Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 12, No. 4, 2024, 465-472.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Development of a Persian Mobile Sales Chatbot based on LLMs and Transformer

Nura Esfandiari, Kourosh Kiani * and Razieh Rastgoo

Electrical and Computer Engineering Department, Semnan University, Semnan, Iran.

Article Info

Article History:
Received 12 September 2024
Revised 23 November 2024
Accepted 07 December 2024

DOI:10.22044/jadm.2024.15067.2609

Keywords:
Large Language Models(LLMs), Chatbot, Transformer, Persian Mobile, Sales Chatbot Persian Mobile, Sales Datasets.

**Corresponding author: Kourosh.kiani@semnan.ac.ir (K. Kiani).*

Abstract

Chatbots are computer programs designed to simulate human conversation. Powered by artificial intelligence (AI), these chatbots are increasingly used to provide customer service, particularly by large language models (LLMs). A process known as fine-tuning LLMs is employed to personalize chatbot answers. This process demands substantial high-quality data and computational resources. In this article, to overcome the computational hurdles associated with fine-tuning LLMs, innovative hybrid approach is proposed. This approach aims to enhance the answers generated by LLMs, specifically for Persian chatbots used in mobile customer services. A transformer-based evaluation model was developed to score generated answers and select the most appropriate answers. Additionally, a Persian language dataset tailored to the domain of mobile sales was collected to support the personalization of the Persian chatbot and the training of the evaluation model. This approach is expected to foster increased customer interaction and boost sales within the Persian mobile phone market. Experiments conducted on four different LLMs demonstrated the effectiveness of the proposed approach in generating more relevant and semantically accurate answers for users.

1. Introduction

In the realm of Artificial intelligence (AI) chatbots have emerged as powerful tools for simulating human conversation. These software applications are designed to provide automated online guidance and support, effectively mimicking natural human discourse [1]. Recently, chatbots are extensively employed in different areas, such as e-commerce and customer service [2]. Human customer service representatives have been the bridge between companies and their customers. Chatbot play a crucial role in building and nurturing these relationships. They can be a powerful tool for enhancing customer service by offering faster resolutions, increased efficiency, and improved customer experiences [2]. Recent advancements in AI and Natural language processing (NLP) techniques have made it easier and more flexible to implement more efficient chatbots. In recent years, chatbots have been developed using various approaches, including, Seq2Seq [3], Transformer [4], BERT [5], and LLMs based model [6]. Each of these approaches has advantages and disadvantages. The recent success of LLM-based chatbots like ChatGPT has spurred researchers to explore new and innovative applications for chatbot technology. The release of OpenAI's ChatGPT marked a significant advancement, pushing the boundaries of LLMs-based chatbots within the AI field [7]. LLMs have improved the world of chatbots, empowering them to interact with humans in ways far beyond what was previously imaginable. Conversational agents, more commonly known as AI chatbots, leverage vast amounts of data to train sophisticated language models. These LLMs then utilize this training to generate novel content (knowledge) in answers to user prompts [8].

While LLMs are trained on massive datasets and possess a broad general knowledge, they often require fine-tuning to excel at specific tasks. Traditionally, smaller models like BERT (450M parameters) or RoBERTa (1.3B parameters) were fully fine-tuned for specific tasks. However, the emergence of significantly larger models like GPT-3.5 has made full fine-tuning computationally prohibitive. Additionally, this process often requires a large amount of high-quality data, including conversation logs, customer queries, and support resolutions in customer service domain [9]. Gathering enough relevant data especially in some language like Persian and Arabic can be difficult, and the quality of the data directly impacts the effectiveness of the fine-tuned model. In addition, Fine-tuning LLMs can be computationally expensive, requiring significant processing power and storage. This can be a barrier for smaller businesses or those with limited resources.

In this article to overcome these challenges, new hybrid approach is proposed by combining the LLMs and transformer models. In this way, the transformer network is an evaluation module to assets LLMs output and select relevant and accurate answers. The intuition behind using the evaluation module is that customize chatbot in customer service domain. Additionally Persian dataset for mobile customer service is collected and prepared for train evaluation module. By developing the proposed approach by the collected data set, the chatbot achieves fluency and natural language understanding in Persian, enabling it to comprehend complex customer inquiries and offer personalized product recommendations. Key contributions from this research are:

Model: we develop a Persian mobile sales chatbot based on LLMs and transformer models. The transformer model assesses LLM-generated answers to select the optimal output. Our model significantly enhances the quality of LLM based chatbot interactions.

Dataset: A Persian dataset containing 23,000 customer service questions and answers for mobile phones was gathered, pre-processed, and prepared to train the evaluation module.

Performance: We conducted experiments using collected Persian dataset on different parameters of the model and obtained the best parameters for training the model. Also, the proposed approach was applied to the outputs of four LLM models with evaluation model and without it. Results indicate that our approach significantly enhances the semantic quality of LLM chatbot answers.

The rest of this article is organized in the following way: a review of relevant research is described in section 2. Section 3 presents our proposed approach in detail. Experimental results are outlined in Section 4, followed by a discussion of these findings in Section 5. Finally, Section 6 drawing conclusions and suggesting potential avenues for future work.

2. Related Works

This section offers a concise overview of recent advancements in AI chatbot technology, categorized into tree primary approaches: Seq2Seq-based, Transformer-based, LLM-based.

2.1. Seq2Seq-based approaches

This approach commonly employs Recurrent Neural Network (RNN) architectures like Long Short-Term Memory (LSTM) and A Gated Recurrent Unit (GRU) for chatbot development [3]. These models struggle to compress all essential information from long sentences and conversations (exceeding 20 words) into a fixed-length vector, a significant challenge in this domain [10]. To address this limitation, researchers have explored various strategies, including architectural modifications such as incorporating word embedding matrices [11], altering encoder or decoder components [12], and integrating attention mechanisms [13]. Rao et al. [14], introduced a hybrid model for text generation combining deep transfer learning with ELMo embeddings, a Variational Autoencoder (VAE), and Bi-directional Long Short-Term Memory (BiLSTM) networks. The model proposed in [15] employs a Bidirectional Recurrent Neural Network (BiRNN) equipped with an attention mechanism to enhance comprehension of input sentences.

2.2. Transformer-based approaches

Some research has explored the use of transformer-based models as an alternative to traditional recurrent layers in sequence-to-sequence architectures. These models employ a multi-head self-attention mechanism [10]. Santra et al. [16] Esfandiari et al. [17] introduced a general framework for hierarchical transformer encoders specifically designed for task-oriented dialogues. It is integrating the power of Reinforcement Learning (RL) and transformer models. One notable research effort proposed TILGAN, a Transformer-based GAN designed specifically for chatbot applications [18]. This model combines a transformer autoencoder with a generative adversarial network (GAN) in the latent area. TILGAN incorporates a new distribution matching method based on Kullback-Leibler divergence, which helps to improve the quality and naturalness of the generated text. Esfandiari et al. [19] introduced a method utilizing Conditional

Wasserstein Generative Adversarial Networks (CWGANs) with a transformer that operations data concurrently within training.

2.3. LLM-based approaches

The landscape of LLMs and chatbots is rapidly evolving, with new models and applications continually emerging. Some prominent examples include: ChatGPT [20], Developed by OpenAI, it's known for its versatility and ability to engage in informative and creative conversations. Bard, Google's response to ChatGPT, aiming for factual grounding and real-world integration. Llama [21], Meta's open-source LLM, has been used as the basis for various chatbot projects. Cohere, an open-source and powerful conversational AI platform developed by Cohere. It's designed to help businesses create engaging and informative chatbots that can interact with users in a natural and human-like way. In addition, researchers have explored the development of personalized chatbots using large language models. S. Yu et al. [22] proposed a chatbot approach by Bidirectional Encoder Representations from Transformers (BERT), an encoder architecture. Zhang et al. [23] introduced DIALOGPT, a transformer-based model pre-trained for generating bot answers. Uddin et al. [24] evaluates the usefulness of ChatGPT, a cutting-edge language model created by OpenAI, in the field of civil engineering education. Developing a chatbot to support victims of sexual harassment using a Large Language Model (LLM) as outlined in [25].

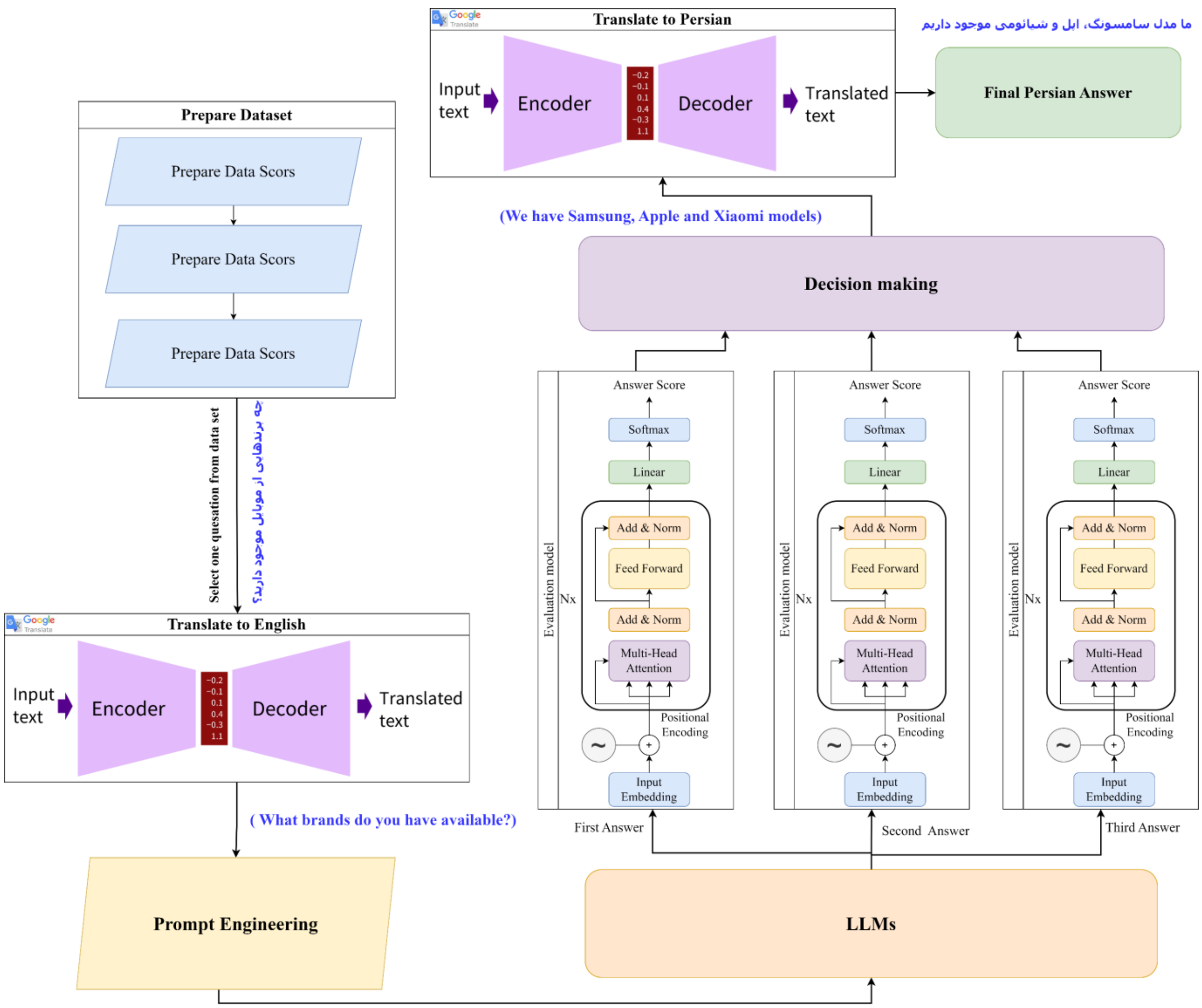

Figure 1. Main architecture of proposed approach.

3. Proposed Approach

In this section, we'll delve into the proposed approach, providing a detailed explanation of its key components and how they work together. We've developed a chatbot in Persian to serve as an intelligent assistant for customer service in mobile store. This chatbot leverages a combination of LLMs and a transformer model for evaluation. Figure 1 illustrates the main architecture of our proposed approach.

The proposed architecture consists of five primary modules: dataset preparation, translation, LLMs, evaluation model, and decision making. The evaluation model is trained on a custom dataset. In the following, we'll explore each module individually and clarify how they interact.

- **Dataset Preparation:**

It involves three main phases: data collection, data pre-processing, and data scoring. In dataset preparation phase a comprehensive dataset of 23,000 real Persian questions and answers related to mobile phone customer service was assembled. Each question was presented to the LLMs three to five times, and the resulting answers were stored in a database. Then to ensure the data is suitable for processing by the LLMs and evaluation models, a series of NLP techniques were applied.

Finally, each LLMs answer was assigned a score using the Bert score metric. The Bert score is a metric used to evaluate the quality of text generation models. It leverages the pre- trained BERT model to compare the generated text to a real text. The calculation of the Bert score is based on the following Equation (1):

$$BertScore = P \times R \times F \quad (1)$$

Where, P (precision) is the number of tokens in the real text that are also in the generated text, divided by the total number of tokens in the generated text. R (recall) is the number of tokens in the real text that are also in the generated text, divided by the total number of tokens in the real text. F (F1-score) is the harmonic mean of precision and recall. The calculated Bert scores were stored in the database for further analysis and to use in the decision-making process.

- **Translation:**

In this module, we leverage the Google Translate API to facilitate language translation. Persian questions are initially translated into English at the beginning of the architecture. Subsequently, the English generated answers are translated back into Persian before being presented to the user.

- **LLMs:**

LLMs are employed to generate comprehensive and contextually relevant answers to the input questions. Each question, along with its corresponding prompt, is given to the LLMs so that between 3 and 5 answers are generated by the LLMs for that question. The prompt specifies the desired length and quantity of the generated answers.

- **Evaluation model:**

The evaluation model assesses the semantic quality of the LLM-generated answers and assigns a score to each answer. This model is comprised of a transformer architecture equipped with an encoder only. It is trained on a custom dataset to effectively evaluate the semantic coherence and relevance of the generated answers. The primary objective of the evaluation model is to learn how to score the answers generated by LLMs in order to select the most appropriate answer in terms of meaning. Each of the answers generated by LLMs (the first, second and third answers) is given to the evaluation model separately and a score is given to each answer by the evaluation model.

- **Decision making:**

The decision module carefully selects the most appropriate answer based on the scores assigned by the evaluation model. The answer with the highest score is deemed the best candidate as follow:

$$Score_{ans} = Max(score_{ans1}, score_{ans2}, score_{ans3}) \quad (2)$$

The best answer is then translated into Persian by the translation module before being presented to the user as the final output.

4. Results

The purpose of this section is to present the details of the proposed approach and the results obtained. We begin by discussing implementation details, followed by a description of the dataset used for training the evaluation model. Next, an ablation study is conducted to evaluate the effectiveness of different components of our method. We then compare our approach to existing methods. Finally, we provide qualitative examples to illustrate the performance of our approach.

4.1. Implementation details

The experiments were conducted on a Microsoft Windows 11 system equipped with an NVIDIA

GeForce RTX 3090 GPU, an Intel Core i5-12600K CPU, and 128GB of RAM. The evaluation model was implemented using PyTorch. Table 1 provides a list of the implementation parameters used in this study.

Table 1. list of the implementation parameters employed in the proposed approach.

Parameters	value	Parameters	value
Learning rate	0.0001	Layers number of Transformer	6
Batch size	32	Dataset split ration	20%
Epoch numbers	100	Max Length of Sentences	30
Processing unit	GPU	Head number of Transformer	8
Dropout	0.5	Number of LLMs answer	3-5

4.2. Datasets

This article aims to evaluate the effectiveness of our proposed approach by implementing it as a Persian chatbot for mobile store customer service. To accomplish this, a dataset of 23,000 real questions and answers from Persian- mobile store customers was gathered. These inquiries covered 25 topics for three brands: Apple, Samsung and Xiaomi that pair with appropriate answers. This dataset will be available upon request after the publication of the article. Some examples of this dataset are shown in Table 2, which provides a representative sample of the data used in this article.

Table 2. Example of raw dataset in Persian.

Question	آیا می‌توانم با شما چت کنم؟
Answer	بله ، من آماده پاسخگویی به سوالات شما هستم
Question	روش های پرداخت در ثبت سفارش چگونه است؟
Answer	صورت آنلاین و یا حضوری و یا درب منزل به
Question	آیا میشه موجودی هر مدل موبایل رو ببینم
Answer	آره قبل از آن ابتدا مدل هر گوشی را انتخاب کرده و سپس موجودی آن مدل را چک کنید
Question	امسال چه برند های جدیدی وارد میکنید؟
Answer	برندهای شیائومی،سامسونگ،اپل وارد میکنیم
Question	چگونه می توان رم گوشی را بررسی کرد؟
Answer	با رفتن به تنظیمات یا خوندن اطلاعات بر روی جعبه گوشی

All questions and answers were then translated into English using Google Translate. Next, each question was presented to the LLMs 3-5 times, and the answers were recorded. To guide the LLMs output, the following prompts were used: 1) limit answers to 30 tokens, 2) keep answers within the realm of mobile sales, and 3) generate 3-5 answers. The resulting dataset was then preprocessed to remove unnecessary punctuation and standardize the text. Finally, the Bert score for each answer generated by the LLMs was calculated based on equation 1 and stored in the database. Table 3 provides an example of this dataset after preprocessed and translation to English language.

Table 3. Example of preprocessed dataset in English.

Question	do you have an existing xiaomi brand
Real Answer	yes, we have an existing xiaomi brand
Answer 1	yes, i do have a few xiaomi models available
Bert Score1	0.68
Answer 2	i can offer you a great deal on the xiaomi 12
Bert Score2	0.66
Answer 3	yes, we have the latest xiaomi releases
Bert Score3	0.77

4.3. Ablation analysis

In this section, we investigate how different parameters affect transformer training performance when evaluated on the collected dataset. To isolate these effects, we've divided our analysis into three categories:

- Category A: Explores the influence of the number of layers while maintaining a fixed head size and learning rate.
- Category B: Examines the impact of head size while keeping the number of layers and learning rate constant.
- Category C: Studies the effects of varying the learning rate while keeping the number of layers and head size fixed.

To measure model performance, we've employed the maximum likelihood estimation (MLE) loss function.

Table 4. Investigating the impact of transformer parameters

Evaluation Model	Type	Layer	LR	Head	MLE Loss
Transformer model	A	2	0.0001	8	0.00336
		4	0.0001	8	0.00336
		6	0.0001	8	**0.00270**
		8	0.0001	8	0.00358
		10	0.0001	8	0.00348
		12	0.0001	8	0.00343
		14	0.0001	8	0.00404
	B	6	0.0001	2	0.00344
		6	0.0001	4	0.00308
		6	0.0001	**8**	**0.00270**
		6	0.0001	16	0.00330
	C	6	0.1	8	0.01955
		6	0.01	8	0.01944
		6	0.001	8	0.01965
		6	**0.0001**	8	**0.00270**
		6	0.00001	8	0.00391

Based on the results in Table 4, we found that the optimal combination of hyperparameters for the transformer model was a head size of 8, 16 layers, and a learning rate of 0.0001.

4.4. Experimental results

Proposed approach aims to enhance the semantic coherence of LLMs-generated answers by incorporating an evaluation model. We compared the performance of our approach against several popular LLMs, including GPT-3.5, GPT-4, Gemini and cohere in the case of using and not using the evaluation model. The BERT score was used to measure the semantic similarity between the generated and real answers. As demonstrated in Table 5, incorporating our proposed evaluation model significantly enhanced the semantic relevance of answers generated by LLMs.

Table 5. Comparison between proposed approach and LLMs

LLM Models	Evaluation Model	Mean Bert Score
GPT-3.5	Yes	0.72
GPT-3.5	No	0.68
GPT-4	Yes	**0.79**
GPT-4	No	0.77
Gemini	Yes	0.77
Gemini	No	0.75
cohere	Yes	0.74
cohere	No	0.71

Table 6 presents a qualitative analysis of our proposed approach's outputs, comparing them to real answer and those generated by LLMs with and without the evaluation model. Table 7 provides a comparison of the answers generated by our proposed approach to the ground truth answers for various questions.

Table 6. Qualitative analysis of proposed approach.

Question	can i chat with you?
Real Answer	yes, i am ready to answer your questions about phon.
GPT-3.5 Answer	How can I assist you today?
GPT-3.5+Evaluation model Answer	yes, what features are you looking for in a phone.
GPT-4 Answer	Of course! How can I help you today?
GPT-4+Evaluation model Answer	hello there how can i help you about phone.
Gemini	Absolutely! Feel free to ask me anything.
Gemini+Evaluation model Answer	Absolutely! I'm ready to chat about phones.
cohere Answer	what can i assist with you
cohere+Evaluation model Answer	of course, are you in the market for a new phone

Table 7. Qualitative analysis of generated answers by proposed approach .

Question	Ground Truth Answer	Generated Answer
شما همیشه فعال هستید؟	بله ، من شبانه روز در حال پاسخگویی هستم، سوال هست بفرمایید	بله من اینجا هستم تا به هر گونه سوالی در ارتباط با تلفن همراه پاسخ دهم
Are you always active?	Yes, I am responding day and night, if you have any questions, please ask	Yes, I'm here to answer any cell phone related questions
تو هر زمینه ای میتونی کمک کنی؟	بلی آماده خدمات رسانی در زمینه مرتبط با تلفن همراه هستم	می‌توانم در مورد هرگونه سؤال مرتبط با تلفن همراه کمک کنم
Can you help in any field?	Yes, I am ready for services related to mobile phones	I can help with mobile phone
به صورت حضوری چگونه میتوان پرداخت کرد؟	به صورت نقدی و یا کارت عابر بانک	برای پرداخت های حضوری می توانیم پول نقد یا کارت عابر بانک بگیریم
How can you pay in person?	In cash or with a bank card	For person payments, we can take cash or bank card
زمان تحویل سفارش من چقدر است؟	بستگی به موقعیت مکانی شما و موجودی انبار دارد	بستگی به موقعیت مکانی شما دارد اما معمولاً ۲۷ روز کاری است
How long does it take to deliver my order?	It depends on your location and stock availability	It depends on your location but usually 2-7 working days
آیا میتوان لیست سفارش خود را افزایش بدهم؟	بله می توان افزایش داد	بله ما می توانیم سفارش شما را افزایش دهیم
Can I increase my order list?	Yes, it can be increased	Yes, we can increase your order
برای ارتقای سیستم عامل گوشی چیکار کنم	شما باید آن را به روز کنید	شما می توانید با انتخاب یک نسخه جدید آن را بروز کنید
What should I do to upgrade the phone's operating system?	You need to update it	You can update it by selecting a new version
تفاوت بین حافظه داخلی و حافظه رم چیست؟	حافظه داخلی برای ذخیره سازی دائمی داده ها استفاده می شود در حالی که حافظه رم به عنوان حافظه موقت استفاده می شود	حافظه داخلی داده ها را برای مدت طولانی ذخیره می کند در حالی که رم برای ذخیره سازی موقت داده ها است
What is the difference between internal memory and RAM memory?	Internal memory is used for permanent data storage while RAM memory is used as temporary memory	Internal memory stores data for a long time while RAM is for temporary data storage
چه رنگ هایی از تلفن همراه محبوب هستند	اول از همه سیاه، سپس سفید و خاکستری رنگ های محبوب هستند	سیاه سفید و خاکستری همیشه انتخاب های محبوبی هستند
What colors of mobile phones are popular?	First of all, black, then white and gray are popular colors	Black, white and gray are always popular choices
آیا حسگر اثر انگشت در همه گوشی ها وجود دارد	نه همه گوشی ها حسگر اثر انگشت ندارند، این ویژگی در برخی از گوشی های هوشمند با مدل های بالاتر موجود است	نه همه گوشی ها حسگر اثر انگشت ندارند، این معمولاً یکی از ویژگی های مدل های بالاتر است
Do all phones have a fingerprint sensor?	Not all phones have a fingerprint sensor, this feature is present in some high-end smartphones	Not all phones have a fingerprint sensor, this is usually a feature of higher models

5. Discussion

Recent LLMs-based chatbots often struggle to maintain semantic relevance in personalized interactions. Fine-tuning, while a common solution, demands substantial data and computational resources. To address these challenges, we propose a transformer-based evaluation model that enhances the semantic alignment between LLM-generated answers and user queries. We collected a Persian dataset on mobile customer service to evaluate our approach. Our model leverages the transformer architecture to learn to score generated answers. Loss functions play a crucial role in deep learning model performance. Our analysis of the training and validation losses in Figure 2 reveals a stable convergence after 200 epochs, indicating effective model training on the mobile customer service dataset.

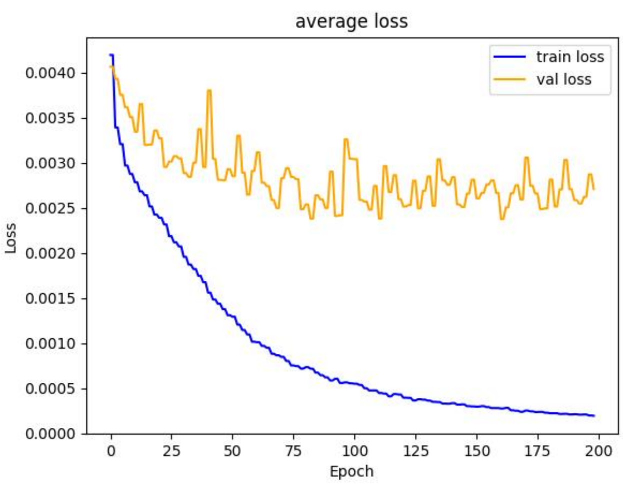

Figure 2. Evaluation model loss curve in mobile customer service dataset.

6. Conclusion and Future work

This study introduces a novel approach to enhance the performance of personalized Persian mobile sales chatbots. By leveraging LLMs and a transformer-based evaluation model, we aimed to improve the quality and relevance of chatbot answers. Proposed evaluation model, which employs the BERT scoring method, effectively assesses the semantic similarity between chatbot answers and user queries. To train the evaluation model in a personalized manner, we collected a Persian language dataset from mobile customer service interactions. Experimental results demonstrate that integrating this evaluation model significantly improves the ability of LLMs to generate accurate and contextually appropriate answers.

While proposed approach shows promising results, it is important to acknowledge its limitations. The current dataset is specific to mobile customer service, and future research could benefit from a more diverse dataset encompassing various domains. Additionally, leveraging a retrieval-augmented generation (RAG) framework as a future work, can improve the accuracy and relevance of chatbot answers by providing access to a knowledge base.

References

[1] G. Caldarini, S. Jaf, and K. McGarry, “A literature survey of recent advances in chatbots,” *Information*, vol. 13, pp. 41, 2022.

[2] M.F. Shahzad, et al., “Assessing the impact of AI-chatbot service quality on user e-brand loyalty through chatbot user trust, experience and electronic word of mouth,” *Journal of Retailing and Consumer Services*, vol. 79, pp. 103867, 2024.

[3] K. Palasundram, et al., “SEQ2SEQ++: A Multitasking-Based Seq2seq Model to Generate Meaningful and Relevant Answers,” *IEEE Access*, vol. 9, pp. 164949-164975, 2021.

[4] R. Zandie and M.H. Mahoor, “Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems,” *in the thirty-third international flairs conference*, 2020.

[5] S. Yu, Y. Chen, and H. Zaidi, “AVA: A Financial Service Chatbot Based on Deep Bidirectional Transformers,” *Frontiers in Applied Mathematics and Statistics*, vol. 7, pp. 604842, 2021.

[6] J.K. Kim, et al., “ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine,” *Journal of Pediatric Urology*, vol. 19(5), pp. 598-604, 2023.

[7] O. Caelen, M.A. Blete, “eveloping Apps with GPT-4 and ChatGPT,” *O'Reilly Media*, August 2023.

[8] C. Lin, A.Y.Q.Huang, S.J.H. Yang, “A Review of AI-Driven Conversational Chatbots Implementation Methodologies and Challenges (1999–2022),” *Sustainability*, vol. 15, 2023.

[9] C. Jeong, “Fine-tuning and utilization methods of domain-specific llms,” *arXiv preprint arXiv:2401.02981*, 2024.

[10] A. Vaswani, et al., “Attention is all you need,” *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000 - 6010, 2017.

[11] I.V. Serban, et al., “A hierarchical latent variable encoder-decoder model for generating dialogues,” *in*

Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 3295–3301, 2017.

[12] J. Gu, et al., "Incorporating Copying Mechanism in Sequence-to-Sequence Learning," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* vol. 1, pp. 1631–1640, 2016.

[13] M. Nuruzzaman, O.K. Hussain, "IntelliBot: A Dialogue-based chatbot for the insurance industry," *Knowledge-Based Systems,* vol. 196, pp. 105810, 2020.

[14] K.Y. RAO, K.S.RAO, "Modeling text generation with contextual feature representation and dimension using deep transfer learning and BI-LSTM," *Journal of Theoretical Applied Information Technology*, vol. 100(9), 2022.

[15] R. Keerthana, G. Fathima and L.Florence, "Evaluating the Performance of Various Deep Reinforcement Learning Algorithms for a Conversational Chatbot," *in 2nd International Conference for Emerging Technology,* pp. 1-8, 2021.

[16] B. Santra, P. Anusha, and P. Goyal, "Hierarchical Transformer for Task Oriented Dialog Systems," *in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 2021.

[17] N. Esfandiari, K. Kiani and R. Rastgoo, "Transformer-based Generative Chatbot Using Reinforcement Learning," *Journal of Artificial Intelligence & Data Mining (JAIDM)*, vol. 12, pp. 349-358, 2024.

[18] S. Diao, et al., "TILGAN: Transformer-based Implicit Latent GAN for Diverse and Coherent Text Generation," *Association for Computational Linguistics (ACL)*, 2021.

[19] N. Esfandiari, K. Kiani, and R. Rastgoo, "A Conditional Generative Chatbot using Transformer Model," preprint arXiv:02074, 2023.

[20] P.P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet of Things and Cyber-Physical Systems,* vol. 3, pp. 121-154, 2023.

[21] H. Touvron, et al., "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[22] S. Yu, Y. Chen, and H. Zaidi, "AVA: A Financial Service Chatbot Based on Deep Bidirectional Transformers," *Frontiers in Applied Mathematics and Statistics*, vol. 7, pp. 604842, 2021.

[23] Y Zhang, et al., "DIALOGPT: Large-Scale Generatie Pre-training for Conversational Response Generation," *In Annual Meeting of the Association for Computational Linguistics*, 2019.

[24] S.M.J. Uddin, et al., "ChatGPT as an educational resource for civil engineering students," *Computer Applications in Engineering Education*, vol. 32(4), pp. e22747, 2024.

[25] S. Vakayil, et al. "RAG-Based LLM Chatbot Using Llama-2," *in 2024 7th International Conference on Devices, Circuits and Systems (ICDCS),* 2024.

توسعه یک چت‌بات فروش موبایل فارسی مبتنی بر مدل‌های زبانی بزرگ (LLM) و ترانسفورمر

نورا اسفندیاری، کوروش کیانی* و راضیه راستگو

[۱]**دانشکده مهندسی برق و کامپیوتر، دانشگاه سمنان، سمنان، ایران.**

ارسال ۲۰۲۴/۰۹/۱۲؛ بازنگری ۲۰۲۴/۱۱/۲۳؛ پذیرش ۲۰۲۴/۱۲/۰۷

چکیده:

چت‌بات‌ها برنامه‌های کامپیوتری هستند که برای شبیه‌سازی گفتگوی انسانی طراحی شده‌اند. این چت‌بات‌ها که توسط هوش مصنوعی (AI) تقویت می‌شوند، به طور فزاینده‌ای برای ارائه خدمات مشتری، به ویژه توسط مدل‌های زبانی بزرگ (LLM) استفاده می‌شوند. فرآیندی به نام "تنظیم دقیق LLM" برای شخصی‌سازی پاسخ‌های چت‌بات به کار می‌رود. این فرآیند نیازمند داده‌های با کیفیت بالا و منابع محاسباتی قابل توجهی است. در این مقاله، برای غلبه بر موانع محاسباتی مرتبط با تنظیم دقیق LLMها، یک رویکرد ترکیبی نوآورانه پیشنهاد شده است. این رویکرد با هدف بهبود پاسخ‌های تولید شده توسط LLMها، به ویژه برای چت‌بات‌های فارسی مورد استفاده در خدمات مشتری موبایل، طراحی شده است. یک مدل ارزیابی مبتنی بر ترانسفورمر برای امتیازدهی به پاسخ‌های تولید شده و انتخاب مناسب‌ترین پاسخ‌ها توسعه داده شد. علاوه بر این، یک مجموعه داده زبان فارسی متناسب با حوزه فروش موبایل جمع‌آوری شد تا از شخصی‌سازی چت‌بات فارسی و آموزش مدل ارزیابی پشتیبانی کند. انتظار می‌رود این رویکرد تعامل بیشتر با مشتریان را تقویت کرده و فروش را در بازار تلفن همراه فارسی افزایش دهد. آزمایش‌هایی که بر روی چهار LLM مختلف انجام شد، اثربخشی رویکرد پیشنهادی را در تولید پاسخ‌های مرتبط‌تر و دقیق‌تر از نظر معنایی برای کاربران نشان داد.

کلمات کلیدی: مدل های زبانی بزرگ (LLM)، چت بات، ترنسفورمر، چت‌بات فروش موبایل فارسی، دیتاست فروش موبایل فارسی.