applied sciences

MDPI

Article

A Multi-Agent Chatbot Architecture for AI-Driven Language Learning

Moneerh Aleedy [1,*], Eric Atwell [2] and Souham Meshoul [1,*]

[1] Department of Information Technology, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[2] School of Computer Science, University of Leeds, Leeds LS2 9JT, UK; e.s.atwell@leeds.ac.uk
* Correspondence: mmaleidi@pnu.edu.sa (M.A.); sbmeshoul@pnu.edu.sa (S.M.)

Abstract

Language learners increasingly rely on intelligent digital tools to supplement their learning experiences, yet existing chatbots often provide limited support, lacking adaptability, personalization, or domain-specific intelligence. This study introduces a novel AI-powered multi-agent chatbot architecture designed to support English–Arabic translation and language learning. Developed through a three-phase methodology, offline preparation, real-time deployment, and evaluation, the system employs both retrieval-based and generative AI models, with specialized agents managing tasks such as translation, example retrieval, user translation review, and learning feedback. The chatbot was developed using a hybrid architecture incorporating fine-tuned Generative Pre-trained Transformer (GPT) model, sentence embedding techniques, and similarity evaluation metrics. A user study involving 40 undergraduate students and 4 faculty members evaluated the system across usability, effectiveness, and pedagogical value. Results show that the multi-agent chatbot significantly enhanced learner engagement, provided accurate and contextually appropriate language support, and was positively received by both students and instructors. These findings demonstrate the value of multi-agent design in language learning applications and highlight the potential of AI-driven chatbots as intelligent educational assistants.

Keywords: multi-agent; chatbot; artificial intelligent; educational assistants; generative AI; retrieval-based AI; translation learning; language learning

Academic Editor: Martin Ebner

Received: 30 August 2025
Revised: 27 September 2025
Accepted: 29 September 2025
Published: 1 October 2025

Citation: Aleedy, M.; Atwell, E.; Meshoul, S. A Multi-Agent Chatbot Architecture for AI-Driven Language Learning. *Appl. Sci.* **2025**, *15*, 10634. https://doi.org/10.3390/app151910634

1. Introduction

The process of learning translation between English and Arabic involves a range of linguistic, cultural, and contextual challenges. These include recognizing accurate word choices, understanding idiomatic expressions, maintaining grammatical structure, and conveying meaning appropriately across two structurally different languages. For students in translation programs, especially at the undergraduate level, acquiring these skills requires continuous practice, guided feedback, and access to authentic examples. However, translation exercises in traditional classroom settings are often constrained by time and instructor availability, limiting opportunities for immediate correction and personalized learning support.

As part of a broader effort to enhance translation education through digital technologies, this study introduces a Translation Learning Chatbot designed to support English–Arabic translation practice and feedback. The system is tailored to address specific pedagogical needs identified in collaboration with faculty members at Princess Nourah bint

Abdulrahman University (PNU) [1], including improving translation accuracy, promoting contextual understanding, raising awareness of common mistakes, and enabling competence self-evaluation.

To meet these objectives, the chatbot was developed using a modular Artificial Intelligence (AI) driven architecture based on a multi-agent design. Each agent is responsible for a specific instructional function, such as generating translations, retrieving examples, reviewing user input, or testing translation competence, enabling the system to provide focused, context-aware support across different learning tasks. The chatbot combines retrieval-based and generative AI approaches, offering both example-based reinforcement and dynamic feedback.

This paper presents the system's full development lifecycle, following a three-phase methodology: an offline preparation phase that includes domain analysis, data preparation, and model development; a deployment phase in which the chatbot is integrated into a user-accessible platform; and an evaluation phase based on structured user studies with students and faculty. The results of this evaluation provide insights into the chatbot's usability, educational value, and potential to support self-directed translation learning.

2. Background

The development of the Translation Learning Chatbot was informed by several key advances in natural language processing, machine translation, and conversational AI. This section provides a concise overview of these foundational technologies, highlighting their relevance to translation learning. It outlines the historical evolution of chatbot systems, the role of machine translation in language education, and modern techniques for evaluating chatbot performance. Together, these areas form the technical and pedagogical foundation for the design choices and methodology presented in this study.

2.1. Evolution of Chatbots and Conversational Agents

The development of chatbots has progressed significantly, from early rule-based systems to modern AI-driven conversational agents. ELIZA, developed in 1966 by Joseph Weizenbaum, simulated a therapist using simple pattern-matching techniques [2]. PARRY, introduced in 1972, extended these ideas with a more sophisticated control structure and simulated emotional responses [3]. Later, ALICE, introduced in 1995 [4], built using AIML, and Mitsuku [5] enhanced conversational capability and personalization through larger response libraries.

Recent systems have adopted hybrid strategies that combine retrieval-based and generative methods. For example, AliMe Chat integrates both approaches using a sequence-to-sequence reranking model to refine outputs [6]. MILABOT, developed by the Montreal Institute for Learning Algorithms, applies deep reinforcement learning to select appropriate responses based on user interactions [7]. CAiRE, introduced in 2019, incorporated empathetic response generation using transformer-based models [8].

In educational contexts, chatbots have been adopted to support teaching and learning through interactive dialogue, immediate feedback, and accessible assistance. In higher education, they have been applied to tasks such as course navigation, student advising, and self-paced practice across various disciplines [9]. For language learning in particular, chatbots can provide low-pressure environments for practice, helping students engage with language tasks in a more flexible and autonomous way.

Early neural models like Bidirectional Recurrent Neural Networks (BRNNs), when combined with attention mechanisms, improved coherence by better capturing context in longer sequences [10]. These developments laid the foundation for transformer-based models used in modern generative AI systems.

The evolution of GPT demonstrates rapid progress in generative AI, from GPT-1 (2018) to GPT-3 (2020), with each version improving in fluency, contextual reasoning, and generation. ChatGPT, introduced in 2022, advanced these capabilities for more interactive and context-aware dialogue [11].

2.2. Machine Translation

Machine Translation (MT) refers to computer-based systems that translate text or speech from one natural language to another. The process involves generating a target sentence that accurately conveys the meaning of the source sentence [12–14]. The MT process is inherently complex, as it must account for the semantic, syntactic, morphological, and grammatical features of both the source and target languages. This complexity is further heightened when the languages involved have significant structural and linguistic differences such as English and Arabic [12].

2.2.1. Categories of Machine Translation

MT has three main categories:

- Rule-based MT: The traditional approach, which uses a set of linguistic rules and bilingual dictionaries to translate text between languages.
- Statistical MT: A data-driven approach that relies on probabilistic models and large-scale parallel corpora rather than grammatical rules.
- Neural MT (NMT): applies artificial neural networks to model translation as a sequence prediction task [13,15].

Comparing with other models, the NMT model requires less linguistic knowledge yet delivers competitive result. Numerous studies have shown that NMT can perform much better than the traditional Statistical MT model [10].

2.2.2. Modern Translation Tools

The most popular online tool for translation today is Google Translate [14]. It has evolved from a rule-based system to a statistical-based, and in 2016, Google introduced a new Neural Network system, enabling a higher quality of translation. The neural translation is widely seen to be an impressive improvement in the quality of MT [14,16]. Bing Translator is also MT owned by Microsoft [17]. It initially used a statistical approach but switched to a neural system more recently.

Furthermore, many other companies have tried to create translators, and one of the companies that succeeded in this is the German company DeepL GmbH [17]. However, DeepL uses convolutional neural networks based on the Linguee database, and it only supports nine languages (all Indo-European) [17].

These tools exemplify the shift toward neural architectures in commercial MT applications. Today, most leading translation systems are powered by deep learning techniques, particularly neural networks, regardless of the specific languages they support.

2.2.3. Machine Translation in Language Education

The use of machine translation in education has a long history, but recent advances in technology, device availability, and access to large language databases have made it significantly more accessible and effective. One of the important roles of MT in education is to use it as an effective supplementary learning tool while writing in a second language. MT helped students write faster and produce more fluent and natural writing with fewer errors [18]. Moreover, the new language students benefited the most from MT; it enabled them to express themselves better and communicate more in writing in a second language.

Several studies have also found that MT can help students get individual feedback about their writing [16,18].

2.3. Chatbot Evaluation Approaches

Chatbot evaluation approaches can be broadly classified into human-based, automatic, and semi-automatic (hybrid) categories, each offering distinct advantages depending on the evaluation goal, scale, and application context [19].

Human-based evaluation remains the gold standard for assessing subjective and nuanced aspects of chatbot interactions, such as empathy, coherence, and user satisfaction. These evaluations are typically performed through surveys or direct annotation by users or experts. A notable example is the Chatbot-Human Interaction Satisfaction Model (CHISM), which was developed to analyze chatbot use in higher education. CHISM incorporates three dimensions: Language Experience (LEX), Design Experience (DEX) and User Experience (UEX) to assess learner perceptions and satisfaction with chatbots [20].

Similarly, the Chatbot Usability Scale (BUS-11) was developed to assess chatbot user experience and is designed to measure users' perceptions of efficiency, accessibility, and engagement during interactions [21]. Modern evaluation approaches also include the measurement of affective components such as emotional appropriateness, social cues, and user trust, which are increasingly recognized as central to user satisfaction [22].

Automatic evaluation methods are broadly categorized into reference-based and reference-free approaches, depending on whether they rely on human-written reference texts [23]. outputs by comparing them to one or more gold-standard references. Surface-level metrics, such as BLEU [24] and ROUGE [25], evaluate similarity based on overlapping n-grams (i.e., sequences of words) between the generated and reference texts. These metrics are valued for their simplicity and speed but often fail to capture semantic meaning or paraphrased content [23]. Embedding-based metrics, such as BERTScore, use contextual token embeddings like the one from Bidirectional Encoder Representations from Transformers (BERT) to evaluate semantic similarity, enabling the detection of paraphrased meanings [23,26]. More advanced trained or model-based metrics utilize fine-tuned pre-trained models calibrated on human quality judgments; a notable example is BLEURT, which builds on BERT to predict human-like evaluation scores [27].

In contrast, reference-free metrics evaluate generated text without the need for reference outputs, relying solely on the generated content itself or its relationship to the input. One key category is language model-based metrics, such as perplexity and cross-entropy, which assess fluency by measuring how likely the text is under a language model [23]. Finally, dialogue-specific metrics are designed to evaluate the quality of chatbot responses. For example, the Unsupervised and Reference-free Score (USR) metric combines several automated scores, such as fluency, coherence, and semantic similarity, without relying on reference answers [28]. Another example is Model-based Automatic Unsupervised Validation Evaluation (MAUVE), which measures how different the distribution of generated chatbot responses is from that of real human responses, helping to detect unnatural or inconsistent outputs [29].

Bridging these two extremes, semi-automatic evaluation combines automated dialogue generation with human interpretation. A notable example is Human Unified with Statistical Evaluation (HUSE), which integrates human ratings and model-based statistics to train a classifier capable of distinguishing human from machine generated responses [30].

These foundational technologies directly inform the multi-agent architecture, hybrid AI components, and task-aligned design of the Translation Learning Chatbot described in the following section.

3. Methodology

This study presents the development and evaluation of the Translation Learning Chatbot, a pedagogically guided tool for English-to-Arabic translation learning. The system is built using vertical multi-agent architecture and a hybrid AI modeling approach, combining generative and retrieval-based techniques to enable modular, context-aware interactions tailored to learners' needs. Within this architecture, a central lead agent orchestrates several specialized agents—each dedicated to a specific pedagogical function such as translation generation, example retrieval, feedback provision, or competence testing—allowing the system to deliver both curated and dynamically generated responses.

To meet the diverse needs of translation learners, the chatbot adopts a hybrid model strategy that combines both retrieval-based and generative approaches. The retrieval-based component is built on a large collection of English–Arabic sentence pairs drawn from the Saudi Learner Translation Corpus (SauLTC) [31]. It is designed to fetch and return semantically relevant sentence pairs based on the specific task, whether presenting a translation example to the user or selecting a sentence for the user to translate. The generative component, powered by a large language model (LLM), operates using carefully designed prompts. It not only generates accurate Arabic translations from English input but also provides structured feedback on user-submitted translations.

To manage this workflow efficiently, the methodology is organized into three main phases, as illustrated in Figure 1:

- Offline Phase: This phase involves domain understanding, data collection and preparation, the development of models, and the design of agents. It establishes the foundational components required for real-time interaction.
- Deployment Phase: This phase covers deploying the chatbot and making it accessible through a user interface, where the lead agent coordinates user inputs and the responses of multiple specialized agents in real time.
- Evaluation Phase: This phase focuses on assessing the chatbot's effectiveness through user interaction and feedback. It involves structured tasks and surveys to evaluate translation accuracy, feedback quality, task usability, and overall user satisfaction using both quantitative and qualitative measures.

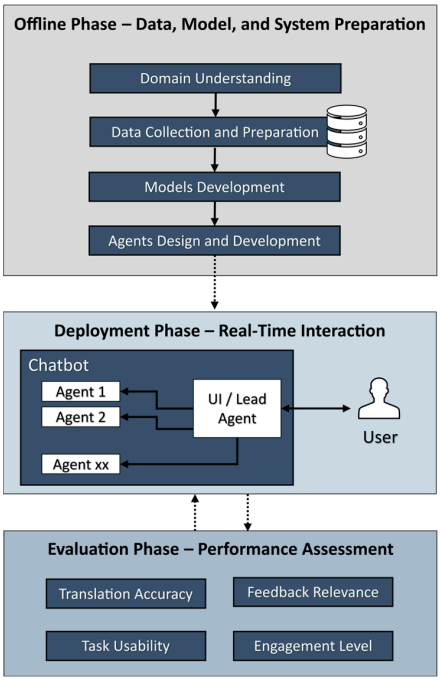

Figure 1. The Three-Phase Methodology of the Translation Learning Chatbot.

3.1. Offline Phase: Data, Model, and System Preparation

This section describes the offline phase of the proposed system, focusing on the preparation of essential components required before deployment and real-time interaction. It covers four key areas: domain understanding, data collection and preparation, models development, and agents design and development.

3.1.1. Domain Understanding

To inform the chatbot's design, we relied on findings from earlier study [32], which included a series of interviews with faculty members from the Translation Department at Princess Nourah University. These interviews identified four key pedagogical challenges in English–Arabic translation learning: ensuring translation accuracy, promoting contextual understanding, increasing awareness of common translation issues, and enabling competence evaluation. These insights directly informed the functional design of the chatbot.

3.1.2. Data Collection and Preparation

The dataset used for training and evaluation in this study was derived from the SauLTC and prepared within the framework of the study published in [32]. The corpus contains translations produced by university students and revised by faculty members. For this research, a subset of health-related English–Arabic texts were selected. The data preparation process included an initial assessment to correct mismatches and mislabelled entries, removal of duplicates and incomplete translations, filtering to retain only the final revised versions, and text preprocessing to normalize formatting inconsistencies. Non-essential metadata was discarded, and the cleaned texts were paired at the document level to link each English source with its corresponding Arabic translation. These paired documents provided the foundation for generating parallel sentence pairs, which were subsequently used for model development and evaluation in this study.

3.1.3. Models Development

The Translation Learning Chatbot employs a hybrid AI architecture that integrates both generative and retrieval-based models to support English–Arabic translation tasks. This dual approach allows the system to generate context-aware translations while also providing learners with authentic example-based support. By combining these two strategies, the chatbot addresses a range of pedagogical needs, from language production to contextual understanding.

Generative models based on transformer architectures, such as GPT-3, were selected for their ability to generate coherent and context-aware responses. Unlike earlier Seq2Seq models built on RNNs (e.g., LSTM, GRU), transformer-based models are better at capturing long-range dependencies and semantic structure. In this work, the GPT-3.5-turbo-0125 model was fine-tuned using OpenAI's API to adapt it to the English–Arabic translation domain. At the time of development, this model offered a strong balance between performance, accessibility, and cost-effectiveness, making it a reasonable and widely adopted choice for experimentation.

The dataset, previously developed and published in [32], comprises 12,000 English–Arabic sentence pairs curated from the SauLTC. These pairs were pre-processed and formatted in JSONL for use in fine-tuning the generative model. Data was split into training (80%), validation (10%), and test (10%) sets. Fine-tuning was conducted with 3 epochs, a batch size of 32, and a learning rate multiplier of 0.05. The final model achieved stable performance (training loss: 0.300; validation loss: 0.347). The overall workflow of the generative model is illustrated in Figure 2.

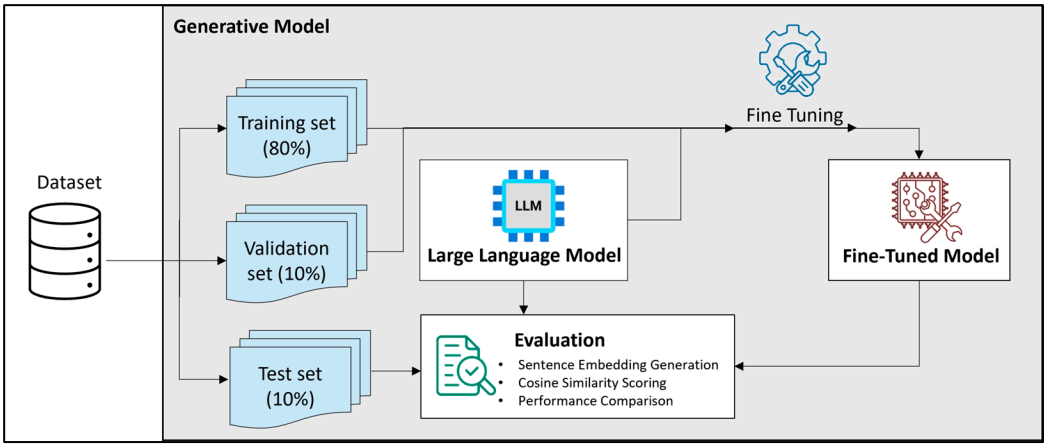

Figure 2. Generative Model Workflow.

To evaluate semantic quality, sentence embeddings were generated using the paraphrase-multilingual-mpnet-base-v2 model, and cosine similarity was computed between each English source sentence and its Arabic translations (original SauLTC reference, base GPT output, and fine-tuned GPT output). The fine-tuned model achieved a higher similarity score (0.889) than both the GPT base model (0.879) and the original SauLTC references (0.872), highlighting the benefits of domain-specific fine-tuning.

The retrieval model supports example-based learning by returning aligned Eng-lish–Arabic sentence pairs that include user-specified keywords. A keyword-based retrieval method was used, leveraging regular expression matching for full-word identification in the source corpus. Matching English sentences are paired with their corresponding Arabic translations from a curated bilingual dataset. The workflow of this retrieval model is shown in Figure 3.

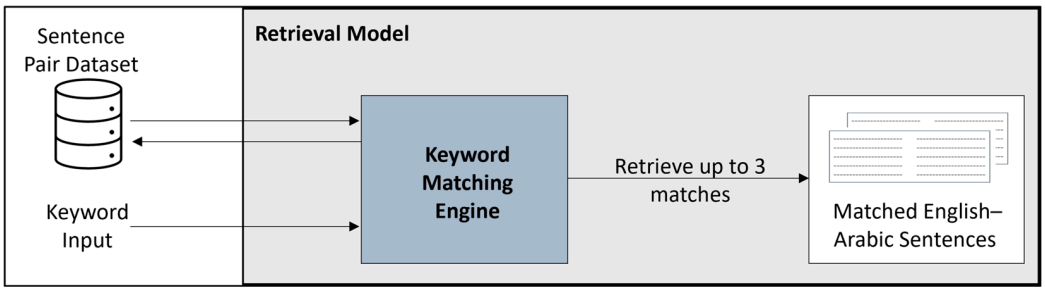

Figure 3. Workflow of the Keyword-based Retrieval Model.

This approach was selected for its interpretability, transparency, and alignment with educational goals. While embedding-based semantic search offers flexibility, the keyword approach allows precise control over lexical focus, making it more suitable for teaching language form and function.

3.1.4. Agents Design and Development

The Translation Learning Chatbot is structured around a multi-agent architecture that enables modular functionality and pedagogical flexibility. Rather than relying on a single AI model to manage all tasks, the system distributes responsibilities across specialized agents, each aligned with a distinct instructional goal. A lead agent coordinates these components, interpreting user input and routing tasks to the appropriate sub-agent. This design promotes clear task separation, system scalability, and the ability to support both guided learning and independent practice.

- Translation Agent: Employs a fine-tuned generative language model to produce accurate Modern Standard Arabic (MSA) translations from English input. Prompt-based guidance ensures linguistic correctness and con-textual appropriateness.
- Retrieval Agent: Supports example-based learning by returning English–Arabic sentence pairs containing user-specified keywords, using regular expressions for precise full-word matching.
- Review Agent: Evaluates user-submitted translations by identifying se-mantic or structural errors and providing targeted feedback, indicating whether the translation is acceptable and why.
- Feedback Agent: Facilitates self-assessment through randomly generated translation tasks, scored using sentence embeddings and cosine similarity to measure alignment with reference outputs.

Table 1 summarizes how each agent contributes to the chatbot's core functionality, highlighting the AI techniques used and the corresponding educational purpose of each component.

Table 1. Functional Mapping of Chatbot Agents to System Roles and Educational Objectives.

Functionality	Agent	Methodology	Purpose
Interactive and Task Routing	Lead Agent	Rule-based	Manage task delegation and system control
English–Arabic Translation	Translation Agent	Generative LLM	Provide context-aware, linguistically accurate translations
Example Retrieval	Retrieval Agent	Keyword-based search	Show how a word or phrase is used in context through real examples
Translation Accuracy Review	Review Agent	Generative LLM	Provide objective feedback on learner translation attempts
Translation Competence Testing	Feedback Agent	Generative LLM + Cosine similarity on embeddings	Enable self-assessment of translation accuracy

The system uses a vertical multi-agent architecture, where a lead agent coordinates several specialized sub-agents. This design was chosen not only for its modularity, which makes the chatbot easier to maintain, but also for its ability to deliver higher-quality output through specialization. Each sub-agent focuses on a specific task, such as translation, example retrieval, feedback, or evaluation. The lead agent manages context, ensuring that each sub-agent receives only the relevant information, which improves coherence and reduces irrelevant responses.

This design also supports more natural, tutor-like behaviour, mirroring how teachers work: first understanding a student's request, then selecting the most suitable teaching strategy before delivering targeted support. The modular structure makes the system scalable, enabling the addition of new agents, for example, grammar correction or back translation, without altering the entire framework. Furthermore, the independence of each agent increases robustness; since no agent depends on the output of another, a failure in one agent does not disrupt the operation of the others.

Together, these components form an integrated system that supports various aspects of translation learning, from guidance to evaluation.

3.2. Deployment Phase—Real Time Interaction

To enable real-time interaction and ensure usability, the Translation Learning Chatbot was deployed through a lightweight, browser-accessible interface built with Streamlit. This deployment integrates the components developed during the offline phase into an interactive system that supports a range of translation-related tasks. The interface was designed for accessibility, minimal setup, and smooth navigation, making it suitable for classroom, lab, or remote use.

At the core of the system is the lead agent, which interprets user input and delegates tasks to four specialized agents. This interaction model enables users to perform activities such as sentence translation, example retrieval, translation review, and competence testing. Each task is handled independently by a dedicated agent, ensuring modularity, real-time response, and system scalability.

Figure 4 illustrates the system's interactive flow, where user inputs are processed by a central lead agent and delegated to specialized agents for translation, retrieval, review, or evaluation. The modular structure ensures task-specific handling, real-time responses, and extensibility of system functionalities.

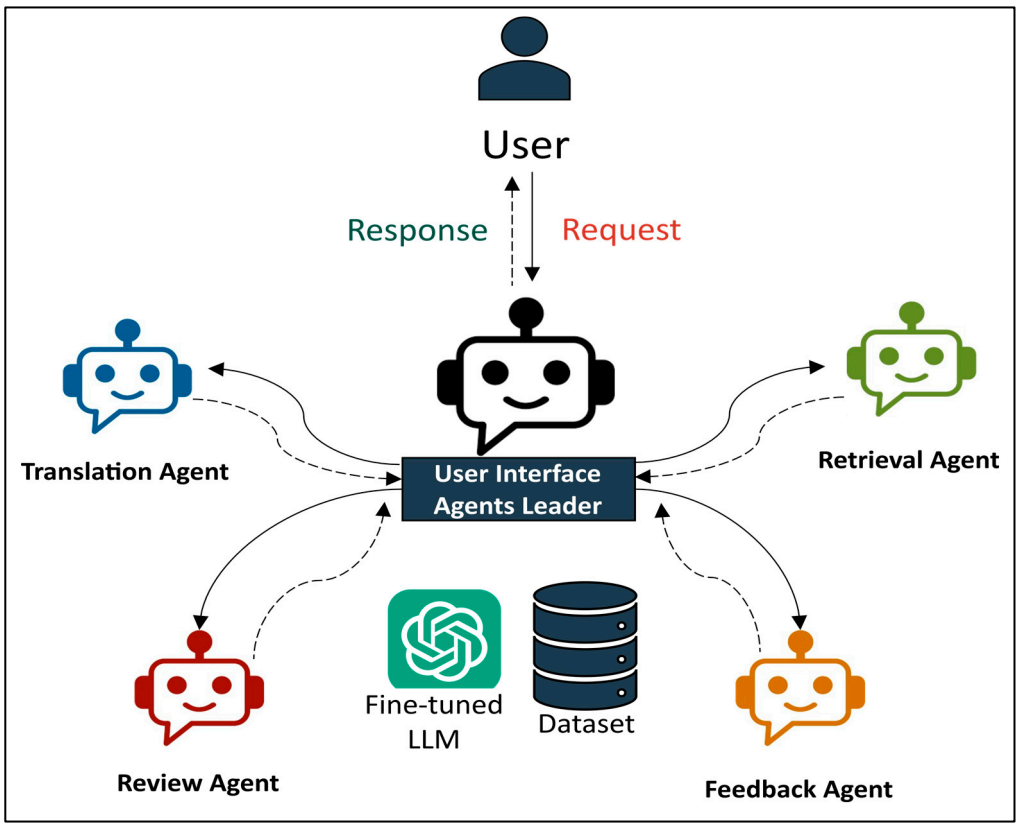

Figure 4. System Architecture of the Translation Learning Chatbot (Multi-Agent Workflow).

The interface emphasizes clarity and ease of use, with labelled task buttons, icons, and structured message formatting to reduce cognitive load and guide learners through task selection, as shown in Figure 5. A conversational flow allows users to submit queries, receive immediate responses, and access contextual feedback. The system's modular design also supports future expansion, allowing new agents or features to be added without modifying the core architecture.

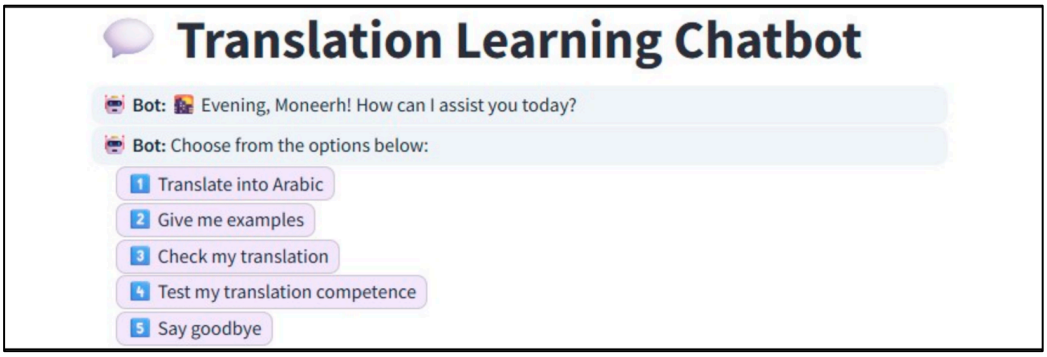

Figure 5. Sample of the Chatbot Interface.

Importantly, the modular structure of the deployed system allows for future extensions. Additional functionalities or new agent types can be introduced without altering the core architecture. This flexibility ensures that the chatbot remains adaptable to evolving learning needs and technological advancements.

3.3. Evaluation Phase: Performance Assessment

The final phase involved evaluating the Translation Learning Chatbot through structured user studies. This evaluation aimed to determine the tool's effectiveness in supporting English-to-Arabic translation learning and reducing instructor workload. The evaluation adopted a mixed-methods approach and was guided by the CHISM. It involved both translation students and faculty members from PNU, who engaged in task-based interactions with the chatbot and completed a post-use survey.

Our intention in including faculty members was not to generate statistically significant results, but rather to obtain qualitative insights and expert perspectives to complement the student data. Accordingly, faculty scores are presented descriptively and interpreted as supplementary insight rather than as statistically comparable outcomes.

3.3.1. Evaluation Objectives

The primary objective of the evaluation was to determine whether the Translation Learning Chatbot provided meaningful and pedagogically valuable support for translation learners. The evaluation focused on collecting structured feedback from users—students and faculty members—based on their direct interaction with the chatbot and their reflections captured through a post-use survey.

The specific aims of the evaluation were to:

- Assess perceived translation quality, as judged by users based on how accurate, fluent, and contextually appropriate the chatbot's Arabic translations were in response to English input.
- Evaluate the usefulness of feedback, including whether participants found the system's responses to their translations clear, relevant, and constructive.
- Measure usability and interface satisfaction, focusing on ease of navigation, response speed, and clarity of chatbot interactions across all four major tasks.
- Explore learning outcomes, particularly whether users felt more confident in their translation abilities after using the tool.
- Engagement and user satisfaction, including enjoyment during use, willingness to use the chatbot again, and likelihood of recommending it to peers or students.
- Gather user suggestions, to inform future improvements and extensions to the system's features and design.

The evaluation was structured as a user study involving two participant groups: undergraduate students and faculty members from PNU. All participants were invited to interact with the deployed chatbot and complete a structured survey designed to capture both usability and pedagogical feedback.

- Students (n = 40):

The student participants were undergraduate learners PNU, primarily in academic levels 6 to 8, and majoring in Translation or Applied Linguistics. These students were introduced to the experiment during scheduled class sessions, where the researcher was physically present. The chatbot and survey links were shared with the students in class, and they completed the tasks on-site using their personal devices. This allowed for brief clarification if needed and helped ensure participation and completion.

- Faculty members (n = 4):

This group included lecturers and assistant professors with specializations in Translation, Linguistics, and Applied Linguistics. All faculty members were familiar with AI tools and had prior experience or interest in using such technologies for translation-related tasks. Participation was entirely remote, with faculty accessing the chatbot and survey independently.

Participants were asked to interact with the chatbot by exploring its four core features:

1. Translate into Arabic: Generating MSA translations from English input using a generative LLM.
2. Give me examples: Retrieving relevant English–Arabic sentence pairs from the SauLTC corpus.
3. Check my translation: Submitting user-generated Arabic translations for feedback and similarity scoring.
4. Test my translation competence: Translating random English sentences and receiving structured performance scores.

This task-based interaction design ensured that each participant engaged with the chatbot across all functional modules.

3.3.2. Survey Structure and Assessment Criteria

The post-interaction survey was composed of two main components. The first one is the closed-ended items based on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree), designed to generate quantitative data based on the CHISM, which provides a widely recognized framework for evaluating conversational agents in educational contexts. The model categorizes user experience into three core dimensions:

- DEX: Measures the technical and visual aspects of the system, including ease of use, interface clarity, and responsiveness.
- LEX: Assesses the linguistic relevance and clarity of chatbot responses, including translation accuracy, feedback quality, and the helpfulness of contextual examples.
- UEX: Evaluates the overall learning experience, focusing on user engagement, confidence in translation tasks, enjoyment, and willingness to reuse or recommend the chatbot.

The second component is the open-ended questions to collect qualitative feedback, allowing participants to elaborate on their experience, highlight strengths, report any issues, and suggest improvements.

This mixed-method design allowed for a comprehensive evaluation of the chatbot's usability, language support, and overall learning experience.

The full list of survey questions is presented below:

- DEX:
  - The system was easy to navigate and use.
  - The chatbot responded quickly and without technical errors.
  - The interface was visually clear and user-friendly.
- LEX:
  - The chatbot's language was clear and easy to understand.
  - The Arabic translations accurately reflected the English input.
  - The examples provided were relevant and helpful.
  - The feedback on user translations was useful.
  - The similarity score helped evaluate the quality of user translations.
- UEX:
  - I felt more confident in my translation skills after using the chatbot.
  - I enjoyed using the chatbot as a learning tool.
  - I would use this chatbot again for translation practice.
  - I would recommend this chatbot to others learning translation.
- Open-ended questions:
  - What did you like most about using the chatbot?
  - What challenges or limitations did you encounter?
  - What suggestions do you have for improving the chatbot?

The faculty and student versions of the survey were identical in content, with only minor wording adjustments to reflect their respective roles.

4. Experiment and Results

Building on the experimental design outlined in Section 3.3, the evaluation involved both students and faculty members who interacted with the chatbot and provided feedback through a mixed-method survey. The analysis focuses on key chatbot functionalities and user experiences to determine the tool's potential in supporting English-to-Arabic translation learning.

While this section reports on user perceptions gathered through the CHISM survey, translation quality was also evaluated earlier in the manuscript using an embedding-based similarity approach, providing an objective complement to the survey findings.

4.1. Quantitative Findings

The quantitative results are based on survey responses from 44 participants: 40 undergraduate students and 4 faculty members. Participants were asked to rate their agreement with a series of statements across three key dimensions—DEX, LEX, and UEX—derived from the CHISM evaluation framework introduced in Section 3.3. Each item was rated using a 5-point Likert scale ranging from strongly disagree (1) to strongly agree (5). The survey also captured participants' self-reported familiarity with AI tools or chatbots used in translation tasks.

Among the student participants, 34 students (85%) indicated they were very familiar with AI tools or chatbots, while 5 students (12.5%) selected somewhat familiar, and 1 student (2.5%) marked neutral. In the faculty group, 2 members reported being very familiar, and the remaining 2 were somewhat familiar. No participants in either group reported unfamiliarity, indicating a generally high level of digital literacy relevant to the tool being evaluated.

The responses across all three dimensions reflected generally positive attitudes toward the chatbot. In the DEX dimension, over 80% of participants either agreed or strongly agreed that the chatbot was easy to navigate, visually clear, and technically responsive.

This was further confirmed by an average score of 4.3 (SD = 0.49) from students and a slightly higher 4.5 (SD = 0.47) from faculty. For the LEX dimension, more than 75% of users reported that the chatbot's language output was accurate and its examples helpful for learning, with students giving it an average score of 4.1 (SD = 0.53) and faculty 4.2 (SD = 0.50). For the UEX dimension, over 70% of participants agreed or strongly agreed that the chatbot boosted their confidence and motivation in translation tasks, reflected in mean scores of 4.0 (SD = 0.58) for students and 4.1 (SD = 0.46) for faculty.

The charts below provide a visual representation of aggregated survey responses across each dimension, with comparisons between student and faculty responses where applicable. These findings offer valuable insights into the chatbot's perceived strengths and areas for improvement.

As shown in Figure 6, faculty members rated the DEX slightly higher than students, while scores for LEX and UEX were comparable. The numerical values for each group are detailed in Table 2.

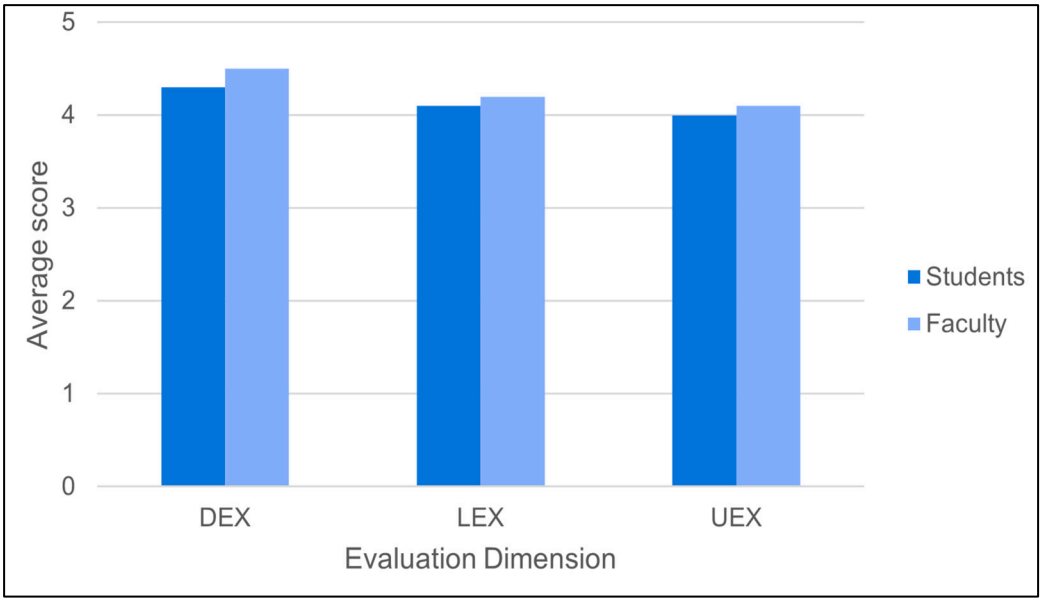

Figure 6. Average Agreement Scores Across Chatbot Evaluation Dimensions.

Table 2. Average Likert-scale scores across evaluation dimensions.

Evaluation Dimension	Students	Faculty
DEX	4.3	4.5
LEX	4.1	4.2
UEX	4	4.1

While the quantitative ratings were largely positive, some response patterns indicated areas for concern. A few participants selected "Neutral" or "Disagree" in response to statements about the usefulness of the examples and feedback, particularly among faculty members. These ratings suggest limitations in the chatbot's ability to provide contextually rich examples. Although these instances were few, they highlight the importance of addressing edge cases and advanced linguistic input in future improvements.

4.2. Qualitative User Feedback

In addition to the Likert-scale responses, participants were invited to provide open-ended feedback on their experience with the Translation Learning Chatbot. These responses were analysed thematically, with answers grouped under broader themes to identify

common trends, strengths, and areas for improvement, offering deeper insights into users' perceptions beyond numerical ratings.

Overall, positive feedback was dominant, with many participants (particularly students) noting that the chatbot was "easy to use," "fast in responding," and "helpful in understanding translation accuracy." Faculty members echoed these sentiments, describing the system as "easy and accurate," and "pedagogically sound". A recurring strength highlighted by both groups was the similarity score feature, which participants found valuable for gauging the closeness of their translations to the reference output. One student commented, "I loved that there's a similarity score, it helped a lot," while a faculty member emphasized its usefulness in assessing translation quality.

To systematically analyze these comments, a thematic coding approach was applied to 44 open-ended responses. The most frequently mentioned theme was "Ease of Use" (10 mentions), mainly raised by students who praised the chatbot's smooth navigation and intuitive design. "Translation Accuracy" (9 mentions) followed closely, with both students and faculty commending the clarity and appropriateness of the Arabic translations, though a few users noted occasional literal renderings. "Feedback Quality" (7 mentions) was another major theme, with faculty highlighting its role in supporting and improving student learning. The "Examples Feature" received 6 mentions and drew mixed feedback. While some users found the examples relevant and clear, others, particularly faculty, reported repetition or unclear context in idiomatic expressions.

Despite these strengths, some users reported specific challenges and limitations. These included repetitive or ambiguous example outputs, difficulty recognizing multi-word expressions or idioms, and rigid evaluation criteria that did not acknowledge acceptable alternative translations. For instance, one faculty member noted that a correct translation was mistakenly marked as incorrect because it did not match the expected wording.

Additional challenges were identified under the themes of "Design Improvement" (5 mentions), "System Limitations" (3 mentions), and "Instructional Clarity" (2 mentions). Suggestions for design improvement focused on enhancing the visual layout and overall interface to boost user engagement. System-related issues included occasional response delays, while instructional clarity concerns reflected the need for better guidance, particularly for new users. Furthermore, "Domain-Specific Support" (1 mention) was suggested by a faculty member, advocating for expansion into specialized translation contexts like medical or legal domains. Figures 7 and 8 present the thematic frequencies for students and faculty, respectively.

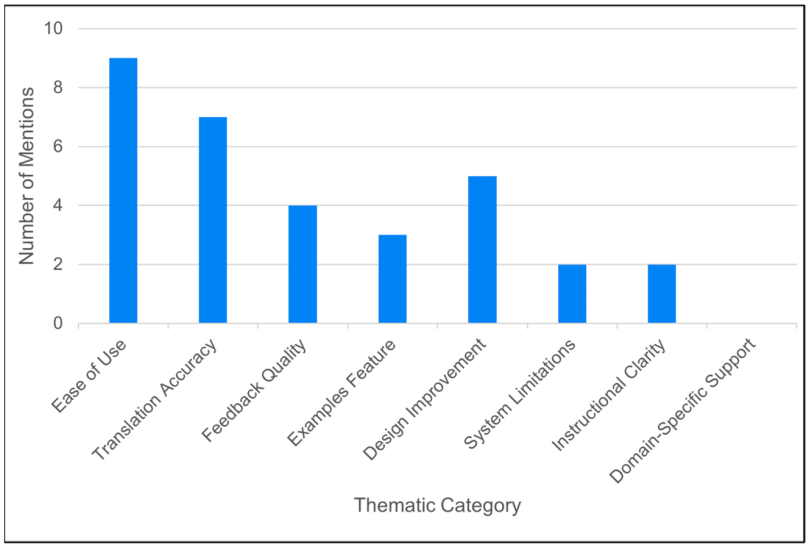

Figure 7. Frequency of Themes Mentioned in Student Survey Responses.

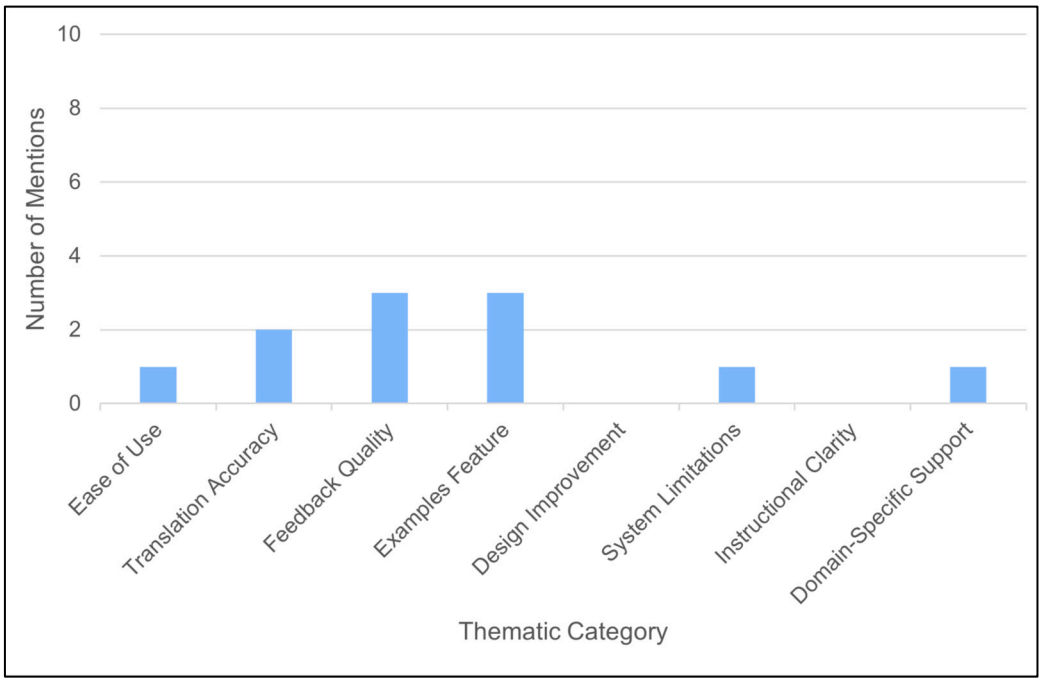

Figure 8. Frequency of Themes Mentioned in Faculty Survey Responses.

These qualitative insights complement the quantitative results by reinforcing the chatbot's core strengths while also identifying concrete areas for refinement.

5. Discussion

The user evaluation results presented in the previous sections highlight both the strengths and areas for refinement in the Translation Learning Chatbot. This discussion aims to interpret the quantitative trends and qualitative insights, offering a deeper analysis of the tool's pedagogical value, design effectiveness, and linguistic capabilities. While the general feedback confirmed the chatbot's strengths in usability and educational support, several user comments also revealed specific limitations rooted in design or linguistic complexity.

For instance, the keyword-based retrieval model sometimes returned results that were contextually unexpected or semantically different from what the user intended. One user searching for the term "winter" expected seasonal context but received sentences where "Winter" appeared as a person's name. This highlights the need for disambiguation mechanisms or refinement of the keyword-matching logic to avoid proper noun misinterpretations.

Additionally, another user noted the limited variability in example generation: "The example provided was not clear, and when I wanted more examples, I got the same one.". This suggests that the example retrieval system could benefit from controlled randomization to surface more varied and relevant contexts per search term.

Some participants also observed limitations in Arabic clarity due to the absence of diacritical marks, which can alter meanings in subtle but important ways. For example, phrases like "She is teaching" and "She is studying" are identical in undiacritized Arabic, potentially confusing learners who depend on accurate distinctions.

On a more positive note, the chatbot demonstrated semantic flexibility in recognizing both British and American English variants. In the sentence: "Do you live in a flat? Yes, it is a small apartment.". The chatbot successfully handled "flat" (British English) and "apartment" (American English) as equivalents, offering a consistent Arabic transla-

tion: "هل تعيش في شقة؟ نعم، إنها شقة صغيرة.". This reflects the system's ability to process regional variations in English and map them accurately to standard Arabic.

Interestingly, some examples revealed a disconnect between accurate human translation and the chatbot's rigid evaluation criteria. In one case, a user translated the sentence "The hare ran down the road for a while and then paused to rest" as: "ركض الأرنب في الطريق لبرهة ثم توقف ليستريح.". The chatbot flagged this translation as inaccurate and recommended replacing "لبرهة" with "لفترة." However, according to classical Arabic dictionaries such as Lisan al-Arab and Jumhurat al-Lugha, the term "برهة" denotes a meaningful or extended period of time and is entirely appropriate in this context. While both terms are valid translations of "for a while," the system failed to recognize this synonymy. This example highlights a limitation in the chatbot's evaluation component, which currently favors rigid lexical matches over nuanced semantic equivalence. Enhancing its tolerance for stylistic variation and synonym usage could significantly improve its ability to assess translation quality more accurately and fairly.

Overall, the evaluation confirms the chatbot's role as a valuable educational tool and highlights user-driven insights that will guide future iterations, ensuring the system evolves with the pedagogical and linguistic needs of its users.

6. Conclusions

This research presented the design, development, and evaluation of the Translation Learning Chatbot, an AI-powered system aimed at supporting English-to-Arabic translation education. The chatbot integrates both retrieval-based and generative language models to provide learners with translation practice, feedback, and self-assessment within an interactive learning environment.

The chatbot adopts a modular, multi-agent architecture composed of distinct components responsible for key functions: translation generation, example retrieval, translation evaluation, and translation competence testing. It leverages a fine-tuned large language model, sentence similarity scoring using multilingual embeddings, and keyword-based search across a bilingual learner corpus. This hybrid framework was designed to address core challenges in translation education, particularly the need for meaningful context, consistency in feedback, and support for self-learning.

The chatbot was evaluated through a mixed-method study involving 44 participants, including students and faculty members. The system received high ratings across all core areas, with average scores exceeding 4.0 out of 5 in usability, language quality, and engagement. Over 80% of users expressed satisfaction with its educational value, and the similarity score feature was especially effective in promoting self-assessment and learner motivation. While overall feedback was positive, some limitations were noted, including difficulty with idiomatic expressions, rigid evaluation rules, and keyword-matching issues. These insights not only validate the effectiveness of the system but also inform priorities for future development.

Future work may focus on several directions to enhance the chatbot's educational value and scalability. These include embedding a structured error classification system to provide more precise, linguistically grounded feedback; expanding domain coverage to specialized translation tasks such as legal, business, or technical texts through domain-specific corpora and fine-tuning; and incorporating multimodal inputs (image, audio, video) to simulate realistic translation scenarios. Additional improvements involve refining, example retrieval with semantic disambiguation and contextual filtering, integrating diacritical marks in Arabic output to support beginner learners, improving the visual layout of the chatbot for enhanced usability and adopting adaptive feedback mechanisms based on reinforcement

learning. Finally, embedding the chatbot into existing Learning Management Systems (e.g., Blackboard) as a plugin would facilitate large-scale institutional adoption, personalized learning, and instructor monitoring, while also exploring advanced evaluation metrics such as BERTScore and related approaches to provide richer and more reliable assessments of translation quality.

Overall, this study contributes a novel, scalable, and pedagogically aligned chatbot framework tailored to the needs of English–Arabic translation learners, demonstrating the potential of AI-driven tools to enhance digital translation education and support self-directed language learning.

Author Contributions: Conceptualization, M.A., E.A. and S.M.; methodology, M.A., E.A. and S.M.; software, M.A.; validation, M.A.; formal analysis, M.A.; investigation, M.A.; data curation, M.A.; writing—original draft preparation, M.A.; writing—review and editing, M.A., E.A. and S.M.; visualization, M.A.; supervision, E.A. and S.M.; funding acquisition, S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R196), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Princess Nourah bint Abdulrahman University (IRB log number: 21-0388 and date of approval: 14 October 2021, category of approval: Exempt).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data will be available upon request from the corresponding authors due to privacy.

Acknowledgments: The authors gratefully acknowledge Maha Al-Harthi and Fatma Alshihri for their valuable support as specialists in linguistics.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BRNN	Bidirectional Recurrent Neural Network
BUS-11	Chatbot Usability Scale
CHISM	Chatbot-Human Interaction Satisfaction Model
DEX	Design Experience
GPT	Generative Pre-trained Transformer
HUSE	Human Unified with Statistical Evaluation
LEX	Language Experience
LLM	Large Language Model
MAUVE	Model-based Automatic Unsupervised Validation Evaluation
MSA	Modern Standard Arabic
MT	Machine Translation
NMT	Neural Machine Translation
PNU	Princess Nourah bint Abdulrahman University
SauLTC	Saudi Learner Translation Corpus
UEX	User Experience
USR	Unsupervised and Reference-free Score

References

1. PNU. Princess Nourah Bint Abdulrahman University. Available online: https://pnu.edu.sa/en/AboutUniversity/Pages/About.aspx (accessed on 26 September 2025).
2. Weizenbaum, J. ELIZA-A computer program for the study of natural language communication between man and machine. *Commun. ACM* **1966**, *9*, 36–45. [CrossRef]
3. Shum, H.Y.; He, X.D.; Li, D. *From Eliza to XiaoIce: Challenges and Opportunities with Social Chatbots*; Zhejiang University: Hangzhou, China, 2018. [CrossRef]
4. Makhkamova, O.; Lee, K.H.; Do, K.; Kim, D. Deep learning-based multi-chatbot broker for q&a improvement of video tutoring assistant. In Proceedings of the 2020 IEEE International Conference on Big Data and Smart Computing, BigComp 2020, Busan, South Korea, 19–22 February 2020; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2020; pp. 221–224. [CrossRef]
5. Nuruzzaman, M.; Hussain, O.K. A Survey on Chatbot Implementation in Customer Service Industry through Deep Neural Networks. In Proceedings of the 2018 IEEE 15th International Conference on e-Business Engineering, ICEBE 2018, Xi'an, China, 12–14 October 2018; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2018; pp. 54–61. [CrossRef]
6. Qiu, M.; Li, F.-L.; Wang, S.; Gao, X.; Chen, Y.; Zhao, W.; Chen, H.; Huang, J.; Chu, W. AliMe Chat: A Sequence to Sequence and Rerank based Chatbot Engine. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 498–503. [CrossRef]
7. Serban, I.V.; Sankar, C.; Germain, M.; Zhang, S.; Lin, Z.; Subramanian, S.; Kim, T.; Pieper, M.; Chandar, S.; Ke, N.R.; et al. A Deep Reinforcement Learning Chatbot. 2017. Available online: http://arxiv.org/abs/1709.02349 (accessed on 1 December 2020).
8. Lin, Z.; Xu, P.; Winata, G.I.; Siddique, F.B.; Liu, Z.; Shin, J.; Fung, P. CAiRE: An Empathetic Neural Chatbot. Available online: http://arxiv.org/abs/1907.12108 (accessed on 17 January 2021).
9. Hien, H.T.; Cuong, P.N.; Nam, L.N.H.; Nhung, H.L.T.K.; Thang, L.D. Intelligent assistants in higher-education environments: The FIT-EBOt, a chatbot for administrative and learning support. In *ACM International Conference Proceeding Series*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 69–76. [CrossRef]
10. Yang, S.; Wang, Y.; Chu, X. A Survey of Deep Learning Techniques for Neural Machine Translation. *arXiv* **2020**, arXiv:2002.07526. Available online: https://arxiv.org/abs/2002.07526v1 (accessed on 4 January 2022). [CrossRef]
11. Wu, T.; He, S.; Liu, J.; Sun, S.; Liu, K.; Han, Q.-L.; Tang, Y. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. *IEEE/CAA J. Autom. Sin.* **2023**, *10*, 1122–1136. [CrossRef]
12. Zakraoui, J.; Saleh, M.; Al-Maadeed, S.; Alja'am, J.M. Evaluation of Arabic to English Machine Translation Systems. In Proceedings of the 2020 11th International Conference on Information and Communication Systems, ICICS 2020, Irbid, Jordan, 7–9 April 2020; pp. 185–190. [CrossRef]
13. Alkhatib, M.; Shaalan, K. The Key Challenges for Arabic Machine Translation. *Stud. Comput. Intell.* **2018**, *740*, 139–156. [CrossRef]
14. Ducar, C.; Schocket, D.H. Machine translation and the L2 classroom: Pedagogical solutions for making peace with Google translate. *Foreign Lang. Ann.* **2018**, *51*, 779–795. [CrossRef]
15. Satpathy, S.; Mishra, S.P.; Nayak, A.K. Analysis of learning approaches for machine translation systems. In Proceedings of the 2019 International Conference on Applied Machine Learning, ICAML 2019, Bhubaneswar, India, 25–26 May 2019; pp. 160–164. [CrossRef]
16. Groves, M.; Mundt, K. A ghostwriter in the machine? Attitudes of academic staff towards machine translation use in internationalised Higher Education. *J. Engl. Acad. Purp.* **2021**, *50*, 100957. [CrossRef]
17. Rescigno, A.A.; Monti, J.; Way, A.; Vanmassenhove, E. A Case Study of Natural Gender Phenomena in Translation: A Comparison of Google Translate, Bing Microsoft Translator and DeepL for English to Italian, French and Spanish. In *Workshop on the Impact of Machine Translation (iMpacT 2020)*; Association for Machine Translation in the Americas (AMTA): Washington, DC, USA, 2020; pp. 62–90. Available online: https://aclanthology.org/2020.amta-impact.4 (accessed on 9 November 2021).
18. Lee, S.-M. The impact of using machine translation on EFL students' writing. *Comput. Assist. Lang. Learn.* **2020**, *33*, 157–175. [CrossRef]
19. Deriu, J.; Rodrigo, A.; Otegi, A.; Echegoyen, G.; Rosset, S.; Agirre, E.; Cieliebak, M. Survey on Evaluation Methods for Dialogue Systems. *Artif. Intell. Rev.* **2019**, *54*, 755–810. [CrossRef] [PubMed]
20. Belda-Medina, J.; Kokošková, V. Integrating chatbots in education: Insights from the Chatbot-Human Interaction Satisfaction Model (CHISM). *Int. J. Educ. Technol. High. Educ.* **2023**, *20*, 63. [CrossRef]
21. Borsci, S.; Schmettow, M. Re-examining the chatBot Usability Scale (BUS-11) to assess user experience with customer relationship management chatbots. *Pers. Ubiquitous Comput.* **2024**, *28*, 1033–1044. [CrossRef]
22. Følstad, A.; Nordheim, C.B.; Bjørkli, C.A. What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study. In *Internet Science*; Bodrunova, S.S., Ed.; Springer International Publishing: Cham, Switzerland, 2018; pp. 194–208.
23. Sai, A.B.; Mohankumar, A.K.; Khapra, M.M. A Survey of Evaluation Metrics Used for NLG Systems. *ACM Comput. Surv.* **2022**, *55*, 1–39. [CrossRef]

24. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. Bleu: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics—ACL '02, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318. [CrossRef]
25. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. 2004. Available online: https://aclanthology.org/W04-1013/ (accessed on 18 June 2025).
26. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. *arXiv* **2019**, arXiv:1904.09675. Available online: http://arxiv.org/abs/1904.09675 (accessed on 18 June 2025).
27. Sellam, T.; Das, D.; Parikh, A. BLEURT: Learning Robust Metrics for Text Generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 7881–7892. [CrossRef]
28. Mehri, S.; Eskenazi, M. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), Online, 5–10 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 681–707. [CrossRef]
29. Pillutla, K.; Swayamdipta, S.; Zellers, R.; Thickstun, J.; Choi, Y.; Harchaoui, Z. MAUVE: Human-Machine Divergence Curves for Evaluating Open-Ended Text Generation. *arXiv* **2021**, arXiv:2102.01454. Available online: https://arxiv.org/abs/2102.01454 (accessed on 18 June 2025).
30. Hashimoto, T.B.; Zhang, H.; Liang, P. Unifying Human and Statistical Evaluation for Natural Language Generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 1689–1701. [CrossRef]
31. Al-Harthi, M.; Al-Saif, A. The Design of the SauLTC application for the English-Arabic Learner Translation Corpus. In *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 80–88. Available online: https://aclanthology.org/W19-5610 (accessed on 8 December 2021).
32. Aleedy, M.; Alshihri, F.; Meshoul, S.; Al-Harthi, M.; Alramlawi, S.; Aldaihani, B.; Shaiba, H.; Atwell, E. Designing AI-Powered Translation Education Tools: A Framework for Parallel Sentence Generation Using SauLTC and LLMs. *PeerJ Comput. Sci.* **2025**, *11*, e2788. [CrossRef] [PubMed]