F3	F4	F5	h^2
PEOU1					0.657	0.479
PEOU2					0.846	0.744
PEOU3					0.575	0.394
PU1				0.774		0.651
PU2				0.861		0.777
PU3				0.532		0.412
QOE1		0.861				0.748
QOE2		0.811				0.682
QOE3		0.589				0.446
TR1	0.905					0.826
TR2	0.855					0.744
TR3	0.655					0.467
BI1			0.729			0.560
BI2			0.940			0.887
BI3			0.557			0.376
Variance (%)	14.416	13.041	12.192	11.575	10.069	

investigate the extent to which QOE, TR, and BI could be estimated from RAGAS metrics.

A. CORRELATION ANALYSIS BETWEEN RAGAS METRICS AND QOE, TR, AND BI

Pearson's linear correlation coefficients (PLCCs) and Spearman's rank correlation coefficients (SRCCs) were calculated to examine the relationships between automated RAGAS metrics and perceived QOE, TR, and BI. The results are presented in Tables 9 to 11.

Across both correlation coefficients, answer relevance exhibited the strongest and most consistent positive associations with QOE and TR. The highest correlations were observed between answer relevance and QOE3, corresponding to positive user experience (PLCC = 0.369, $p < 0.001$; SRCC = 0.400, $p < 0.001$). Similar results were found for the TR variables, with the strongest Pearson's correlation found between answer relevance and TR2 (PLCC = 0.372, $p < 0.001$), whereas the strongest Spearman's correlation was observed between answer relevance and TRmean (SRCC = 0.390, $p < 0.001$).

The overall score showed consistent positive associations with QOE and TR, although weaker than those for answer relevance. Context precision showed weaker and less consistent relationships with QOE and TR, while faithfulness showed only weak relationships with both variables. In contrast, RAGAS metrics showed only weak correlations with BI, with PLCC and SRCC remaining close to zero across all BI variables.

B. REGRESSION MODELS FOR QOE, TR, AND BI ESTIMATION

Models using all four RAGAS metrics as predictors, as well as models using answer relevance as an individual predictor, were evaluated for QOE and TR estimation. Based on the correlation analysis, answer relevance was selected as the

primary individual predictor. QOE_{mean} and TR_{mean} were computed as the arithmetic mean of the three QOE or TR questionnaire items. Several regression models were evaluated using nested cross-validation, including a baseline mean predictor, linear regression, quadratic polynomial regression, ridge regression, lasso regression, elastic net, decision trees, k-nearest neighbors, and support vector regression (SVR) [55], as well as XGBoost [56]. This broad set of models was included as an exploratory benchmark to compare linear, regularized, and nonlinear prediction approaches under the same nested cross-validation framework. The baseline predictor estimates all observations using the mean value of the target variable in the training set. The outer loop consisted of repeated five-fold cross-validation with 100 repetitions, resulting in 500 train-test splits. RMSE was calculated from the mean MSE across the outer test folds, whereas SRCC was averaged across the valid outer test folds. A five-fold cross-validation procedure was used in the inner loop for model selection and hyperparameter tuning.

Tables 12 and 13 report the best-performing nested cross-validation models using answer relevance as the individual predictor. The relative improvement in RMSE over the baseline predictor was calculated using (2).

$$\text{Improvement (\%)} = 100 \cdot \frac{\text{RMSE}_{\text{baseline}} - \text{RMSE}_{\text{model}}}{\text{RMSE}_{\text{baseline}}} \quad (2)$$

where $\text{RMSE}_{\text{baseline}}$ and $\text{RMSE}_{\text{model}}$ denote the RMSE of the baseline predictor and the evaluated regression model, respectively. The results indicate that linear and quadratic regression models generally outperformed the more complex alternatives. The strongest predictive performance among the QOE models was observed for QOE3 (SRCC = 0.392), which is consistent with the correlation analysis (Table 9), where QOE3 exhibited the strongest associations with the RAGAS metrics. RMSE improvements of 2.6%–6.3% were obtained for QOE variables, while QOE_{mean} achieved a 5.5% improvement over the baseline predictor. For the TR variables, predictive performance was generally similar to that observed for the QOE variables and yielded consistent improvements over the baseline predictor. The strongest regression result was obtained for TR2 (SRCC = 0.370), which is broadly consistent with the correlation analysis (Table 10), where TR2 (PLCC) and TR_{mean} (SRCC) exhibited the strongest associations with answer relevance. RMSE improvements of 2.7%–6.3% were obtained for TR variables, while TR_{mean} achieved a 5.1% improvement over the baseline predictor.

Models using all four RAGAS metrics as predictors generally did not improve predictive performance over the simpler models based solely on answer relevance. Among the evaluated QOE and TR targets, only QOE3 exhibited a marginally lower RMSE (0.898 vs. 0.899) using a tuned ridge regression model, whereas all remaining targets were predicted more accurately using only answer relevance. None of the evaluated regression models for BI outperformed the baseline mean predictor, indicating that the RAG evaluation metrics provided little predictive value for BI.

TABLE 8. Regression results for hypothesis testing. Simple models report bivariate associations, whereas multiple models report unique predictor contributions. Result refers to individual predictor significance, while Overall model reports model significance, F_{model} , and p_{model} . Bold indicates the evaluated predictor in multiple models and the newly added predictor in hierarchical models.

Hyp.	Relationship	Model	β	t	p	R^2	ΔR^2	Result	Overall model
H1a	SE $\rightarrow$ PEOU	Simple	0.372	3.758	< 0.001	0.138	–	Supported	–
H1b	SA $\rightarrow$ PEOU	Simple	0.535	5.938	< 0.001	0.286	–	Supported	–
H1c	EE $\rightarrow$ PEOU	Simple	0.296	2.911	0.005	0.088	–	Supported	–
H1a	SE + SA + EE $\rightarrow$ PEOU	Multiple	0.187	1.979	0.051	0.348	–	Not supported	Significant $F_{\text{model}}(3, 86) = 15.33$ $p_{\text{model}} < 0.001$
H1b	SE + SA + EE $\rightarrow$ PEOU	Multiple	0.442	4.744	< 0.001	0.348	–	Supported	
H1c	SE + SA + EE $\rightarrow$ PEOU	Multiple	0.144	1.567	0.121	0.348	–	Not supported	
H1d	PEOU $\rightarrow$ PU	Simple	0.378	3.825	< 0.001	0.143	–	Supported	–
H1e	SE $\rightarrow$ PU	Simple	0.186	1.774	0.079	0.035	–	Not supported	–
H1f	SA $\rightarrow$ PU	Simple	0.334	3.321	0.001	0.111	–	Supported	–
H1g	EE $\rightarrow$ PU	Simple	–0.008	–0.075	0.941	0.000	–	Not supported	–
H1d	PEOU + SE + SA + EE $\rightarrow$ PU	Multiple	0.302	2.493	0.015	0.189	–	Supported	Significant $F_{\text{model}}(4, 85) = 4.94$ $p_{\text{model}} = 0.001$
H1e	PEOU + SE + SA + EE $\rightarrow$ PU	Multiple	0.056	0.519	0.605	0.189	–	Not supported	
H1f	PEOU + SE + SA + EE $\rightarrow$ PU	Multiple	0.189	1.612	0.111	0.189	–	Not supported	
H1g	PEOU + SE + SA + EE $\rightarrow$ PU	Multiple	–0.156	–1.493	0.139	0.189	–	Not supported	
H2a	PEOU $\rightarrow$ QOE	Simple	0.480	5.130	< 0.001	0.230	–	Supported	–
H2b	PU $\rightarrow$ QOE	Simple	0.630	7.617	< 0.001	0.397	–	Supported	–
H2a	PEOU + PU $\rightarrow$ QOE	Multiple	0.282	3.331	0.001	0.466	–	Supported	Significant $F_{\text{model}}(2, 87) = 37.89$ $p_{\text{model}} < 0.001$
H2b	PEOU + PU $\rightarrow$ QOE	Multiple	0.524	6.189	< 0.001	0.466	–	Supported	
H3	QOE $\rightarrow$ TR	Simple	0.665	8.362	< 0.001	0.443	–	Supported	–
H4a	QOE $\rightarrow$ BI	Simple	0.572	6.542	< 0.001	0.327	–	Supported	–
H4b	TR $\rightarrow$ BI	Simple	0.583	6.738	< 0.001	0.340	–	Supported	–
H4a	QOE + TR $\rightarrow$ BI	Multiple	0.330	2.968	0.004	0.401	–	Supported	Significant $F_{\text{model}}(2, 87) = 29.12$ $p_{\text{model}} < 0.001$
H4b	QOE + TR $\rightarrow$ BI	Multiple	0.364	3.273	0.002	0.401	–	Supported	
H5a	PU $\rightarrow$ BI	Simple	0.603	7.088	< 0.001	0.363	–	Supported	–
H5b	PU + QOE $\rightarrow$ BI	Hierarchical	0.319	3.041	0.003	0.425	0.061	Supported	–
	PU + QOE $\rightarrow$ BI	Hierarchical	0.402	3.838	< 0.001	0.425	0.061	–	–
H5c	PU + QOE + TR $\rightarrow$ BI	Hierarchical	0.325	3.107	0.003	0.483	0.058	Supported	–
	PU + QOE + TR $\rightarrow$ BI	Hierarchical	0.370	3.686	< 0.001	0.483	0.058	–	–
	PU + QOE + TR $\rightarrow$ BI	Hierarchical	0.123	1.039	0.302	0.483	0.058	–	–

TABLE 9. PLCC and SRCC between RAGAS metrics and QOE variables.

Corr.	Question	RAGAS1	RAGAS2	RAGAS3	RAGAS4
PLCC	QOE1	0.197	0.293**	0.259*	0.291**
	QOE2	0.160	0.317**	0.178	0.249*
	QOE3	0.207	0.369***	0.312**	0.343***
	QOEmean	0.201	0.349***	0.266*	0.315**
SRCC	QOE1	0.222*	0.295**	0.248*	0.277**
	QOE2	0.234*	0.363***	0.218*	0.293**
	QOE3	0.246*	0.400***	0.340**	0.364***
	QOEmean	0.255*	0.377***	0.289**	0.335**

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

TABLE 10. PLCC and SRCC between RAGAS metrics and TR variables.

Corr.	Question	RAGAS1	RAGAS2	RAGAS3	RAGAS4
PLCC	TR1	0.073	0.268*	0.162	0.188
	TR2	0.094	0.372***	0.178	0.240*
	TR3	0.164	0.305**	0.204	0.256*
	TRmean	0.122	0.343***	0.199	0.249*
SRCC	TR1	0.118	0.317**	0.174	0.223*
	TR2	0.137	0.378***	0.178	0.244*
	TR3	0.218*	0.354***	0.238*	0.298**
	TRmean	0.181	0.390***	0.214*	0.284**

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

VI. DISCUSSION

A. ACCEPTANCE OF THE RAG-BASED UNIVERSITY CHATBOT

The results generally support the proposed extension of the Technology Acceptance Model for evaluating a RAG-based university chatbot. The findings indicate that students' acceptance of the system is associated not only with the traditional TAM constructs of PEOU and PU, but also with the experi-

ential and trust-related constructs of QOE and TR.

Both PEOU and PU were individually associated with QOE and remained significant unique predictors when considered jointly. These relationships suggest complementary mechanisms through which traditional TAM constructs may shape the experienced quality of interaction: PEOU reflects the effort required to use the system, whereas PU reflects the instrumental value users obtain from it. This finding is

TABLE 11. PLCC and SRCC between RAGAS metrics and BI variables.

Corr.	Question	RAGAS1	RAGAS2	RAGAS3	RAGAS4
PLCC	BI1	−0.049	0.058	0.093	0.035
	BI2	−0.085	−0.056	0.072	−0.027
	BI3	−0.109	0.073	0.024	−0.014
	BI _{mean}	−0.092	0.029	0.069	−0.003
SRCC	BI1	−0.034	0.075	0.100	0.041
	BI2	−0.107	−0.054	0.037	−0.052
	BI3	−0.073	0.090	0.019	−0.010
	BI _{mean}	−0.086	0.046	0.049	−0.014

TABLE 12. Best nested cross-validation regression models for QOE estimation using answer relevance. Best model (from all tested models) according to the RMSE is bolded.

Target	Model	RMSE _{baseline}	RMSE _{model}	Impr. (%)	SRCC
QOE1	Linear	1.047 ± 0.127	1.012 ± 0.120	3.333	0.288
	Quadratic		1.020 ± 0.121	2.576	0.285
QOE2	Linear	1.060 ± 0.134	1.013 ± 0.152	4.371	0.356
	Quadratic		1.014 ± 0.149	4.357	0.323
QOE3	Linear	0.959 ± 0.131	0.899 ± 0.155	6.254	0.392
	Quadratic		0.910 ± 0.158	5.170	0.390
QOE _{mean}	Linear	0.952 ± 0.128	0.899 ± 0.142	5.490	0.366
	Quadratic		0.905 ± 0.140	4.848	0.365

consistent with previous chatbot research showing that both perceived ease of use and perceived usefulness positively influence users' chatbot experience [41]. QOE therefore provides an experiential component linking traditional technology acceptance factors with users' evaluation of the actual interaction.

The relationship between QOE and TR further extends this mechanism by suggesting that trust in an AI chatbot may develop partly through the quality of the experienced interaction. A positive experience that meets users' expectations may provide cues about the reliability and competence of the system, thereby contributing to greater TR. This finding is consistent with previous research identifying user experience as an antecedent of trust in generative AI chatbots [42]. This relationship is particularly relevant for RAG-based conversational systems, where users must evaluate not only whether the technology is useful and easy to use, but also whether they can rely on the information it provides.

Both QOE and TR were individually associated with BI and remained significant unique predictors when considered jointly. This suggests that students' intention to use the application is associated with both the quality of their experience and their trust in the chatbot. In particular, the significant relationship between TR and BI is consistent with previous findings linking trust to continuance and reuse intentions in AI chatbot contexts [42], [43]. These findings support the inclusion of QOE and TR as additional constructs for explaining BI in the proposed RAG-based chatbot.

The hierarchical regression analysis further clarifies the complementary roles of PU, QOE, and TR in explaining BI. QOE provided explanatory value beyond PU, while the addition of TR further increased the explained variance in BI.

TABLE 13. Best nested cross-validation regression models for TR estimation using answer relevance. Best model (from all tested models) according to the RMSE is bolded.

Target	Model	RMSE _{baseline}	RMSE _{model}	Impr. (%)	SRCC
TR1	Linear	1.030 ± 0.141	1.002 ± 0.145	2.695	0.311
	Quadratic		0.995 ± 0.146	3.432	0.313
TR2	Linear	0.950 ± 0.109	0.891 ± 0.123	6.291	0.370
	Quadratic		0.899 ± 0.124	5.462	0.370
TR3	Linear	1.066 ± 0.131	1.024 ± 0.138	3.915	0.345
	Quadratic		1.021 ± 0.139	4.194	0.352
TR _{mean}	Linear	0.929 ± 0.119	0.881 ± 0.130	5.139	0.380
	Quadratic		0.881 ± 0.132	5.104	0.380

Notably, the unique contribution of QOE was no longer statistically significant after TR was included, whereas PU and TR remained significant, suggesting some overlap between the explanatory contributions of QOE and TR to BI. Thus, the proposed extension contributes to TAM by distinguishing between perceived usefulness, the quality of the experienced interaction, and trust in the AI-generated information as related but distinct factors associated with BI.

The external factors showed a differentiated pattern in relation to PEOU. While SE, SA, and EE were significant individual predictors of PEOU, SA remained the only significant unique predictor when the three external factors were considered jointly. This suggests that system accessibility and reliable operation may be more relevant to perceived ease of use than users' self-efficacy or prior chatbot experience in this setting.

For PU, among the external factors examined individually, only SA was a significant predictor, while neither SE nor EE showed a significant relationship with PU. When considered jointly with PEOU, none of the external factors provided a significant unique contribution to PU, whereas PEOU remained significant. The loss of significance of SA suggests that its relationship with PU may overlap with the relationship between PEOU and PU, although this does not establish mediation. Since participants interacted with the application within a single study session rather than through sustained use, they had limited opportunity to encounter usage problems or learn the system through extended use, which may explain the limited role of SE in their immediate perceptions of usefulness. Furthermore, PU referred to the UNINwalk application as a whole, whereas EE reflected prior chatbot use, which may explain the limited relationship between EE and PU.

B. ESTIMATING PERCEIVED QOE, TR, AND BI FROM AUTOMATED RAG METRICS

The second objective of this study was to investigate whether automated RAG evaluation metrics could serve as indicators of perceived QOE, TR, and BI. The results suggest that RAGAS-based evaluation and subjective evaluations are related for QOE and TR, although the observed relationships were generally modest rather than strong. In contrast, RAGAS metrics exhibited only weak relationships with BI.

Among the evaluated metrics, answer relevance consistently exhibited the strongest associations with both QOE and TR variables. This suggests that answer relevance is more closely related to users' perceived QOE and TR than other evaluated RAG dimensions, particularly faithfulness and context precision. These findings are broadly consistent with previous research. While the study by [33] focused on correlations between RAG metrics and expert human evaluations of answer quality, this study examined relationships between RAG metrics and end-user perceptions of QOE and TR. Although both studies report limited to moderate relationships between automated RAG metrics and human assessments, their findings are not directly comparable because the respective human evaluations represent different constructs. More broadly, human-centered AI evaluation research emphasizes that automated evaluation methods may not fully capture subjective user perceptions and should complement rather than replace human evaluation [57].

The regression results further suggest that RAGAS metrics provide some predictive information about perceived QOE and TR, but little predictive value for BI. While answer relevance exhibited a measurable relationship with users' perceived QOE and TR, the predictive performance remained modest. Perceived QOE is likely affected by additional factors that are not captured by RAGAS metrics, including interface design, user expectations, and contextual aspects of interaction [8], [58]. Similarly, TR may be associated with broader aspects of user experience that are not directly captured by RAGAS metrics [42].

C. STUDY LIMITATIONS

RAGAS-based evaluation relies on LLM-based judgments, and the resulting scores may therefore depend on the performance and characteristics of the evaluator model [2]. In this study, the same Llama model was used for response generation and RAGAS-based evaluation, which may introduce evaluator dependence and model-specific bias. Future work should compare RAGAS scores obtained using different evaluator models to assess the robustness of the observed relationships.

The user-facing presentation of RAGAS scores represents another limitation of the study. Although the evaluation panel was collapsed by default, participants who expanded it could view the scores before providing their subjective evaluations and may therefore have incorporated this information into their judgments of QOE, TR, and BI. Panel expansion was not recorded, so the number of participants who viewed the scores cannot be determined. The presentation of these scores was an intentional design choice to support system transparency; nevertheless, future studies should separate subjective assessment from the presentation of automated evaluation scores or explicitly record participants' exposure to the evaluation panel.

Furthermore, although the questionnaire items were based on established constructs and prior literature, a formal content

validity assessment was not conducted prior to data collection.

VII. CONCLUSION AND FUTURE WORK

This study proposed an extended TAM that integrates QOE and TR constructs to explain students' BI. The model was evaluated using the UNINwalk university application, and the empirical results largely supported the proposed hypotheses. PU, QOE, and TR emerged as significant predictors of BI, while TR contributed additional explanatory power beyond PU and QOE, highlighting the importance of TR in AI-assisted educational environments.

The study also investigated relationships between automated RAG evaluation metrics and subjective user evaluations. Among the evaluated metrics, answer relevance exhibited the strongest and most consistent associations with QOE and TR. Regression analysis demonstrated that answer relevance provided modest but consistent predictive information about perceived QOE and TR, yielding RMSE improvements of 2.6%–6.3% and 2.7%–6.3%, respectively, over baseline prediction. In contrast, RAGAS metrics provided little predictive value for BI. These findings suggest that automated RAG evaluation metrics are more closely related to users' perceived QOE and TR than to their BI. However, RAGAS metrics alone should not be considered substitutes for user-centered evaluation, as a substantial proportion of the variance in perceived QOE and TR remained unexplained.

The findings have implications for both chatbot evaluation and educational AI deployment. From a practical perspective, answer relevance may provide a useful automated indicator of certain aspects of perceived QOE and TR during iterative system development. At the same time, the results indicate that higher-level acceptance outcomes such as BI cannot be reliably inferred from RAGAS metrics alone.

Future work should examine larger and more diverse student populations, additional RAG evaluation metrics, and multi-feature prediction models. The robustness of RAGAS-based evaluation should also be assessed by comparing scores obtained using different evaluator models. The retained five-factor measurement structure should also be confirmed using confirmatory factor analysis on a larger independent sample, ideally following more sustained use of the chatbot. In addition to RAGAS scores, future approaches could incorporate response length, token usage, lexical characteristics, semantic similarity measures, and other interaction-level features that may better capture the factors influencing perceived QOE, TR, and BI.

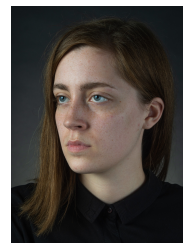
VIII. DATA AVAILABILITY STATEMENT

The anonymized survey responses and RAGAS evaluation data used in this study are publicly available at <https://msl.unin.hr/uninwalk-1/results> (accessed on 10 September 2026). The UNINwalk application and its integrated chatbot are publicly available at <https://msl.unin.hr/uninwalk-1> (accessed on 10 September 2026).

REFERENCES

- [1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 9459 – 9474. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf
- [2] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAs: Automated evaluation of retrieval augmented generation," in *EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of System Demonstrations*, 2024, pp. 150 – 158. [Online]. Available: <https://doi.org/10.18653/v1/2024.eacl-demo.16>
- [3] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly: Management Information Systems*, vol. 13, no. 3, pp. 319 – 339, 1989. [Online]. Available: <https://doi.org/10.2307/249008>
- [4] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly: Management Information Systems*, vol. 27, no. 3, pp. 425 – 478, 2003. [Online]. Available: <https://doi.org/10.2307/30036540>
- [5] H. Liu, M. A. H. Jafri, S. Xu, and M. F. Shahzad, "The impact of artificial intelligence on consumers' willingness to use CBDCs: evidence from the chinese banking sector," *Humanities and Social Sciences Communications*, vol. 12, no. 1, 2025. [Online]. Available: <https://doi.org/10.1057/s41599-025-05067-5>
- [6] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50 – 80, 2004. [Online]. Available: <https://doi.org/10.1518/hfes.46.1.50.30392>
- [7] I. Celik, H. Muukkonen, and S. Siklander, "Teacher – artificial intelligence (AI) interaction: The role of trust, subjective norm and innovativeness in teachers' acceptance of educational chatbots," *Policy Futures in Education*, vol. 24, no. 1, pp. 164 – 188, 2026. [Online]. Available: <https://doi.org/10.1177/14782103251348551>
- [8] ITU-T P.852, "Subjective quality evaluation of text-based chatbots," International Telecommunications Union, Tech. Rep., July 2022. [Online]. Available: <https://www.itu.int/rec/T-REC-P.852/en>
- [9] V. Venkatesh and F. D. Davis, "Theoretical extension of the technology acceptance model: Four longitudinal field studies," *Management Science*, vol. 46, no. 2, pp. 186 – 204, 2000. [Online]. Available: <https://doi.org/10.1287/mnsc.46.2.186.11926>
- [10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600 – 612, 2004. [Online]. Available: <https://doi.org/10.1109/TIP.2003.819861>
- [11] E. Dumić, K. Sakic, and L. A. da Silva Cruz, "Crowdsourced subjective 3D video quality assessment," *Multimedia Systems*, vol. 25, no. 6, pp. 673 – 694, 2019. [Online]. Available: <https://doi.org/10.1007/s00530-019-00619-7>
- [12] P. Rabuzin and E. Dumić, "AI-generated video quality assessment: Dataset and analysis," in *International Symposium on Image and Signal Processing and Analysis, ISPA*, 2025, pp. 193 – 198. [Online]. Available: <https://doi.org/10.1109/ISPA66905.2025.11259101>
- [13] S. Palazari and E. Dumić, "Towards reliable evaluation of underwater image enhancement using subjective and objective analysis," *Electronics (Switzerland)*, vol. 15, no. 11, 2026. [Online]. Available: <https://doi.org/10.3390/electronics15112412>
- [14] D. Ramandanis and S. Xinogalos, "Designing a chatbot for contemporary education: A systematic literature review," *Information (Switzerland)*, vol. 14, no. 9, 2023. [Online]. Available: <https://doi.org/10.3390/info14090503>
- [15] L. Labadze, M. Grigolia, and L. Machaidze, "Role of AI chatbots in education: systematic literature review," *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, 2023. [Online]. Available: <https://doi.org/10.1186/s41239-023-00426-1>
- [16] T. Debets, S. K. Banihashem, D. Joosten-Ten Brinke, T. E. Vos, G. Maillette de Buy Wenniger, and G. Camp, "Chatbots in education: A systematic review of objectives, underlying technology and theory, evaluation criteria, and impacts," *Computers and Education*, vol. 234, 2025. [Online]. Available: <https://doi.org/10.1016/j.compedu.2025.105323>
- [17] C. McGrath, A. Farazouli, and T. Cerratto-Pargman, "Generative AI chatbots in higher education: a review of an emerging research area," *Higher Education*, vol. 89, no. 6, pp. 1533 – 1549, 2025. [Online]. Available: <https://doi.org/10.1007/s10734-024-01288-w>
- [18] W. Huang, J. Jiang, R. B. King, and L. K. Fryer, "Chatbots and student motivation: a scoping review," *International Journal of Educational Technology in Higher Education*, vol. 22, no. 1, 2025. [Online]. Available: <https://doi.org/10.1186/s41239-025-00524-2>
- [19] K. Peyton, S. Unnikrishnan, and B. Mulligan, "A review of university chatbots for student support: FAQs and beyond," *Discover Education*, vol. 4, no. 1, 2025. [Online]. Available: <https://doi.org/10.1007/s44217-025-00397-7>
- [20] H. Dinh and T. K. Tran, "Educhat: An AI-based chatbot for university-related information using a hybrid approach," *Applied Sciences (Switzerland)*, vol. 13, no. 22, 2023. [Online]. Available: <https://doi.org/10.3390/app132212446>
- [21] S. Neupane, E. Hossain, J. Keith, H. Tripathi, F. Ghiasi, N. A. Golilarz, A. Amiratifi, S. Mittal, and S. Rahimi, "From questions to insightful answers: Building an informed chatbot for university resources," in *Proceedings - Frontiers in Education Conference, FIE*, 2024. [Online]. Available: <https://doi.org/10.1109/FIE61694.2024.10892994>
- [22] A. T. Neumann, Y. Yin, S. Sowe, S. Decker, and M. Jarke, "An LLM-driven chatbot in higher education for databases and information systems," *IEEE Transactions on Education*, vol. 68, no. 1, pp. 103 – 116, 2025. [Online]. Available: <https://doi.org/10.1109/TE.2024.3467912>
- [23] K. F. Hew, W. Huang, J. Du, and C. Jia, "Using chatbots to support student goal setting and social presence in fully online activities: learner engagement and perceptions," *Journal of Computing in Higher Education*, vol. 35, no. 1, pp. 40 – 68, 2023. [Online]. Available: <https://doi.org/10.1007/s12528-022-09338-x>
- [24] W. Huang and K. F. Hew, "Facilitating online self-regulated learning and social presence using chatbots: Evidence-based design principles," *IEEE Transactions on Learning Technologies*, vol. 18, pp. 56 – 71, 2025. [Online]. Available: <https://doi.org/10.1109/TLT.2024.3523199>
- [25] M. D. L. Roca, M. M. Chan, A. Garcia-Cabot, E. Garcia-Lopez, and H. Amado-Salvatierra, "The impact of a chatbot working as an assistant in a course for supporting student learning and engagement," *Computer Applications in Engineering Education*, vol. 32, no. 5, 2024. [Online]. Available: <https://doi.org/10.1002/cae.22750>
- [26] M. Farrukh Shahzad, S. Xu, A. R. Bilal, Quratulain, and S. Akram, "Exploring role of digital literacy and theory of planned behavior in shaping generative AI behavior intentions for educational sustainability," *International Journal of Sustainability in Higher Education*, 2026. [Online]. Available: <https://doi.org/10.1108/IJSHE-10-2025-1237>
- [27] F. Subhani, S. A. Khan, M. A. Sandhu, and M. F. Shahzad, "What factors affect the adoption intention and actual use of ChatGPT in higher education? the moderating role of academic integrity," *TechTrends*, vol. 69, no. 5, pp. 1056 – 1071, 2025. [Online]. Available: <https://doi.org/10.1007/s11528-025-01096-8>
- [28] J. H. Gruenhagen, P. M. Sinclair, J.-A. Carroll, P. R. Baker, A. Wilson, and D. Demant, "The rapid rise of generative AI and its implications for academic integrity: Students' perceptions and use of chatbots for assistance with assessments," *Computers and Education: Artificial Intelligence*, vol. 7, 2024. [Online]. Available: <https://doi.org/10.1016/j.caeai.2024.100273>
- [29] A. M. Vieriu and G. Petrea, "The impact of artificial intelligence (AI) on students' academic development," *Education Sciences*, vol. 15, no. 3, 2025. [Online]. Available: <https://doi.org/10.3390/educsci15030343>
- [30] W. Villegas-Ch, A. Arias-Navarrete, and X. Palacios-Pacheco, "Proposal of an architecture for the integration of a chatbot with artificial intelligence in a smart campus for the improvement of learning," *Sustainability (Switzerland)*, vol. 12, no. 4, 2020. [Online]. Available: <https://doi.org/10.3390/su12041500>
- [31] M. Rukhiran, N. Phaokla, and P. Netinant, "Adoption of environmental information chatbot services based on the internet of educational things in smart schools: Structural equation modeling approach," *Sustainability (Switzerland)*, vol. 14, no. 23, 2022. [Online]. Available: <https://doi.org/10.3390/su142315621>
- [32] J. Swacha and M. Gracel, "Retrieval-augmented generation (RAG) chatbots for education: A survey of applications," *Applied Sciences (Switzerland)*, vol. 15, no. 8, 2025. [Online]. Available: <https://doi.org/10.3390/app15084234>
- [33] E. Oro, F. M. Granata, A. Lanza, A. Bachir, L. De Grandis, and M. Ruffolo, "Evaluating retrieval-augmented generation for question answering with large language models," in *CEUR Workshop Proceedings*, vol. 3762, 2024, pp. 129 – 134. [Online]. Available: <https://ceur-ws.org/Vol-3762/495.pdf>

- [34] S. Roychowdhury, S. Soman, H. G. Ranjani, N. Gunda, V. Chhabra, and S. K. Bala, "Evaluation of RAG metrics for question answering in the telecom domain," in *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. [Online]. Available: <https://icml.cc/virtual/2024/39469>
- [35] M. Khasanova Zafar kizi and Y. Suh, "Design and performance evaluation of LLM-based RAG pipelines for chatbot services in international student admissions," *Electronics (Switzerland)*, vol. 14, no. 15, 2025. [Online]. Available: <https://doi.org/10.3390/electronics14153095>
- [36] M. Hassenzahl, "User experience (UX): Towards an experiential perspective on product quality," in *ACM International Conference Proceeding Series*, 2008, pp. 11 – 15. [Online]. Available: <https://doi.org/10.1145/1512714.1512717>
- [37] D. Gefen, E. Karahanna, and D. W. Straub, "Trust and TAM in online shopping: An integrated model," *MIS Quarterly: Management Information Systems*, vol. 27, no. 1, pp. 51 – 90, 2003. [Online]. Available: <https://doi.org/10.2307/30036519>
- [38] A. Polypotis and N. Pahos, "Understanding students' adoption of the ChatGPT chatbot in higher education: the role of anthropomorphism, trust, design novelty and institutional policy," *Behaviour and Information Technology*, vol. 44, no. 2, pp. 315 – 336, 2025. [Online]. Available: <https://doi.org/10.1080/0144929X.2024.2317364>
- [39] R. H. Mustofa, T. G. Kuncoro, D. Atmono, H. D. Hermawan, and Sukirman, "Extending the technology acceptance model: The role of subjective norms, ethics, and trust in AI tool adoption among students," *Computers and Education: Artificial Intelligence*, vol. 8, 2025. [Online]. Available: <https://doi.org/10.1016/j.caeai.2025.100379>
- [40] T. Nazaretsky, P. Mejia-Domenzain, V. Swamy, J. Frej, and T. Käser, "The critical role of trust in adopting AI-powered educational technology for learning: An instrument for measuring student perceptions," *Computers and Education: Artificial Intelligence*, vol. 8, 2025. [Online]. Available: <https://doi.org/10.1016/j.caeai.2025.100368>
- [41] I. Lubbe, M. Roberts-Lombard, and J. Langerman, "Millennials' experiences and satisfaction with chatbots: a study of self-service technology in emerging markets," *European Business Review*, vol. 37, no. 4, pp. 741 – 769, 2025. [Online]. Available: <https://doi.org/10.1108/EBR-06-2024-0190>
- [42] H. Yu, Y. Sang, and E. Han, "Digital competence of AI chatbot users: scale development and its relationships with smart customer experience, trust, and reuse intention," *Telematics and Informatics*, vol. 106, 2026. [Online]. Available: <https://doi.org/10.1016/j.tele.2026.102392>
- [43] S. Alzahrani and A. K. Bhunia, "How ai chatbot service quality drives continuance intention: an s-o-r perspective in saudi higher education," *Frontiers in Education*, 2026. [Online]. Available: <https://doi.org/10.3389/feduc.2026.1803390>
- [44] P. L. Callet, S. Möller, and A. Perkins, "Qualinet white paper on definitions of quality of experience (QoE)," European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC1003), White Paper Version 1.2, 2013. [Online]. Available: https://www.qualinet.eu/wp-content/uploads/2021/04/QoE_whitepaper_v1.2.pdf
- [45] International Telecommunication Union, "ITU-T recommendation p.10/g.100 amendment 2: Vocabulary for performance, quality of service and quality of experience," International Telecommunication Union, Tech. Rep., 2024. [Online]. Available: <https://www.itu.int/rec/T-REC-P.10-202405-1!Amd2/en>
- [46] P. A. Pavlou, "Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model," *International Journal of Electronic Commerce*, vol. 7, no. 3, pp. 101 – 134, 2003. [Online]. Available: <https://doi.org/10.1080/10864415.2003.11044275>
- [47] A. Choudhury and H. Shamszare, "Investigating the impact of user trust on the adoption and use of chatgpt: Survey analysis," *Journal of Medical Internet Research*, vol. 25, 2023. [Online]. Available: <https://doi.org/10.2196/47184>
- [48] S. Taylor and P. A. Todd, "Understanding information technology usage: A test of competing models," *Information Systems Research*, vol. 6, no. 2, pp. 144 – 176, 1995. [Online]. Available: <https://doi.org/10.1287/isre.6.2.144>
- [49] D. R. Compeau and C. A. Higgins, "Computer self-efficacy: Development of a measure and initial test," *MIS Quarterly: Management Information Systems*, vol. 19, no. 2, pp. 189 – 210, 1995. [Online]. Available: <https://doi.org/10.2307/249688>
- [50] W. H. DeLone and E. R. McLean, "The DeLone and McLean model of information systems success: A ten-year update," *Journal of Management Information Systems*, vol. 19, no. 4, pp. 9 – 30, 2003. [Online]. Available: <https://doi.org/10.1080/07421222.2003.11045748>
- [51] V. Venkatesh, "Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model," *Information Systems Research*, vol. 11, no. 4, pp. 342 – 365, 2000. [Online]. Available: <https://doi.org/10.1287/isre.11.4.342.11872>
- [52] H. F. Kaiser and J. Rice, "Little jiffy, Mark IV," *Educational and Psychological Measurement*, vol. 34, no. 1, pp. 111–117, 1974. [Online]. Available: <https://doi.org/10.1177/001316447403400115>
- [53] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang, "Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses," *Behavior Research Methods*, vol. 41, no. 4, pp. 1149 – 1160, 2009. [Online]. Available: <https://doi.org/10.3758/BRM.41.4.1149>
- [54] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2013. [Online]. Available: <https://doi.org/10.4324/9780203771587>
- [55] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction to Statistical Learning: with Applications in Python*. Springer, 2023. [Online]. Available: <https://doi.org/10.1007/978-3-031-38747-0>
- [56] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, 2016, pp. 785 – 794. [Online]. Available: <https://doi.org/10.1145/2939672.2939785>
- [57] P. A. Aoyagui, S. Ferguson, and A. Kuzminykh, "Exploring subjectivity for more human-centric assessment of social biases in large language models," in *Proceedings of the 1st HEAL Workshop (Human-centered Evaluation and Auditing of Language Models) at CHI 2024*, 2024, pp. 1 – 4. [Online]. Available: <https://arxiv.org/abs/2405.11048>
- [58] S. Amershi, D. Weld, M. Vorvoreanu, A. Fournier, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz, "Guidelines for human-AI interaction," in *Conference on Human Factors in Computing Systems - Proceedings*, 2019. [Online]. Available: <https://doi.org/10.1145/3290605.3300233>



NELA DURANEC received the title of baccalaurea of engineering in multimedia and graphic technology (bacc. ing. techn. graph.) in 2023, and the title of university master of engineering in multimedia and graphic technology (univ. mag. ing. techn. graph.), in 2025, both from University North, Varaždin, Croatia.

Since 2023, she has been working as a Frontend Developer, focusing on the design and implementation of web applications. Her research interests

include human–computer interaction, user interface, user experience and the evaluation of AI-based conversational systems in educational applications.



EMIL DUMIĆ received M.Sc. and Ph.D. degrees in electrical engineering from the University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia, in 2007 and 2011, respectively.

He is currently a Full Professor with the University North, Varaždin, Croatia. He has authored or coauthored more than 70 peer-reviewed publications, including journal articles, conference papers, and book chapters. His current research interests include multimedia, image processing, image quality assessment, point clouds, and human–computer interaction.

Prof. Dumić is a Member of the IEEE and the ACM.

...