Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 17 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i17.62615

PAPER

Developing a Chatbot for Soft Skills Education: Comparing ReLU-LSTM and GloAT-Transformer Models

Chaimaa Bouafoud[1] (✉), **Abdellah Madani**[1], **Khalid Zine-Dine**[2]

[1]LAROSERI Laboratory, Chouaïb Doukkali University – El Jadida, Morocco

[2]Faculty of Sciences Rabat (FSR), Mohammed V University – Rabat, Morocco

bouafoud.chaimaa@ucd.ac.ma

ABSTRACT

Recently, soft skills, such as communication, teamwork, and problem-solving, have become more important in higher education. In Moroccan universities, fostering these skills is essential to better prepare students for the demands of the job market and modern work environments. However, traditional educational methods often lack interactive and personalized approaches to develop these competencies. This led us to develop a conversational AI system to support soft skills learning through intent classification using natural language processing (NLP) and neural networks (NN). Two deep learning models were developed and evaluated; a ReLU-Enhanced long short-term memory (LSTM), optimized for stable sequential processing, and a GloAT-Transformer model leveraging global self-attention to capture contextual relationships.

KEYWORDS

artificial intelligence (AI), deep learning, natural language processing (NLP), chatbots, education, soft skills, classification intent, neural network (NN), long short-term memory (LSTM), transformers

1 INTRODUCTION

Conversational artificial intelligence (AI) bots represent a significant advancement in information technology, designed to enhance user interaction by simplifying communication between services and end-users through chat interfaces. These bots use natural language processing (NLP) to understand user queries. They can provide text-based answers or perform actions such as booking flights or checking emails. Industries such as customer service and banking are increasingly adopting this technology to streamline operations and improve user experience [1, 2].

In the educational sector, chatbots hold immense potential as virtual assistants supporting academic and administrative activities from students and staff alike. In the Moroccan higher education context, the National Pact for Higher Education Acceleration (PACTE ESR 2030) emphasizes the digitalization of learning and the reinforcement of soft skills [38]. However, Moroccan universities face structural and pedagogical challenges, including large and heterogeneous class sizes, multilingual learning

Bouafoud, C., Madani, A., Zine-Dine, K. (2026). Developing a Chatbot for Soft Skills Education: Comparing ReLU-LSTM and GloAT-Transformer Models. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(17), pp. 4–25. https://doi.org/10.3991/ijim.v20i17.62615

Article submitted 2026-05-10. Revision uploaded 2026-07-01. Final acceptance 2026-07-01.

environments (Arabic, French, and English), and limited opportunities for individualized learner support. These constraints hinder effective soft skills development.

The massive influx of students increases student-to-teacher ratios, reinforcing the need for scalable and adaptive educational solutions. The integration of AI-based tools becomes particularly relevant for enhancing student engagement and supporting personalized learning pathways (see Figure 1). This study addresses this gap by providing an automated, scalable, retrieval-based chatbot solution tailored to the informational and pedagogical needs of Moroccan university students.

Artificial intelligence replicates human-like intelligence, including learning, reasoning, and understanding natural language. The journey of AI began in the 1950s with early programs such as ELIZA, which simulated human conversation. Over the decades, advances have led to sophisticated chatbots capable of handling complex queries across various domains including healthcare, consumer services, and education [3].

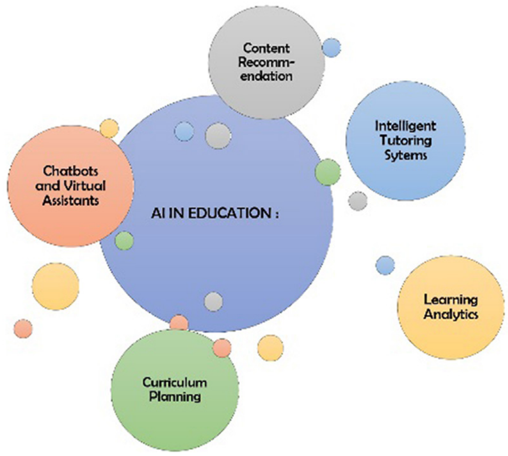

Fig. 1. Benefits of AI in personalized learning

The use of intelligent chatbot systems, often based on frameworks such as information technology infrastructure library (ITIL), is increasingly being explored to improve information technology governance in universities [35].

The development of a conversational agent capable of supporting the learning of soft skills represents a complex engineering challenge, the first crucial step of which is intention modeling. We consider intention to be the generation of a semantic label that determines the final response. The goal of our research is to identify the most effective neural network (NN) architecture for mapping a textual query to this semantic label, thereby conditioning the retrieval of an appropriate response from a structured knowledge base.

The specific contributions of this article are structures as follows:

- Conceptualization and implementation of an intent classification system for a conversational agent providing information on soft skills, based on a retrieval-based architecture.
- Conducting an empirical comparative analysis between two deep NN architectures—a standard LSTM model and a Transformer model—to evaluate their respective effectiveness in the task of intent classification.
- Critical discussion and perspective on the methodological challenges inherent in the use of small, domain-specific datasets in the context of deep learning applied to education.

The rest of the paper is structured as follows: Section 2 analyzes the existing approaches for building chatbots, highlighting their role in the education sector, with a focus on various deep learning and NLP architectures applied to the question-answering task. Sections 3 and 4 present the proposed system architecture and interpret the experimental results. Section 5 concludes the paper by summarizing key findings and potential future work.

2 RELATED WORK

This section provides an overview of the current state of the various types of chatbots and their applications, focusing on their taxonomy, their deployment in field of education. It distinguishes between open-domain and closed-domain chatbots. Adding to that, the section discusses the potential of LSTM and transformer models for revolutionizing education through personalized learning, homework assistance, language learning, research support, and virtual mentoring.

2.1 Taxonomy and application domains of chatbots

Conversational AI aims to create agents that converse such as humans. This means the generation of responses that sound natural, it also needs to ensure that such responses are informative and relevant to whatever the user has queried about. According to [6], the answers should be grounded in external knowledge to ensure accurate and contextually appropriate information. In addition to naturalness, conversational AI has to execute some degree of control over its generated outputs. This includes maintaining a certain style or tone throughout the responses. The effect of controlled style in communication will be important, particularly in areas related to customer care where consistency in the voice of brands is imperative.

Conversational AI systems must be capable of two things: understanding the user input—key questions or statements—and responding appropriately. This is deeper than just keyword recognition, it requires an understanding of language, context, and user intent. According to [7], the model should understand conversational history to engage in efficient multi-turn dialogue. That means the system has to remember previous utterances in a conversation so that its responses are coherent and contextually relevant.

Multi-turn conversations pose a very interesting challenge in this regard for conversational AI. For good dialogue management, the system needs to follow the course of a conversation, remember how the dialogue has gone so far, and react accordingly. This makes it way more user-friendly since this is tantamount to having a natural, free-flowing conversation, such as among humans.

The area of human-computer interaction, represented by chatbots, can be regarded as one of the fastest-growing ones, implemented in a line of sectors: e-commerce [8, 9], education [10], banking [11], and healthcare [12]. They enhance the essence of such interactions because, working among them, an interactive response to users takes place.

Fields of application. Chatbots are mainly distinguished by their scope of application and their response generation mechanism [5]. Generally speaking, there are mainly two approaches to developing chatbots: open-domain and closed-domain (see Figure 2).

- **Open-domain:** These systems are designed to support unconstrained dialogues on a multitude of topics (e.g., large generative language models). They require sophisticated generative architectures and massive training corpora.

- **Closed domain:** The thematic scope of the conversation is strictly defined (e.g., customer support, educational tutoring). Our work falls squarely into this category, targeting information on soft skills.

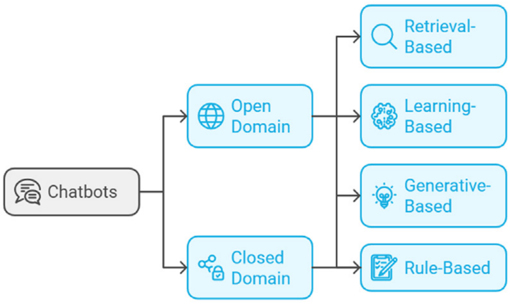

Fig. 2. Chatbot categorization

Response mechanisms. Chatbot systems can also be divided into rule-based and AI-based mechanisms. Rule-based chatbots have limited flexibility due to the pre-programming of their responses. Conversely, AI-based chatbots apply enhanced retrieval, learning, and generative techniques. A retrieval based chatbot would search for the best response among the entries in the dataset, whereas a learning-based model learns patterns from provided training data in deciding on what to reply in context. As its name indicates, a generative model generates responses by predicting sequences of words conditioned on input.

- **Rule-based** chatbots analyze the inputs provided by the user and install some predefined responses, which are wholly based on the recognition of particular variables by using pre-set rules [13]. Their methodologies can be divided into the techniques of pattern matching and task-oriented systems [14]. The task-oriented systems lead the user through various steps to enable them to perform certain pre-defined tasks. One of the significant weaknesses of rule-based chatbots includes silliness and an inability to learn anything new, depending on pre-canned responses to answer questions.
- **AI-based** chatbots; these systems use machine learning to optimize the relevance of interactions. They fall into three main categories:
    - **Retrieval-based** approaches that aim at an ideal response chosen among a previously assembled dataset of conversational knowledge [15]. It classifies the user input into a specific intent, then retrieves the most appropriate response from pre-existing knowledge base (FAQ). This is the methodological paradigm adopted for this study.
    - **Learning-based** methods gained popularity in academia, they face difficulty in representing all useful information within a fixed-length vector in cases where either the sentence or the dialogue is longer than 20 words. Different approaches were suggested to solve this problem, including structural modifications, improvement of encoder and decoder architectures, and attention mechanisms [16, 17]. The best solution so far seems to be methods that use transformers, since they brought a lot of improvements, acting through multi-headed self-attention.
    - **Generative approaches**, such as SeqGAN [18] and StepGAN [19], rely on reinforcement learning but face challenges such as slow speed and convergence, especially for long inputs. [20] proposes an effective novel architecture based on cWGAN and uses a transformer model as the backbone to process the responses generated from a retriever component to return more context-specific responses. While huge progress has been achieved in models such as ChatGPT,

there are still many challenges in closely mimicking human conversational features for particular domains. This is especially true for specialized domains, such as religious texts, where pretraining models such as BERT on specific translations (e.g., Qur'an and Hadith) is necessary to improve performance [36]. Furthermore, the development of context-based AI in chatbots is crucial for conveying personalized interdisciplinary knowledge to users [37].

2.2 Chatbots in instructional design

Due to the growing influence of AI and its widespread applications across almost every field, it has been come increasingly important to develop an AI-focused curriculum. [4] sought to create an educational environment for AI and identified three key competencies for primary and secondary school students: knowledge, skills, and attitudes. AI knowledge refers to understanding fundamental concepts and various types of AI, AI skills emphasize the ability to use AI-based tools effectively, and AI attitudes reflect an awareness of AI's societal impact and a willingness to collaborate with intelligent systems. This framework aims to enhance students' overall understanding of AI, as even those who do not intend to pursue careers in technology should possess a basic comprehension of how AI systems function and are applied.

Current efforts to promote AI literacy at the primary and secondary school levels mainly focus on algorithmic reasoning and practical uses of AI. Tools such as Google's "Teachable Machine" are frequently used to illustrate how AI models acquire knowledge from datasets, while learners are introduced to everyday AI applications such as YouTube's recommendation algorithm [5]. Although these methodologies assist students in developing a foundational comprehension of AI, a gap remains between the educational content provided and the expectations of AI practitioners.

At the higher education level, the integration of AI into learning environments is expanding rapidly, offering solutions for cognitive assistance, language learning, and personalized mentoring [8]. Recent studies confirm the potential of conversational agents to personalize learning paths and provide immediate, formative feedback [9].

In the specific context of soft skills, the objective extends beyond the simple transmission of definitions toward the simulation of complex and contextual interaction. Although our proposed system is retrieval-based, it represents a critical initial step in structuring domain knowledge and validating intent classification, which is the foundation of any structured dialogue system. The explicit incorporation of pedagogical frameworks (e.g., Bloom's taxonomy and Kolb's experiential learning cycle) is a necessary perspective for evolving chatbots into effective active learning tools [10].

2.3 Chatbot types deep learning architectures for NLP

Owing to the advancements in Deep Learning, NN have established themselves as the preeminent methodology across a multitude of fields, encompassing Image Recognition and NLP. Within the domain of NLP, NN are employed for a diverse array of functions, including text classification, machine translation, and named entity recognition (NER).

Recurrent Neural Networks (RNN) & Long Short-Term Memory (LSTM). Neural networks utilized in NLP can be delineated into two fundamental operational categories: they may serve for end-to-end system training, as evidenced by generative chatbot models, or they may be incorporated into discrete components of a system. For example, recurrent neural networks (RNNs) possess a distinguished capability for managing sequential data, rendering them particularly effective for applications such as text generation and sentiment analysis [21, 22].

Furthermore, contemporary innovations have underscored the efficacy of transformer-based architectures, including BERT and GPT, which have profoundly transformed the capabilities of language comprehension and generation within conversational agents [23]. These models not only augment the efficacy of chatbots but also enable more sophisticated interactions by assimilating knowledge from an extensive corpus of text data.

This incorporation of NN within NLP has markedly altered the mechanisms by which machines interpret and comprehend human language, facilitating a broad spectrum of applications that persist in advancing through ongoing scholarly inquiry and technological development in this area.

The long short-term memory (LSTM) model is a specialized type of recurrent neural network (RNN) designed to handle sequential data. It employs gate mechanisms to regulate the flow of information, effectively addressing the issues of vanishing and exploding gradients commonly found in traditional RNNs [24].

The basic LSTM architecture, depicted in Figure 3, consists of three types of gates: input gates, forget gates, and output gates. These gates selectively control the passage of information through the network, enabling the LSTM to manage long sequences and retain recurring patterns over extended periods.

This unique feature makes LSTM particularly well suited for financial price modeling, as it can capture long term dependencies present in market data. Numerous researchers and industry professionals have applied this foundational architecture in the financial domain. For instance, several studies (e.g., [25], [26], [27], and [28]) have used LSTM for predicting financial market trends. In some cases, such as [27], the basic LSTM structure was enhanced by combining it with stacked autoencoders, yielding improved results. Other approaches, such as [29], have employed bidirectional LSTM (BiLSTM) to boost prediction accuracy by learning temporal dependencies in both forward and backward directions within time series data. This LSTM variant is especially useful when the problem requires a comprehensive understanding of the entire time series at various iterations.

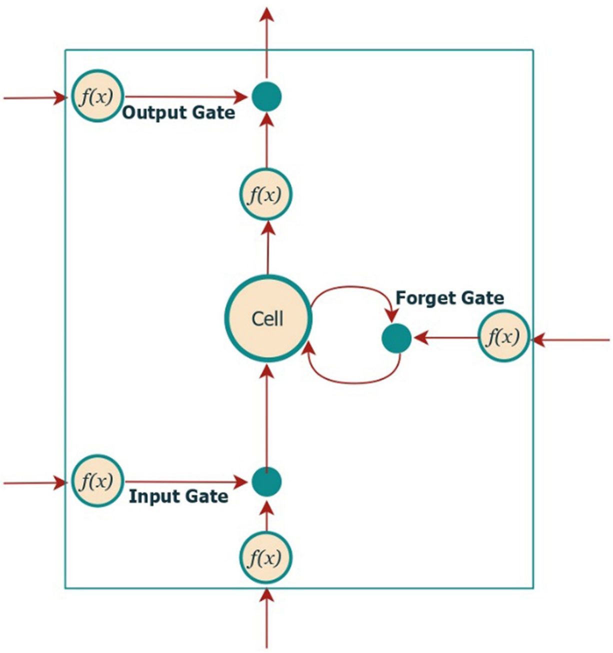

Fig. 3. A basic LSTM structure

Transformer-based model. While RNNs and LSTMs have been recognized for a long time as powerful paradigms in language modeling, they have the disadvantage of relying on computation along the symbol positions of input and output sequences in a serial manner. This, by nature, inherently prevents parallelization of training across samples, which slows down the learning process. To tackle this issue, there have also been bidirectional variants developed for such models, allowing sentences to be encoded both from left to-right and from right-to-left. However, according to the study by [30], such an action does not help in showing any significant improvement when one is dealing with long-range dependencies. Attention mechanisms [31] were developed to model word dependencies without length constraints. Then, the transformer architecture cropped up as a revolutionary model that solely relies on attention mechanisms without recurrent neural networks.

The model consists of two main components: the encoder and the decoder (see Figure 4), which can be used independently for specific tasks. For instance, encoder-only models, which can be pre-trained with a few more training or fine-tuning, are ideal for tasks that require insight with deeper input analysis, such as sentence classification and NER. Decoder-only models are for generative tasks, including text generation. These applications, such as translation or summarization, rely on understanding the input and generating the output, which is enabled by the encoder-decoder models, also called sequence-to-sequence models. Most of these models have been used to improve performance in the GLUE bench mark tests, which test natural language understanding on ten tasks, including sentiment analysis and question answering, among others [32].

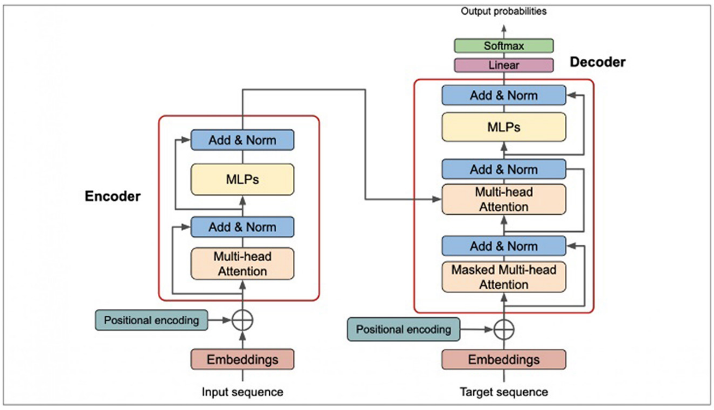

Fig. 4. The transformer model architecture [31]

Some transformers, such as GPT-3, are trained on massive datasets such as Common Crawl and have hundreds of billions of parameters [33], others, such as DistilBERT, were designed on a much smaller scale and for more convenient application in real world scenarios [34]. This great variety of transformer architectures allows for a wide range of applications to be provided within NLP, serving both the most complex and simple tasks with ease.

3 METHODOLOGY

This section explains the functional architecture of our system, details how we'll deploy our chatbot application, and outlines the workflow of the entire solution. It begins by addressing the problem of building a chatbot designed to understand and respond to user queries. The dataset used for training and evaluation is described, including its structure and source. The system's roadmap involves comparing two neural architectures—ReLU-Enhanced LSTM and Global Attention Transformer (GloAT-Transformer)—to enhance performance in classification tasks.

3.1 Problem definition

The study suggests an alteration of the architectures and parameters associated with these deep learning models, with the objective of enhancing outcomes. The investigation was systematically executed in four phases. Initially, the collection of data was initiated, resulting in the acquisition of the dataset from multiple sources. Secondly, the procedure of data preprocessing is crucial for the enhancement of the gathered data, ensuring its conformity with the criteria established by the specified models. Third, the architectures of ReLU-Enhanced LSTM and GloAT-Transformer were designed, implemented, and trained through parameter optimization to achieve optimal results. Finally, predictions are made based on the trained models, and their performance is evaluated in terms of accuracy, efficiency, and contextual understanding for NLP tasks.

This paper conducts a comparative study between LSTM and Transformer-based models to build a chatbot capable of understanding user queries and retrieve a corresponding response through intent classification. The chatbot is trained on a dataset of predefined intents, each intent represents a category of user queries associated with examples of patterns and corresponding responses. The study aims to evaluate the effectiveness of sequential modeling using LSTM architectures versus the ability of Transformer-based models to capture long-range dependencies and contextual relationships in textual data. To improve performance, architectural modifications and parameter optimization were applied to both models, namely the Relu-Enhanced LSTM and the Global Attention Transformer.

The research methodology was done in four stages. First, data collection, where the dataset was obtained from various sources. Second, data preprocessing ensures that the collected data is cleaned, structured and prepared to meet the input requirements of the defined models. Third, the modified architectures of ReLU-Enhanced LSTM and GloAT-Transformer are designed, implemented, and trained through parameter optimization to achieve optimal results. Finally, predictions are generated used the trained models, and their performance is evaluated in terms of accuracy, efficiency, and contextual understanding for NLP tasks.

3.2 System architecture

The developed system is a Retrieval-based chatbot with a strictly sequential processing flow:

- **User input:** Receipt of the text query.
- **Preprocessing:** Tokenization and preparation of the sequence for model input.
- **Intent classification:** The query is processed by the deep learning model (LSTM or Transformer), which predicts one of N = 38 predefined intents.
- **Response retrieval:** The classified intent is used as a key to retrieve a corresponding response from the knowledge base (JSON file).
- **Response delivery:** Display of the response to the user.

As illustrated in Figure 7, the system architecture is designed to handle NLP tasks, with a focus on intent extraction and response retrieving. At its core, the system relies on a NN model trained to process and understand text data, performing tasks such as intent classification and language parsing.

The training process uses a dataset that enables the model to handle diverse inputs effectively. The NLP component extracts the intent from user input, which is then analyzed to give appropriate responses. If the system cannot find a suitable answer internally, it may search external sources such as Wikipedia using tools such as Selenium Web Driver, as shown in the figure. User communication is managed through an HTTP server, and fallback responses are provided when the system fails to interpret the input correctly.

3.3 Data corpus and preprocessing

The data corpus constitutes a central component of our study, as its characteristics directly affect the validity and applicability of the experimental results. To train the NN models used in the chatbot, a dataset containing predefined intents and corresponding responses was required.

Corpus origin and structure. The initial dataset was aggregated from heterogeneous sources, including technical forums and online educational resources, notably e-learning FAQ data derived from Stack Overflow and related platforms available on Kaggle. Raw data was then collected and converted from CSV to JSON format to ensure compatibility with the system architecture. The dataset is structured as JSON file, comprising a list of objects. Each object represents an intent and contains three main fields: tag (intent label), patterns (a list of sample phrases associated with the intent), and responses (a list of predefined corresponding responses). An overview of this structure is illustrated in Figure 5.

Following data aggregation, a rigorous manual filtering and adaptation process was conducted to content with 38 predefined soft skills related intents. The resulting corpus contains approximately 300 question-answer pairs, corresponding to 1278 user questions and 300 responses, distributed across the 38 intent classes, as shown in Figure 6.

Preprocessing and data partitioning. Prior to model training, the text data underwent a preprocessing phase consisting primarily of tokenization and sequence preparation. The corpus was then divided into training, validation, and testing subsets using a standard split of 80%, 10%, and 10%, respectively. This partitioning strategy is essential for assessing the models' ability to generalize to unseen data and for ensuring reliable performance evaluation.

```
"intents": [
    {
        "tag": "soft_skills_definition",
        "patterns": [
            "What are soft skills?",
            "Can you define soft skills?",
            "Explain soft skills",
            "What does soft skills mean?",
            "Tell me about soft skills"
        ],
        "responses": [
            "Soft skills are interpersonal skills that help you interact effectively
            "Soft skills include communication, teamwork, problem-solving, and adapt
            "They are non-technical skills that influence how you work and interact
            "Soft skills are personal attributes that enhance your relationships and
        ],
        "context": [""]
    },
```

Fig. 5. Dataset overview

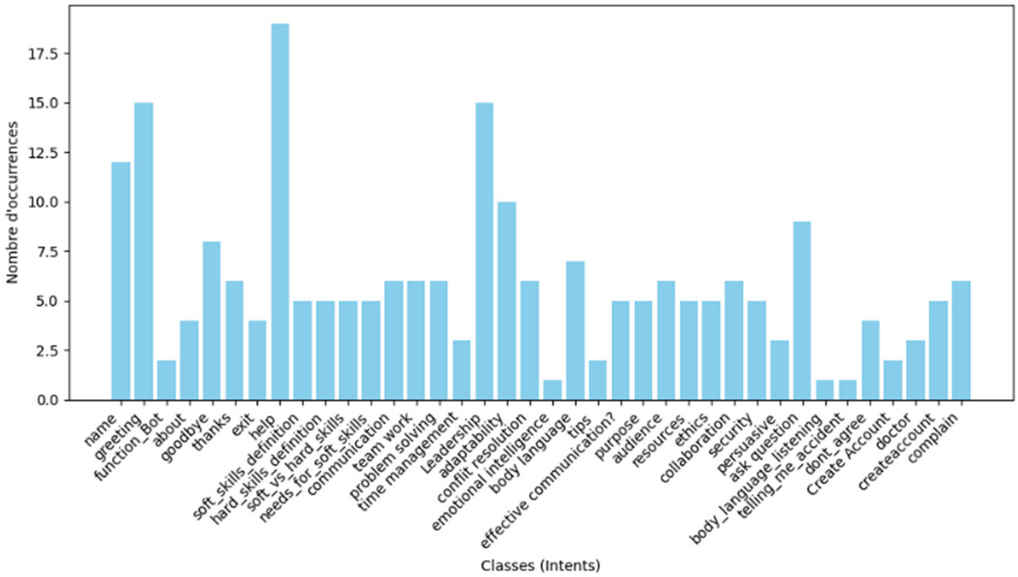

Fig. 6. Class distribution

3.4 Classification model specifications

The proposed system relies on two deep learning architectures for intent classification: a LSTM-based model, well suited for sequential data modeling, and a Transformer-based model, which excels at capturing long-range dependencies and complex relationships in textual data. This comparative approach ensures as objective evaluation of the strengths and limitations of recurrent and attention-based approaches in the context of soft skills classification. Figure 7 provides a visual representation of the system's workflow and components, highlighting the interaction between user input, model inference, and response providing.

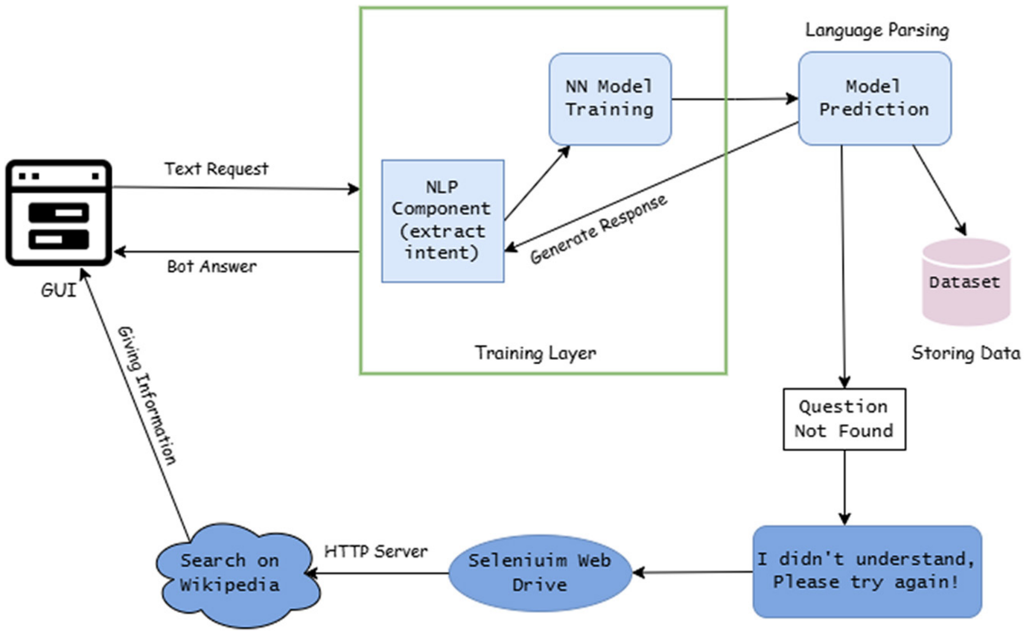

Fig. 7. The system architecture

ReLU-enhanced LSTM. The proposed model architecture extends the standard LSTM framework by integrating fully connected layers and ReLU activation functions, as illustrated in Figure 8. While the LSTM layer retains its core components—Input Gate, Forget Gate, Cell State, and Output Gate—ensuring effective modeling of sequential dependencies. The term ReLU-Enhanced refers to the inclusion of dense intermediate layers with ReLU activations, a widely adopted and non-novel practice in deep learning aimed at introducing non-linearity, improving learning dynamics, and mitigating vanishing gradient effects.

The model is implemented in PyTorch using a custom NeuralNet class, which defines key parameters such as input size = vocabulary size = 1247, hidden size = 128, and output size = 38 (number of classes). This architecture integrates one LSTM layer followed by two fully connected layers, each activated by ReLU function, the final layer uses Softmax activation for probability output over the 38 intents. The training process employs the Adam optimizer (lr = 10^{-3}) and Cross-Entropy loss. Figure 8 clearly provides a visual representation of the interaction between the LSTM and fully connected layers within the model.

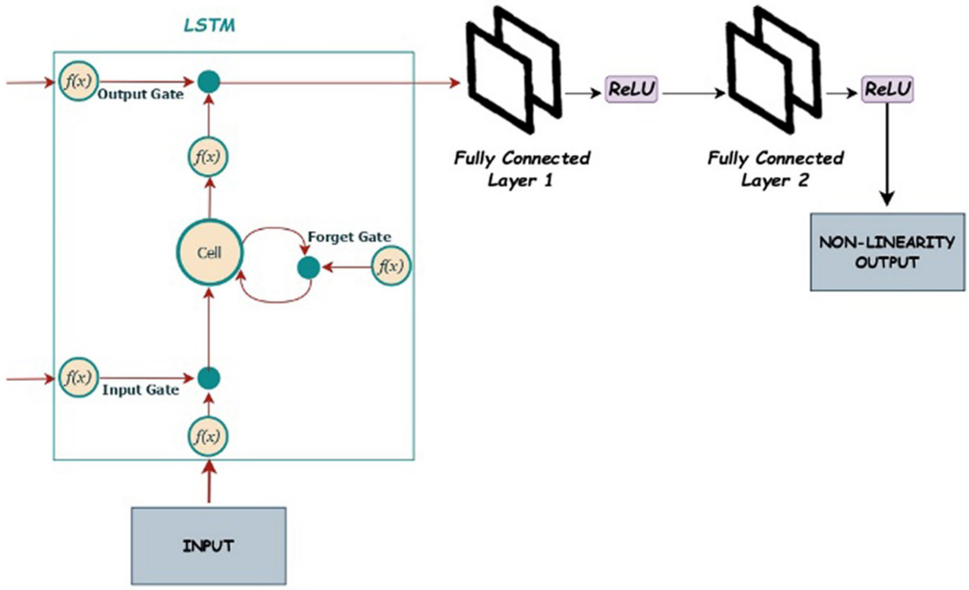

Fig. 8. LSTM neural network with ReLU activation

GloAT-Transformer. The second architecture GloAT-Transformer is based on the standard Transformer Encoder model [12], as illustrated in Figure 9. Its operation employs a multi-head self-attention mechanism that assigns different levels of importance to each token in the input sequence relative to all other tokens, enabling the model to capture complex contextual dependencies.

This architecture operates as follows: (1) input tokens are first mapped to 128-dimentional dense embedding; (2) these embedding are processed through two transformer encoder layers, each using 8-head multi-head self-attention which assigns different importance weights to all token pairs and captures complex contextual dependencies; (3) the resulting contextualized token representations are aggregated via global average pooling, which compresses the variable-length sequence into a fixed-size vector by averaging across all token positions-this "global" pooling is the key innovation giving the model its name (GloAT = Global Attention Transformer); (4) the pooled vector is passes through a linear layer and Softmax to produce a probability distributions over the 38 intent classes.

The key design advantages are simplicity, parameter efficiency, and robustness to variable query phrasing-properties particularly suited to our small, domain-specific dataset. Training uses the Adam optimizer (lr = 10^{-4}) and Cross-Entropy Loss.

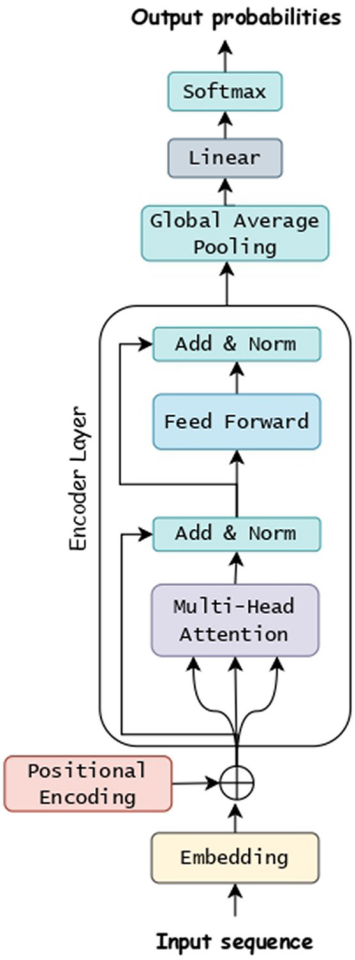

Fig. 9. Global Attention Transformer (GloAT) architecture

Training specifications and reproducibility. To ensure reproducibility of results, training hyperparameter are recorded in Table 1. The remaining hyperparameters were selected empirically to ensure a fair comparison between the ReLU-Enhanced LSTM and GloAT-Transformer models while maintaining training stability and computational efficiency.

Table 1. Training hyperparameters

Hyperparameter	ReLU-Enhanced LSTM Model	GloAT-Transformer Model
Vocabulary size	1247	1247
Embedding size	128	128
Batch size	32	32
Number of epochs	100	100
Learning rate	10^{-3} (Adam)	10^{-4} (Adam)
Attention heads	N/A	8
Encoder layers	N/A	2

4 RESULT AND DISCUSSION

This section presents the quantitative evaluation results and a detailed analysis of the proposed ReLU-Enhanced LSTM and GloAT-Transformer models for intent classification, their performance, robustness and limitations. The evaluation focuses on key metrics such as precision, recall, F1-score, and accuracy. Additionally, we discuss the impact of varying hyperparameters on model performance and provide insights into the trade-offs between model complexity, training time, and generalization of the obtained.

4.1 System implementation overview

Experimental Setup The chatbot system was implemented using PyTorch, an open-source, high-level Python library. The following steps outline the process of creating and training the models:

a) Neural network class definition;
- For the LSTM-based model; we created a class named NeuralNet with parameters such as input_size, hidden_size, and output_size. The model consists of two hidden layers with the ReLU activation function, which helps train the NN and reduce model error. Each linear layer determines the inputs and outputs for the model.
- For the Transformer-based model; we implemented a Transformer class that includes multi-head attention mechanisms and a global average pooling layer. The model leverages the MultiHeadAttention class to capture long-range dependencies in the input data, making it more effective for complex sequential tasks.

b) Optimizer and loss function;
- For both models; we used the Adam optimizer to minimize the error and improve model performance. For loss calculation, we employed the Cross-Entropy Loss function, which is optimized during training. The primary objective is to minimize this loss function, as lower loss indicates a better-performing model.

c) Training and saving the model;
- For both models; were trained and saved as a "data.pth" file. Once trained, the models can be used to make predictions. To test the chatbot's performance, we provided inputs such as queries about soft skills advices. For the LSTM-based model, we inherited a class from torch.nn.Module to define the architecture. For the Transformer-based model, we implemented additional components such as multi-head attention to enhance its ability to process sequential data.

d) User interaction workflow;
- The user sends a query to the system, asking the bot for specific information. This input message serves as the basis for the chatbot's response.
- The system filters the class/tag for the user input based on a probability threshold of 0.75. It creates a list of all filtered tags and sorts them in reverse order, ensuring that the class/tag with the highest probability appears first. Finally, the function returns the list, including classes and their corresponding probability values.
- The system identifies the class/tag with the highest probability, as returned by the previous function. It then compares this tag to the ones defined in the

"intents.json" file and selects a response associated with the identified class/tag at random.

- The bot's response is displayed on the UI bot interface. Thanks to the Q&A system implemented using the Flask framework.

4.2 Interpretation of model performance

The chatbot uses two advanced models: ReLU-Enhanced LSTM and GloAT-Transformer, both trained in PyTorch. These models replace the previously used Feedforward NN, offering improved performance in NLP tasks.

When a user submits a query, the chatbot processes the input using NLP techniques and retrieves the most relevant response. The ReLU-Enhanced LSTM excels in sequential data modeling, while the GloAT-Transformer leverages multi-attention mechanisms to capture long range dependencies, making it highly effective for complex queries. The system evaluates the confidence level of the predicted response, ensuring that only answers with a confidence threshold above 0.75 are returned to the user.

The chatbot's architecture is built using the following frameworks and libraries:

- **PyTorch**; for implementing and training the ReLU Enhanced LSTM and GloAT-Transformer models.
- **NumPy**; for numerical computations and data manipulation
- **Flask**; for handling user interactions.
- **Scikit-learn**; for additional data preprocessing and evaluation metrics.

Figures 10 and 11 are screenshots of the chatbot in action, demonstrating its user interface and interaction capabilities using both the ReLU-Enhanced LSTM and GloAT-Transformer models.

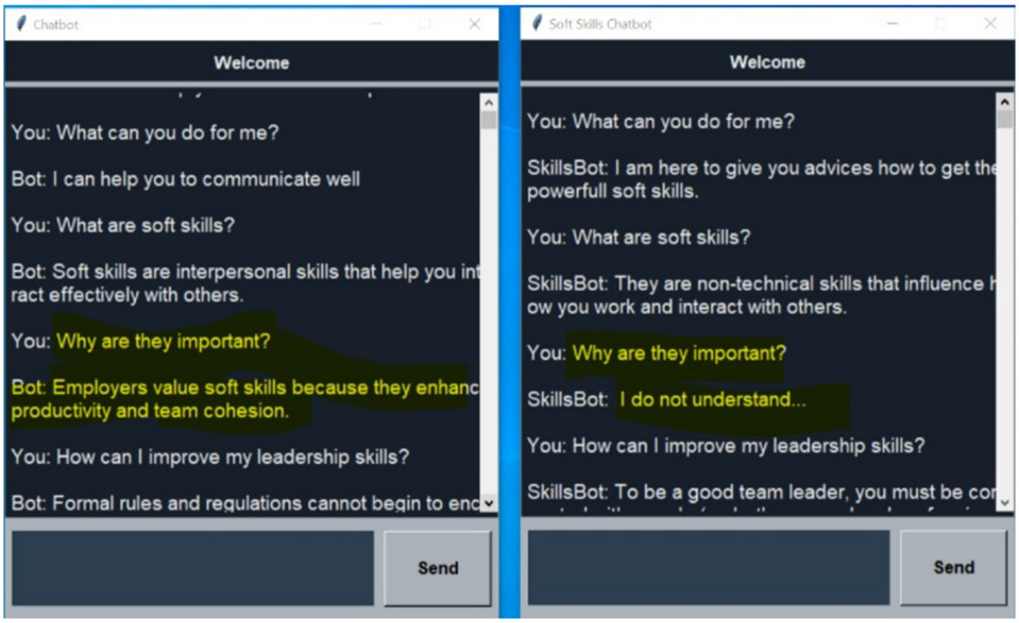

Fig. 10. Interaction with a Soft Skills Chatbot

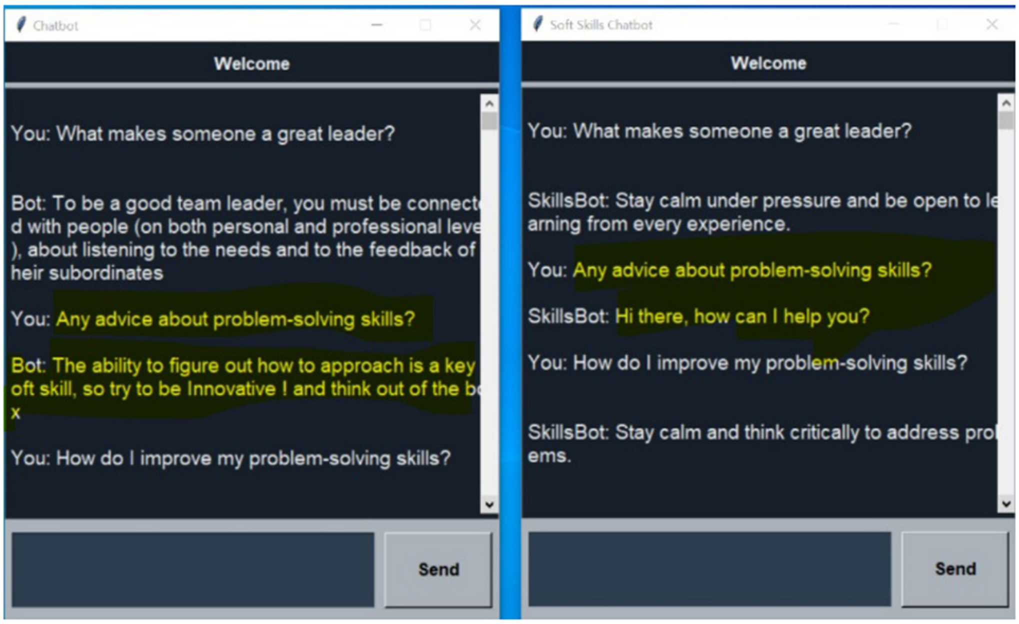

Fig. 11. Leadership and problem-solving advice

4.3 Evaluation and performance metrics

The performance of the two architectures was evaluated on the test set using multi-class classification metrics; Accuracy, Precision, Recall, and F1-score, which are widely used in NLP tasks to assess both classification correctness and class-level balance. The quantitative results are summarized in Table 2 illustrating the key differences in performance between the two architectures.

- GloAT-Transformer Model Evaluation; the Transformer model achieves an accuracy of 0.9776, a precision of 0.9848, a recall of 0.9776, and an F1 score of 0.9777.
- ReLU-Enhanced LSTM Model Evaluation; the LSTM model achieves an accuracy of 0.9731, a precision of 0.9690, a recall of 0.9731, and an F1 score of 0.9693.

Table 2. Comparison between ReLU-Enhanced LSTM and GloAT-Transformer Models

Metric	ReLU-Enhanced LSTM Model	GloAT-Transformer Model
Accuracy	0.9731	0.9776
Precision	0.9690	0.9848
Recall	0.9731	0.9776
F1 Score	0.9693	0.9777

The results indicate that the GloAT-Transformer model outperforms the ReLU-Enhanced LSTM model across all metrics, demonstrating its superior ability to handle complex language tasks. These GloAT-Transformer's higher scores reflect better contextual understanding and slight enhancement in balanced classification performance.

4.4 Confusion matrix analysis

To further analyze classification behavior, confusion matrices were examined for both models (see Figures 12 and 13), the GloAT-Transformer and the ReLU-Enhanced LSTM models and they illustrate their performance in intent classification.

The GloAT-Transformer model exhibits a stronger diagonal pattern, indicating a high rate of correct predictions with minimal misclassifications. This suggests that the GloAT-Transformer effectively captures contextual dependencies and enhances classification accuracy. The ReLU-Enhanced LSTM model also shows a well-defined diagonal structure but with slightly more off-diagonal entries, suggesting occasional misclassifications between closely related intents. Specifically, the off-diagonal entries in Figures 12 and 13 highlight the challenge of distinguishing between semantically similar soft skills (e.g., ‘Leadership’ vs ‘Teamwork’), a common issue in intent classification on domain-specific datasets.

A detailed confusion matrix analysis is essential to identify systematically misclassified intents and determine whether errors arise from semantic ambiguity, class overlap, or dataset imbalance. Such analysis provides deeper insight into model weaknesses beyond aggregate performance metrics.

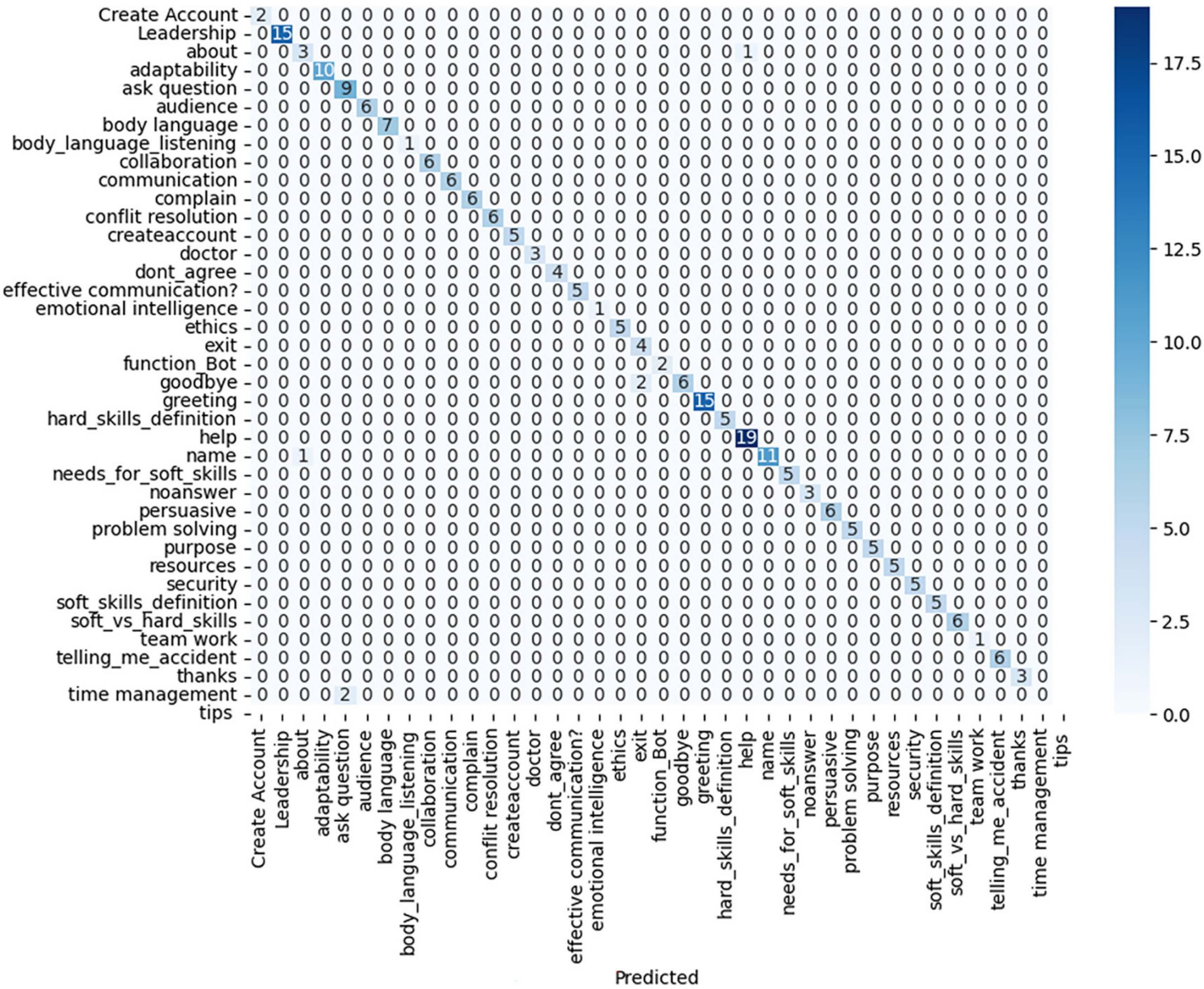

Fig. 12. Confusion matrix of ReLU-enhanced LSTM

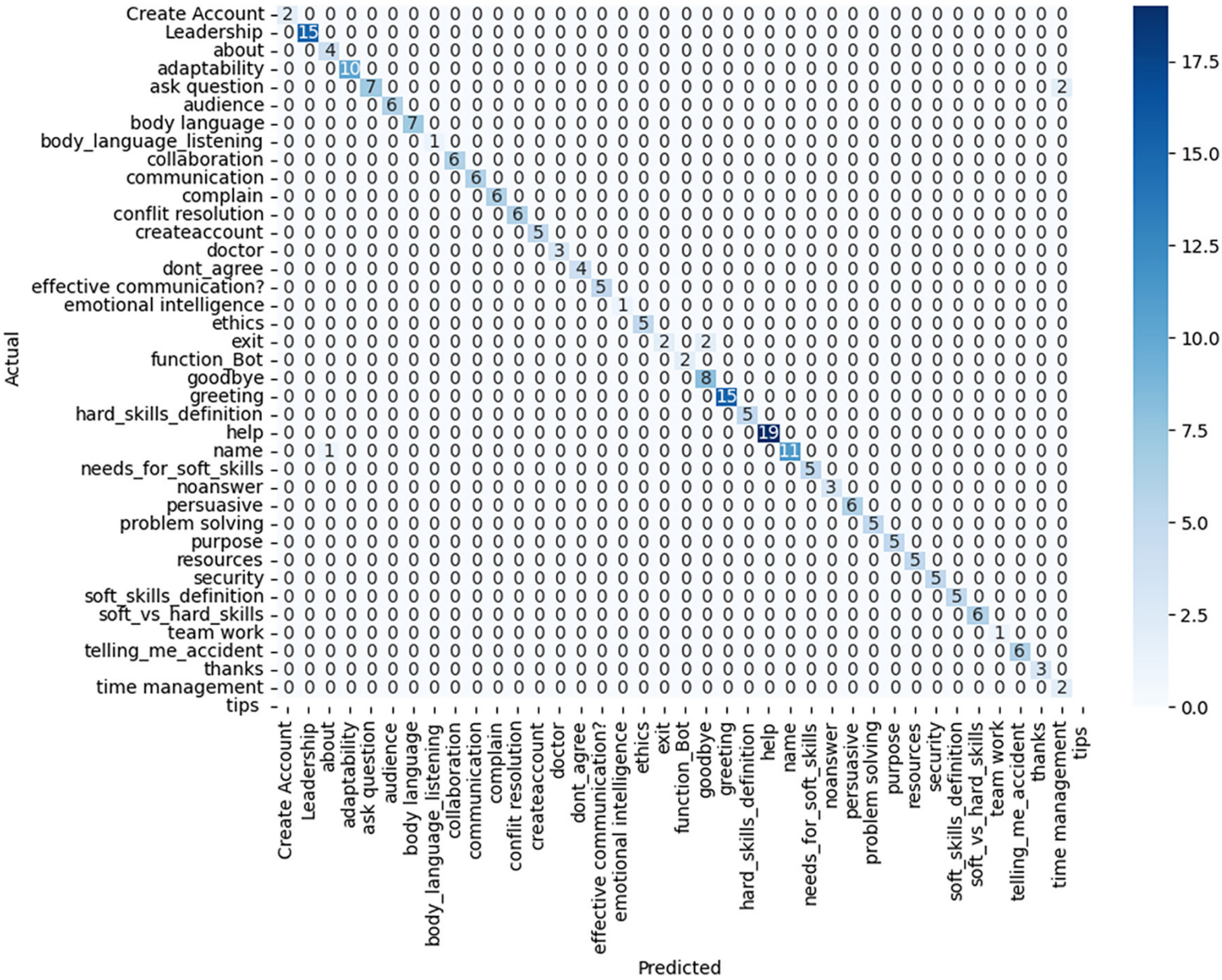

Fig. 13. Confusion matrix of GloAT-Transformer

4.5 Hyperparameter sensitivity and robustness

Experimental observations indicate that the GloAT-Transformer demonstrates greater robustness to hyperparameter variations compared to the ReLU-Enhanced LSTM as illustrated in Figure 14. The ReLU-Enhanced LSTM model is simpler to train, shows more variability in performance, particularly with suboptimal hyperparameter configurations. While the transformer maintains stable performance across tested settings, showing up its architectural resilience. This visual evidence reinforces the conclusion that the GloAT-Transformer architecture is not only marginally more accurate but also significantly more reliable for deployment in a real-world educational environment where optimal tuning may be challenging.

This robustness comes at the cost of higher computational complexity and increased inference time, which may impact deployment decisions in resource-constrained environments.

4.6 Discussion and limitations

The results underscore the advancements achieved by integrating Transformer-based and enhanced LSTM approaches in providing users with predefined responses based on recognized intents for soft skills informational support.

The Transformer model's superior performance in capturing long-range dependencies and contextual nuances makes it particularly effective for chatbot applications. On the other hand, the ReLU-Enhanced LSTM architecture represents a significant advancement in sequential data processing, particularly for educational applications. By combining the strengths of LSTM—its ability to model long-term dependencies—with the flexibility of fully connected layers and ReLU activation functions, the model achieves a balance between complexity and efficiency.

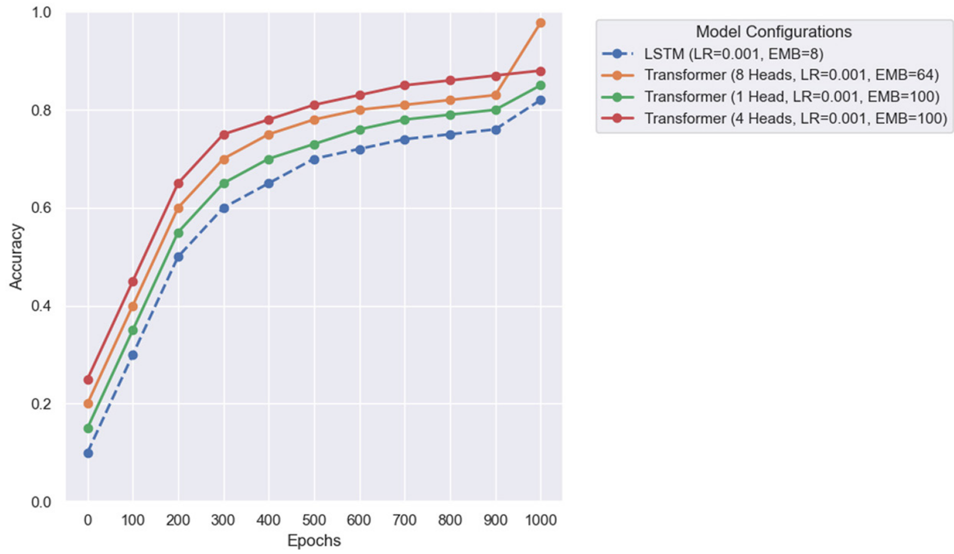

Fig. 14. GloAT-Transformer vs. ReLU-Enhanced LSTM performance comparison with different hyperparameters setting

Despite these advantages, several limitations must be acknowledged, regarding generalization capability, the high performance achieved is likely dataset-specific, and scalability to larger, more diverse corpora remains unverified. The current system is primarily focused on intent recognition and predefined response selection, which constrains, its ability to support more dynamic, context-aware, and personalized interactions. In terms of computational cost, although slightly more accurate, the transformer model introduces higher computational overhead compared to the LSTM model, which is critical factor for real-time or large-scale deployment.

However, these challenges remain areas for further research. Future work could focus on refining the model's efficiency, expanding the dataset for improved generalization, and integrating external knowledge sources to enhance chatbot responses.

Figure 14 compares the performance of both models across different hyperparameter settings. The GloAT Transformer model consistently outperforms the ReLU Enhanced LSTM model in all configurations, achieving higher accuracy and better contextual understanding. This figure highlights the robustness of the Transformer-based architecture, which is less sensitive to hyperparameter variations compared to the LSTM-based model, while simpler to train, shows more variability in performance, particularly with suboptimal hyperparameter settings.

5 CONCLUSION

This study highlights the potential of AI-powered chatbots to enhance student engagement and soft skills development through the implementation and evaluation of two deep learning models: the ReLU-Enhanced LSTM and the GloAT-Transformer. The results revealed that the Transformer-based model achieves marginally superior performance compared to the LSTM-based approach, at the level of accuracy, recall, and F1-score. That demonstrate its ability to capture long-range dependencies and contextual nuances in conversations. This underscores the effectiveness of Transformer chatbots in delivering interactive and personalized learning experiences.

The ReLU-enhanced LSTM architecture also represents a significant advancement in sequential data processing, particularly for educational applications. By combining sequential modeling with ReLU-based non-linearity, the model achieves a balance between complexity and efficiency, making it suitable for real-time educational applications.

This architecture is promising for Moroccan universities, especially for AI-driven chatbots focused on soft skills. Such chatbots can provide personalized feedback to students, helping them improve essential skills such as communication, teamwork, and problem-solving.

The model's ability to understand context and provide nuanced responses makes it an ideal tool for fostering student growth in both academic and professional settings. But a substantial augmentation and diversification of the dataset is required to enhance model generalization and mitigate overfitting. Then the system will evolve from intent classification toward a more comprehensive pedagogical chatbot, by leveraging Transformer-based text generation to enable richer, more context-aware interactions and scenario-based learning. Additionally, we plan to deploy this chatbot in an e-learning platform for computer science students, providing access to educational content and personalized recommendations. We also aim to extend the system, including voice-based interactions using speech recognition APIs and a recommendation system that suggests modules and advice based on user profiles.

Overall, this study contributes to the development of intelligent educational assistants that support students in acquiring and strengthening soft skills, preparing them for future academic and professional challenges. By integrating advanced AI technologies, we aim to create a more engaging and effective learning environment, ensuring that students have access to the resources they need to succeed.

6 ACKNOWLEDGMENT

This work was supported by the National Center for Scientific and Technical Research (CNRST) through the "PhD Associate Scholarship–PASS" program.

7 REFERENCES

[1] M.-A. Alt, I. Vizeli, and Z. Săplăcan, "Banking with a chatbot–A study on technology acceptance," *Studia Universitatis Babes Bolyai Oeconomica*, vol. 66, no. 1, pp. 13–35, 2021. https://doi.org/10.2478/subboec-2021-0002

[2] R. Jackson, "Understanding (and using) ChatGPT in banking," *American Bankers Association, ABA Banking Journal*, vol. 115, no. 3, pp. 16–17, 2023.

[3] M. R. King and ChatGPT, "A conversation on artificial intelligence, chatbots, and plagiarism in higher education," *Cellular and Molecular Bioengineering*, vol. 16, pp. 1–2, 2023. https://doi.org/10.1007/s12195-022-00754-8

[4] S. Kim *et al.*, "Why and what to teach: AI curriculum for elementary school," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, 2021, pp. 15569–15576. https://doi.org/10.1609/aaai.v35i17.17833

[5] A. Sabuncuoglu, "Designing one year curriculum to teach artificial intelligence for middle school," in *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education*, 2020, pp. 96–102. https://doi.org/10.1145/3341525.3387364

[6] T. Fu, S. Gao, X. Zhao, J.-R. Wen, and R. Yan, "Learning towards conversational AI: A survey," *AI Open,* vol. 3, pp. 14–28, 2022. https://doi.org/10.1016/j.aiopen.2022.02.001

[7] C. Tao, J. Feng, R. Yan, W. Wu, and D. Jiang, "A survey on response selection for retrieval-based dialogues," in *Proceedings of IJCAI*, 2021, pp. 4619–4626. https://doi.org/10.24963/ijcai.2021/627

[8] A. D. Tran, J. I. Pallant, and L. W. Johnson, "Exploring the impact of chatbots on consumer sentiment and expectations in retail," *Journal of Retailing and Consumer Services*, vol. 63, p. 102718, 2021. https://doi.org/10.1016/j.jretconser.2021.102718

[9] A. Miklosik, N. Evans, and A. M. A. Qureshi, "The use of chatbots in digital business transformation: A systematic literature review," *IEEE Access,* vol. 9, pp. 106530–106539, 2021. https://doi.org/10.1109/ACCESS.2021.3100885

[10] C. W. Okonkwo and A. Ade-Ibijola, "Chatbots applications in education: A systematic review," *Computers and Education: Artificial Intelligence,* vol. 2, p. 100033, 2021. https://doi.org/10.1016/j.caeai.2021.100033

[11] E. Mogaji, J. Balakrishnan, A. C. Nwoba, and N. P. Nguyen, "Emerging-market consumers' interactions with banking chatbots," *Telematics and Informatics,* vol. 65, p. 101711, 2021. https://doi.org/10.1016/j.tele.2021.101711

[12] S. Ayanouz, B. A. Abdelhakim, and M. Benhmed, "A smart chatbot architecture based NLP and machine learning for health care assistance," in *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, 2020, pp. 1–6. https://doi.org/10.1145/3386723.3387897

[13] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Machine Learning with Applications,* vol. 2, p. 100006, 2020. https://doi.org/10.1016/j.mlwa.2020.100006

[14] Z. Peng and X. Ma, "A survey on construction and enhancement methods in service chatbots design," *CCF Transactions on Pervasive Computing and Interaction,* vol. 1, no. 3, pp. 204–223, 2019. https://doi.org/10.1007/s42486-019-00012-3

[15] R. Yan, Y. Song, and H. Wu, "Learning to respond with deep neural networks for retrieval-based human-computer conversation system," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2016, pp. 55–64. https://doi.org/10.1145/2911451.2911542

[16] X. Zhou *et al.*, "Multi-turn response selection for chatbots with deep attention matching network," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,* vol. 1 (Long Papers), 2018, pp. 1118–1127. https://doi.org/10.18653/v1/P18-1103

[17] C. Shu *et al.*, "Open domain response generation guided by retrieved conversations," *IEEE Access,* vol. 11, pp. 99365–99375, 2023. https://doi.org/10.1109/ACCESS.2022.3225647

[18] L. Yu, W. Zhang, J. Wang, and Y. Yu, "SeqGAN: Sequence generative adversarial nets with policy gradient," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017. https://doi.org/10.1609/aaai.v31i1.10804

[19] Y.-L. Tuan and H.-Y. Lee, “Improving conditional sequence generative adversarial networks by stepwise evaluation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 788–798, 2019. https://doi.org/10.1109/TASLP.2019.2896437

[20] N. Esfandiari, K. Kiani, and R. Rastgoo, “A conditional generative chatbot using transformer model,” *arXiv preprint arXiv:2306.02074*, 2023.

[21] T. T. Quan, “N/A modern approaches in natural language processing,” *VNU Journal of Science: Computer Science and Communication Engineering*, vol. 39, no. 1, 2022. https://doi.org/10.25073/2588-1086/vnucsce.302

[22] E. Fathi and B. M. Shoja, “Deep neural networks for natural language processing,” in *Handbook of Statistics*, vol. 38, 2018, pp. 229–316. https://doi.org/10.1016/bs.host.2018.07.006

[23] Z. Lin, “Deep neural networks for natural language processing and its acceleration,” Montréal, QC, Canada, 2020.

[24] S. Hochreiter, “Long short-term memory,” *Neural Computation MIT-Press*, vol. 9, no. 8, pp. 1735–1789, 1997. https://doi.org/10.1162/neco.1997.9.8.1735

[25] M. Roondiwala, H. Patel, and S. Varma, “Predicting stock prices using LSTM,” *International Journal of Science and Research (IJSR)*, vol. 6, no. 4, pp. 1754–1756, 2017. https://doi.org/10.21275/ART20172755

[26] J. Cao, Z. Li, and J. Li, “Financial time series forecasting model based on CEEMDAN and LSTM,” *Physica A: Statistical Mechanics and its Applications*, vol. 519, pp. 127–139, 2019. https://doi.org/10.1016/j.physa.2018.11.061

[27] W. Bao, J. Yue, and Y. Rao, “A deep learning framework for financial time series using stacked autoencoders and long-short term memory,” *PLoS ONE*, vol. 12, no. 7, p. e0180944, 2017. https://doi.org/10.1371/journal.pone.0180944

[28] T. Fischer and C. Krauss, “Deep learning with long short-term memory networks for financial market predictions,” *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, 2017. https://doi.org/10.1016/j.ejor.2017.11.054

[29] S. Siami-Namini, N. Tavakoli, and A. S. Namin, “A comparative analysis of forecasting financial time series using ARIMA, LSTM, and BiLSTM,” *arXiv preprint arXiv: 1911.09512*, 2019.

[30] E. Kiperwasser and Y. Goldberg, “Simple and accurate dependency parsing using bidirectional LSTM feature representations,” *Trans Actions of the Association for Computational Linguistics*, vol. 4, pp. 313–327, 2016. https://doi.org/10.1162/tacl_a_00101

[31] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, Long Beach, CA, USA, vol. 30, 2017, pp. 5998–6008.

[32] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, “GLUE: A multi-task benchmark and analysis platform for natural language understanding,” in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium, 2018, pp. 353–355. https://doi.org/10.18653/v1/W18-5446

[33] T. Brown *et al.*, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, H. Larochelle et al., Eds., vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[34] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108*, 2020.

[35] S. Ahriz, H. Gharbaoui, N. Benmoussa, A. Chahid, and K. Mansouri, “Enhancing information technology governance in universities: A smart chatbot system based on information technology infrastructure library,” *Engineering Technology & Applied Science Research*, vol. 14, no. 6, pp. 17876–17882, 2024. https://doi.org/10.48084/etasr.8878

[36] I. Darmawan, H. Elmunsyah, and D. D. Prasetya, "ALBERTIR: A BERT-based pretraining for Indonesian religious texts using qur'an and hadith translations," *Engineering Technology & Applied Science Research*, vol. 15, no. 5, pp. 28307–28312, 2025. https://doi.org/10.48084/etasr.12977

[37] A. Raza, M. Latif, M. U. Farooq, M. A. Baig, and M. A. Akhtar, "Enabling context-based AI in chatbots for conveying personalized interdisciplinary knowledge to users," *Engineering Technology & Applied Science Research*, vol. 13, no. 6, pp. 12231–12236, 2023. https://doi.org/10.48084/etasr.6313

[38] Ministère de l'Enseignement Supérieur, de la Recherche Scientifique et de l'Innovation, "PACTE ESRI 2030: Plan National d'Accélération de la Transformation de l'Écosystème de l'Enseignement Supérieur, de la Recherche Scientifique et de l'Innovation," 2023.

8 AUTHORS

Chaimaa Bouafoud is a doctoral researcher and teaching assistant at the Faculty of Sciences, Chouaïb Doukkali University, El Jadida, Morocco. She is actively engaged in research at the intersection of artificial intelligence, educational technologies, and data science. Her work focuses on developing intelligent systems and adaptive learning models that enhance student engagement and performance in higher education. She has contributed to several research projects and academic events, including activities related to new university students. Her current research explores the integration of AI-driven adaptive learning platforms, such as chatbots and recommendation systems, within the Moroccan educational context (E-mail: bouafoud.chaimaa@ucd.ac.ma).

Pr. Abdellah Madani is a lecturer and researcher at the Department of Computer Science, Chouaïb Doukkali University, El Jadida, Morocco. He is the Director and an active member of the LAROSERI research laboratory. His research interests include artificial intelligence, machine learning, natural language processing, and cybersecurity, with a particular focus on intelligent systems and their real-world applications. He has authored numerous scientific publications and actively participates in academic and research activities (E-mail: madani.a@ucd.ac.ma).

Pr. Khalid Zine-Dine obtained his Ph.D. in 2000 from Mohammed V University. He is currently a Full Professor at the Faculty of Sciences, Mohammed V University, Rabat, Morocco. His research interests include system and network security, network architectures and communication protocols, wireless ad hoc and sensor networks, cloud computing, smart grids, and artificial intelligence applied to cybersecurity. He is actively involved in Ph.D. supervision and national and international research projects (E-mail: k.zinedine@um5r.ac.ma).