Available online at www.garmian.edu.krd

Passer Journal of Basic and Applied Sciences

Passer 7 (Issue 1) (2025) 506-512

Original Article

A Chatbot for Frequently Asked Questions using TF-IDF and Query Expansion Techniques

Mohammed Salah Ibrahim*[1], Anas Diab Sallibi[2], Ahmed Adnan Hadi[3], Ahmed Adil Nafea[1]

[1]*Department of Artificial Intelligence, College of Computer Science and IT, University of Anbar, Ramadi 31001, Iraq.*
[2]*Department of Computer Sciences, College of Sciences, University of Al Maarif, Ramadi 31001, Iraq.*
[3]*Artificial Intelligence Sciences Department, College of sciences, Al-Mustaqbal University, Babil 51001, Iraq.*

Received 7 Nowember 2024; revised 19 march 2025;
accepted 3 April 2025; available online 26 June 2025

DOI: 10.24271/PSR.2025.487515.1805

ABSTRACT

The Frequently Asked Questions (FAQs) chatbot plays an important role in providing information, especially in academic fields. FAQs are created to address common concerns frequently raised and answered by domain experts. Answers to such FAQs should be precise and related to the question asked. This paper presents a chatbot system that uses a Term Frequency-Inverse Document Frequency (TF-IDF) method and enhances this using a semantic query expansion technique (TF-IDF-Expa) to rank and retrieve accurate responses. Two question-answering datasets are utilized to evaluate the system, SQuAD and MS-Marco. Moreover, 50 synthetic academic questions are used to show the system's performance in real-time. Both methods, TF-IDF and TF-IDF-Expa, reported 100% accuracy; However, the confidence rate for TF-IDF-Expa was around 70% due to the additional information presented during query expansion, which can add ambiguity. The trade-off between accuracy and confidence raises concerns and suggests more optimisation chances for real-world applications.

Keywords: Generative AI applications, Frequently Asked Questions, Chatbots, Question-Answering, Query Expansion, Chatbot development.

1 Introduction

With the recent development of communication services, especially chatbots and other related conversational interfaces, the need for accurate and advanced chatbots has become paramount. Chatbots, commonly known as conversational agents, belong to the field of question-answering area of study. The main goal of building such agents is to create a system that can process user questions and provide accurate answers [1]. The chatbot market is growing rapidly. According to DemandSage, a well-known statistics organisation, the chatbot market could reach 10.3 billion dollars in 2025 [2]. Other statistics, like Cherniak (2024), showed that market revenue will jump from 106 million in 2022 to 454 million in 2027. Figure 1 shows how the revenue curves increase [3]. Frequently Asked Questions (FAQs) systems play a fundamental role in providing instantaneous and relevant information to users. As institutions and organisations encourage customer engagement through digital channels, the need for intelligent FAQ tools is increasing. The purpose is to provide timely customer service with less time and human effort.

* Corresponding author
E-mail address moh.salah@uoanbar.edu.iq (Instructor).
Peer-reviewed under the responsibility of the University of Garmian.

Chatbots come in various forms, from the most basic ones used to schedule appointments on Facebook Messenger to fully functional AI assistants built into gadgets (such as Siri or Alexa). These chatbots are not mere substitutes for FAQ pages. They can assist with staff orientation, offer customised products and services, and take and track orders, depending on their level of sophistication [4].

Chatbots are robust communication systems. They are helpful in fields such as e-commerce, business, education, and information retrieval [5]. With the development of large language models (LLMs), chatbots became more accurate with advances in how they communicate and provide information. The recent models are all based on the concept of a transformer and an attention mechanism with more than 100 million parameters [6]. LLM chatbots such as ChatGPT [7], Gemini [8], DeepSeek [9], and many other chatbots are artificial intelligence (AI) chatbots built based on LLM. They are very potent tools that can formulate human-like answers based on user queries [10].

In this research, large language models are not required to formulate chatbot answers because FAQ answers are already listed for any question asked. The proposed system processes the question and any other text that helps to find accurate and already

listed answers. Due to that, different Natural Language Processing (NLP) tools and algorithms are used with the help of query expansion to reach the best-matched answers. The main goal of this research is to design and implement a classic chatbot specifically designed for Frequently Asked Questions, leveraging advanced NLP techniques. This system should be able to answer user questions quickly without the need for a user to dig into all the FAQs to find the answer they are looking for. The aim is to integrate TF-IDF (Term Frequency-Inverse Document Frequency) ranking for precise information retrieval with query expansion tools to improve the chatbot's understanding of user queries. The TF-IDF technique is lightweight compared to LLMs, which require high computation resources, massive datasets, and millions to billions of parameters. Additionally, this technique provided a high accuracy, enough to retrieve FAQs. From these objectives, we aim to address the limitations of conventional FAQ systems and contribute to the advancement of conversational agents.

This paper is organized into several sections. The literature review examines existing research, while the methodology section details the chatbot's design and techniques. Subsequent sections present the system evaluation results, discuss, and conclude.

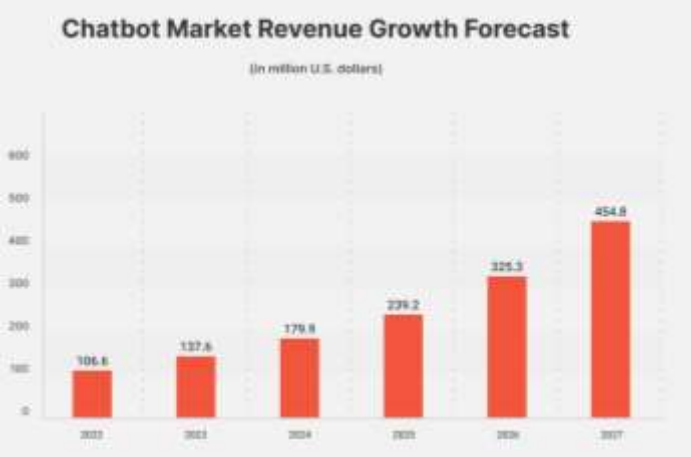

Figure 1: The expected revenue of the Chatbot market [3].

2 Literature Review

Many studies have been proposed to provide or improve FAQ chatbot systems. An early study [11] presented a semantic chatbot called FAQ FINDER. This chatbot used WordNet, a lexical dictionary of word semantics, to increase the degree of match between a question and its answer. The systems did not get a good recall percentage, around 67%. Another early agent in this field is ELIZA [12]. ELIZA was proposed during the 1960s to provide a close-to-human psychotherapist's conversation. However, ELIZA did not find its way in the area due to its ability to learn and the lack of knowledge it has [13]. In 1995, the ALICE chatbot was presented. It was built on the concept of direct matching between question and answer and used AIML, which is short for Artificial Intelligence Markup Language, a corpus of around 40,000 documents. Despite that, ALICE was not intelligent enough to capture many human questions.

Peyton et al. [14] Investigated using chatbots in online learning to support academic institutions managing increased student enrolment. They emphasise the importance of chatbots in addressing common questions related to qualifications and registration, allowing staff to focus on more complex interactions. The study compares the intent classification results from two chatbot frameworks with a Sentence BERT (SBERT) model. The methodology involves preparing a university FAQ dataset for training, revealing that the SBERT model outperformed the frameworks with an F1-score of 0.99. At the same time, Google Dialogflow and Microsoft QnA Maker scored 0.96 and 0.95, respectively, consistent with existing benchmarks in Natural Language Understanding (NLU).

Yigc et al. [15] examined the impact of large language models (LLMs) on natural language processing, focusing on their evolution and the emergence of generative chatbots like ChatGPT. Since its launch in 2022, ChatGPT has gained over 100 million users in just two months, highlighting the rapid adoption of this technology. The study explores LLM applications in various fields, notably higher education, where they offer personalized learning experiences and support asynchronous learning. However, concerns about academic integrity, misinformation, and bias remain prevalent. The authors discuss the development and characteristics of LLM-based chatbots, their potential benefits in education, and the challenges associated with their use, emphasizing the need for strategies to address these issues.

Mzwri et al. [16] explored the growing popularity of chatbots in automating tasks and providing consistent support in various domains, including education. However, they note the limitations of traditional chatbots in adapting to the evolving nature of knowledge. This research proposes an Internet wizard to enhance the knowledge base of an open-domain question-answering chatbot by leveraging search engines like Google. The chatbot can dynamically access real-time information from the Web by utilising features such as snippets and knowledge graphs, delivering relevant answers to user queries. A pilot study in higher education demonstrated the chatbot's effectiveness in generating responses across a wide range of topics, with positive feedback highlighting its ability to engage students and facilitate exploration of additional resources beyond the curriculum.

PI and Majid [17] investigated the growing demand for chatbots, particularly in academic settings, distinguishing between intelligent and simple chatbots. Intelligent chatbots facilitate enhanced interactions between users and institutions and are increasingly integrated into university websites to improve service levels. However, there was a lack of guidelines for developing these advanced chatbots. This study addresses this gap by identifying the core components necessary to create an intelligent chatbot for academic purposes. Through a literature

review covering the years 2017 to 2020, the researchers compiled a list of components that will be evaluated via surveys. The ultimate goal is to establish an intelligent chatbot academic model as a guideline for universities in chatbot development.

This field of study developed with the development of machine learning and deep learning models, especially in fields of understanding natural language and tracking dialogue, which established a different solution for machines to understand human dialogue and provide answers to [18]. A new era of conversational agents and chatbots has developed with the advancement of LLMs. These models can capture more than the semantic meaning of words, but also their accompanying information and representations in a sequence of words. ChatGPT, Gemini, Copilot, and many others have proved their ability to answer human questions and provide information in different domains, such as healthcare [19–21], education [22–24], military [25], climate change [26], machine translation [27] and many more.

The key limitations of the previous studies include either low accuracy or requiring significant computational resources with many parameters (billions of parameters), rather than the complexity of algorithms used, such as in the case of current LLMs. The proposed system uses classical NLP methods to process FAQs to determine the most relevant answers. This research did not use any LLMs because frequently asked questions are common in the domain, and the domain expert already provided answers. Therefore, the proposed system applies techniques, such as TF-IDF ranking and semantics, to rank answers and provide the most pertinent information.

3 Methodology

Figure 2 shows the steps of our proposed system. The chatbot receives a prompt from an end user. This prompt represents the question to which a user is seeking an answer. This prompt goes through many steps. First, it is lowercase and then preprocessed from stopwords, punctuation, numbers, and unnecessary characters. Extra steps are applied, such as lemmatisation and word expansion, which will be explained in the following sections. These are the methods used to process the prompt presented by a user.

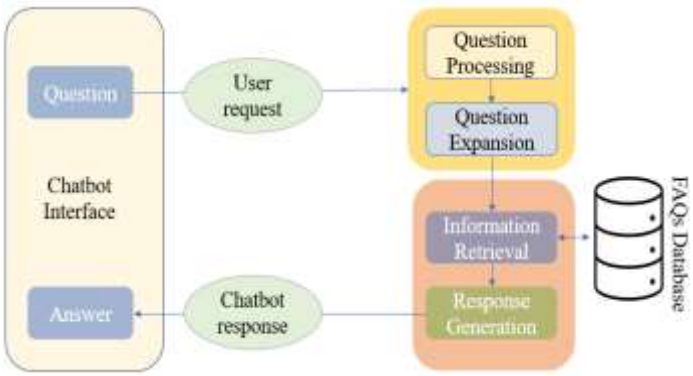

Figure 2: The proposed FAQs chatbot system.

3.1 Dataset Description

To evaluate our proposed system, we used two datasets. The first one is the Stanford Question Answering Dataset (**SQuAD**). This dataset consists of 100,000 question-answer pairs extracted from around 500 Wikipedia articles. The crowd workers created these questions and their answers. Each question in this dataset is paired with an answer that is a specific segment, or span, of text from the corresponding article [28]. This dataset is suitable for training and evaluating models for natural language understanding. This research employed the SQuAD dataset to assess chatbots' accuracy in providing related answers.

The other dataset is the **MS-Marco** (Microsoft MAchine Reading COmprehension) dataset. The dataset was introduced by Nguyen et al. in 2018. It has around 1,000,000 search queries asked via Bing or Cortana associated with related text passages from retrieved documents [29].

3.2 Preprocessing Steps

Several necessary steps are taken in the initial pre-processing to enhance clarity and relevance. These steps are explained below.

a) **Lowercasing:** Convert all text to lowercase to ensure consistency and avoid mismatches due to case differences (e.g., "Chatbot" vs "chatbot").

b) **Stopword Removal:** Remove common words that do not contribute significant meaning to the query, such as "and," "the," and "is," to focus on key terms.

c) **Punctuation removal:** Eliminate punctuation marks (e.g., commas, periods) from the text, as they are generally not helpful in meaning in this context.

d) **Number Removal:** To reduce noise, remove any numeric values unless they are essential to the query's meaning.

e) **Unnecessary Character Removal:** Strip any irrelevant characters or symbols (e.g., special characters) that could interfere with processing.

f) **Lemmatisation:** Transform words into their base or root form (e.g., changing "running" to "run") to unify different variations of a word, aiding in accurate matching.

3.3 TF-IDF Ranking in FAQ Chatbots

The TF-IDF is a popular concept in information retrieval and has been used in many applications requiring ranking [30]. In our work, we will use this concept to rank the words that are the most similar to the query words and to provide the most frequent

answer to the questions asked. Equation (1) of TF-IDF is as follows:

$$TF - IDF = TF(t,d) * IDF(t) \quad (1)$$

Here, TF(t,d) is the frequency of the term t in document d, representing how often the term appears in that document. The *IDF(t)* is the inverse document frequency of term t.

3.4 TF-IDF with Query Expansion

For this experiment, we will use the pre-trained Google vector embeddings generated by the Word2Vec algorithm trained on a collection of Google News (100 billion words) [31]. For every word *w* in a query *q*, we collect a set *s* of similar words where $s = \{x_1, x_2, \ldots, x_n\}$, where *x* is a word identical to a query word *w* and *n* is the number of similar words. For example, a query like "What is the cost of studying a master's in information technology?" This query will be processed from stopwords and punctuation and will be tokenised. So, the final query will be {cost, study, master, information, technology}. In our research, we chose *n*=2. The reason for selecting the top *n* =2 is that we want only the top and most similar words to the word *w*. Adding more similar words will add more noise to the query. Moreover, we did experiment on different *n* ranks and found that two is the best for our study. Carpineto and Romano (2012) reported that expanding the query with the top-ranked words is critical to retrieval precision, and too many similar terms can reduce retrieval effectiveness [32].

When we choose *n* = 2, the top two similar words will be added for every word in a query set. So, in the case of our previous example, the final query will look like: {cost, price, toll, study, perusal, perusing, master, maestro, overlord, information, information, data, technology, engineering, applied_science}. We can see how the query expansion helped by adding some good words like price, perusing, information, and others to the query. However, it also added some noise, such as overload, toll, and other noise, that could mess with the results. All that will be studied here in our analysis, and we will see if it adds better results.

3.5 System Evaluation

For the evolution of this project, we used accuracy, precision, recall, and F1-score metrics. These are standard metrics commonly used for measuring many retrieval systems. In addition to these metrics, the confidence metric was used to show the maximum similarity score between the expanded and the real question in the ground truth dataset. Here, we explain the equations and parameters of every metric used:

Accuracy represents the total of correct predictions divided by the total number of predictions [33].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \quad (2)$$

Recall measures the system's ability to capture relevant instances. It calculates the ratio of correct predictions to the total number of actual predictions in the ground truth dataset [34].

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

Precision measures the accuracy of optimistic predictions by indicating the proportion of correctly identified positive cases out of the total predicted positives.

$$\text{precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (4)$$

F1-score is a metric that combines precision and recall into a single value to provide a balanced evaluation of the model's performance [35].

$$F - measure = \frac{2 \times (Precision * Recall)}{Precision + Recall} \quad (5)$$

The final metric used is **confidence,** which represents a maximum cosine similarity score of how the expanded question $\hat{q}$ matches the real question q_i in the ground truth dataset. Here is the confidence equation:

$$confidence\ \hat{q} = \max(cosine - similarity(\hat{q}, qi)) \quad (6)$$

Where $\hat{q}$ is the user's expanded question and qi represents every question in the ground truth dataset.

4 Results

The work was implemented on a Kaggle notebook with a 4-core CPU and 30 GB of RAM. The work was done on the two datasets mentioned before, with the first step running without query expansion and the other with query expansion. The first dataset has around 12,000 unanswered questions, so we removed them because our FAQ chatbot deals with questions that should have answers. So, we only kept the 87599 questions along with their answers. The second dataset is full of questions and their answers so that we will use it directly. The results of the experiments for TF-IDF and query expansion (TF-IDF-Expa) methods with the two datasets are reported in Table 1 with 100% accuracy.

Table 1: Evaluation of the TF-IDF method with the SQuAD and MS-Marco datasets.

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Confidence (%)
SQuAD	100	100	100	99	99
MS-Marco	100	100	100	100	100

We can see from Table 1 that the TF-IDF method reported a good accuracy over the two datasets, with an accuracy of 100%. This means the FAQs chatbot can answer questions like those listed in the dataset. But what if the user asked for a query with the semantic meaning of a question listed on the FAQs dataset? Is the FAQs chatbot system going to be able to recognise such questions and answer them? To help the FAQs Chatbot answer such questions, we expanded the query with similar words from pretrained embeddings built using Word2Vec. This method with query expansion is called TF-IDF-Expa. Table 2 reports the results of the implementation of such a method.

Table 2: Evaluation of TF-IDF-Expa method with the SQuAD and MS-Marco datasets.

Dataset	Accuracy(%)	Precision(%)	Recall (%)	F1-Score (%)	Confidence (%)
SQuAD	100	100	100	99	70
MS-Marco	100	100	100	100	75

Table 2 shows that the results are the same as those reported in Table 1 regarding retrieval with an accuracy of 100%. However, we can see that the average confidence score is around 70, which means some differences exist between the original and expanded questions. This is true because the expanded question has more information than the original question. The expanded questions have two top-ranked similar words for every word in that question. Therefore, when measuring confidence using its formula in Equation (6), the similarity score between the user's asked questions and the questions in the ground truth dataset has some differences that lower the confidence score. Figure 3 shows the distribution of confidence scores between 65% and 75%.

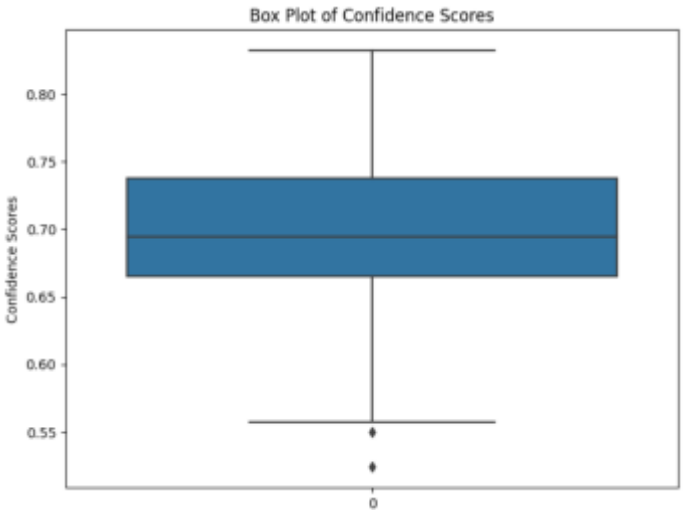

Figure 3: Distribution of confidence for the TF-IDF-Expa method.

The results show that the system did not miss any answer to any question asked with 100% accuracy. The advantage lies in the fact that, regardless of whether related or unrelated terms are included in the query, the system can still retrieve the relevant answer owing to the amount of information present in the question posed. The traditional TF-IDF method showed high confidence scores over all datasets, while the TF-IDF-Expa reported variants over the two datasets. This difference is due to the additional information inserted during the query expansion. This led to a more diverse set of matching terms and lower original term weights. Although both methods provided a high accuracy, the variation in confidence levels suggests that there may be trade-offs between expanding the semantic range of queries and maintaining high confidence in specific matches. This finding highlights the complexity of balancing semantic richness with precision in natural language processing tasks.

5 Discussion

It is important to compare our results with state-of-the-art studies, which we presented in Table 3, highlighting the performance of our proposed model, which achieved a high accuracy of 100%, surpassing all previous methods. Compared to traditional DL models, such as BERT-based architectures (Amer et al. [36]; Pandey and Roy [37]) and TaskBERT Gunnam et al. [38], our model shows a more effective retrieval mechanism. Although various previous approaches leveraged powerful transformers and attention mechanisms, they faced limitations regarding generalization capabilities and required extensive fine-tuning. Syed et al. [39] applied transfer learning and achieved an accuracy of 87.1%, highlighting the constraints in feature extraction and generalization. Arora et al. [40] proposed a heuristic approach with 78% accuracy, highlighting the challenges of rule-based approaches in large-scale retrieval tasks. On the other hand, the Retrieval Augmented Generation model through Vakayil et al. [41] reached 95% accuracy, which shows the strength of retrieval-based enhancements. Despite applying intelligent question-answering mechanisms, the CIQA mode presented by Hung et al. [42] achieved only 75.61% accuracy, showing a challenge with contextual reasoning. In contrast, our approach utilises TF-IDF, a statistical technique that effectively extracts the most relevant terms, ensuring precise document retrieval and answer selection. Additionally, our model was evaluated on both SQuAD and MS MARCO datasets, reinforcing its strength and flexibility across different benchmarks, while most previous methods focused on a single dataset. The lightweight nature of the TF-IDF approach also reduces model complexity, making it less prone to overfitting while maintaining high interpretability. These findings highlight the efficiency of retrieval-based methods over DL models in specific contexts and suggest that hybrid models integrating statistical and DL methods could further enhance performance in future research.

Conclusion

This research provided a chatbot system for the FAQs using two approaches: one using traditional TF-IDF and another combining

TF-IDF with query expansion (TF-IDF-Expa). The query is expanded using Google's pre-trained vectors, where each word is enhanced with the top 2 similar words from Google's vectors. Both methods achieved 100% accuracy in answering user queries, which shows a high effectiveness in finding answers to the asked question within the FAQ dataset. However, the results showed that the confidence levels differed between the two approaches. The traditional TF-IDF method showed high confidence scores over all datasets, while the TF-IDF-Expa reported variants over the two datasets. This difference is due to the additional information inserted during the query expansion. For future work, we suggest evaluating the system on larger datasets. We also recommend applying a hybrid approach using recent embedding models such as BERT or GPT and seeing how the improvement can be achieved.

Table 3: A comparison of our proposed system with other studies.

Reference	year	Proposed	Dataset	Accuracy
Syed et al. [39]	2021	Transfer Learning	SQuAD	87.1%
Amer et al. [36]	2021	Pre-trained Google BERT language model	SQuAD	96%
Arora et al. [40]	2022	Heuristic approach	ms Marco	78%
Gunnam et al. [38]	2024	TaskBERT	SQuAD	92.4%
Vakayil et al. [41]	2024	Retrieval Augmented Generation	ms Marco	95%
Pandey and Roy [37]	2024	BERT embedding, GRU, and attention mechanisms.	SQuAD	95%
Hung et al. [42]	2025	Intelligent question answering (CIQA) mode	SQuAD	75.61%
Proposed model	2025	TF-IDF	SQuAD and ms Marco	100%

Authors contribution

All authors of the paper contributed equally to this paper.

Conflict of interest

The paper presents the authors' contribution and the idea. There is no conflict of interest with any other research.

References

[1] Hussain, S., Ameri Sianaki, O. & Ababneh, N. A Survey on Conversational Agents/Chatbots Classification and Design Techniques. *in Web Artif. Intell. Netw. Appl*. **927**, 946–956, doi:10.1007/978-3-030-15035-8_93 (2019).

[2] Kumar, N. 65 Chatbot Statistics for 2025 — New Data Released. *DemandSage*. https://www.demandsage.com/chatbot-statistics/ (2025).

[3] Cherniak, K. Chatbot Statistics: What Businesses Need to Know About Digital Assistants. *Master of Code Global*. https://masterofcode.com/blog/chatbot-statistics (2024).

[4] R. R. Singh and P. Singh, "Chatbot: Chatbot Assistant ", *J. Manage. Serv. Sci*. **4**, pp. 1–19, https://doi.org/10.54060/a2zjournals.jmss.57 (2024).

[5] Shawar, B. A. & Atwell, E. Chatbots: Are they Really Useful?. *J. Lang. Technol. Comput. Linguist.* **22**, 29–49, https://doi.org/10.21248/jlcl.22.2007.88 (2007).

[6] Zhang, Z., & Huang, X. The impact of chatbots based on large language models on second language vocabulary acquisition. *Heliyon*. **10**, e25370, https://doi.org/10.1016/j.heliyon.2024.e25370 (2024).

[7] Chen, Z., He, Z., Liu, X. & Bian, J. Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases. *BMC Med. Inform. Decis. Mak*. **18**, 65, https://doi.org/10.1186/s12911-018-0630-x (2018).

[8] Chowdhery, A. et al. PaLM: Scaling Language Modeling with Pathways. *Preprint* at https://doi.org/10.48550/arXiv.2204.02311 (2022).

[9] DeepSeek-AI et al. DeepSeek-V3 Technical Report. *Preprint* at https://doi.org/10.48550/arXiv.2412.19437 (2024).

[10] Ibrahim, M. S., Eleiwy, J. A., Muhi-Aldeen, H. M., Al-Yasiri, Y. & Nafea, A. A. Human to Chatbot Text Classification Using Multi-Source AI Chatbots and Machine Learning Models. *J. Intell. Syst. Internet Things* **16**, 152–165, doi:10.54216/JISIoT.160113 (2025).

[11] Burke, R. D. et al. Question answering from frequently asked question files: Experiences with the faq finder system. *AI Mag.* **18**, 57–57, https://doi.org/10.1609/aimag.v18i2.1294 (1997).

[12] Weizenbaum, J. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM*. **9**, 36–45, https://doi.org/10.1145/365153.365168 (1966).

[13] Adamopoulou, E. & Moussiades, L. Chatbots: History, technology, and applications. *Mach. Learn. Appl*. **2**, 100006, http://dx.doi.org/10.1016/j.mlwa.2020.100006 (2020).

[14] Peyton, K. & Unnikrishnan, S. A comparison of chatbot platforms with the state-of-the-art sentence BERT for answering online student FAQs. *Results Eng*. **17**, 100856, https://doi.org/10.1016/j.rineng.2022.100856 (2023).

[15] Yigci, D., Eryilmaz, M., Yetisen, A. K., Tasoglu, S. & Ozcan, A. Large Language Model-Based Chatbots in Higher Education. *Adv. Intell. Syst.* **7**, 2400429, doi:10.1002/aisy.202400429 (2024).

[16] Mzwri, K. & Turcsányi-Szabo, M. Internet Wizard for Enhancing Open-Domain Question-Answering Chatbot Knowledge Base in Education. *Appl. Sci*. **13**, 8114, https://doi.org/10.3390/app13148114 (2023).

[17] Pi, S. N. M. S. & Majid, M. A. Components of Smart Chatbot Academic Model for a University Website. *in 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*. 1–6. doi:10.1109/ETCCE51779.2020.9350903 (2020).

[18] Wang, X. & Yuan, C. Recent advances on human-computer dialogue. *CAAI Trans. Intell. Technol.* **1**, 303–312, http://dx.doi.org/10.1016/j.trit.2016.12.004 (2016).

[19] Biswas, S. S. Role of Chat GPT in Public Health. *Ann. Biomed. Eng*. **51**, 868–869, https://doi.org/10.1007/s10439-023-03172-7 (2023).

[20] Kung, T. H. et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health*. **2**, e0000198, https://doi.org/10.1371/journal.pdig.0000198 (2023).

[21] Meyer, A., Soleman, A., Riese, J. & Streichert, T. Comparison of ChatGPT, Gemini, and Le Chat with physician interpretations of medical laboratory questions from an online health forum. *Clin. Chem. Lab. Med. (CCLM)*. **62**, 2425-2434, https://doi.org/10.1515/cclm-2024-0246 (2024).

[22] Sharma, S. & Yadav, R. Chat GPT–A technological remedy or challenge for education system. *Glob. J. Enterp. Inf. Syst.* **14**, 46–51, https://www.gjeis.com/index.php/GJEIS/article/view/698/640 (2022).

[23] Benčina, K. & Dubac Nemet, L. Is feedback in the eye of the beholder? Chat GPT vs. Google Gemini in giving feedback on students' oral presentation scripts. in *e-methodology: IX International Academic Conference on Internet Research in Medicine and Beyond.* https://www.croris.hr/crosbi/publikacija/prilog-skup/827852 (2024).
[24] Durmaz Engin, C., Karatas, E. & Ozturk, T. Exploring the Role of ChatGPT-4, BingAI, and Gemini as Virtual Consultants to Educate Families about Retinopathy of Prematurity. *Children.* **11**, 750, https://doi.org/10.3390/children11060750 (2024).
[25] Biswas, S. Prospective Role of Chat GPT in the Military: According to ChatGPT. *Qeios*, https://doi.org/10.32388/8WYYOD (2023).
[26] Biswas, S. S. Potential Use of Chat GPT in Global Warming. *Ann. Biomed. Eng.* **51**, 1126–1127, https://doi.org/10.1007/s10439-023-03171-8 (2023).
[27] Muhealddin, D. R., Salih, Y. M. M. & Taher, F. J. A Linguistic Evaluation of Google Translate Between Human and Machine Translation for English-Sorani Kurdish Languages Using Natural Language Processing (NLP). *Passer J. Basic Appl. Sci.* **6**, 351–368, https://doi.org/10.24271/psr.2024.422619.1415 (2024).
[28] Rajpurkar, P., Zhang, J., Lopyrev, K. & Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv*. https://doi.org/10.48550/arXiv.1606.05250 (2016).
[29] Bajaj, P. et al. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv*. https://doi.org/10.48550/arXiv.1611.09268 (2018).
[30] Salh, D. A. & Nabi, R. M. Kurdish Fake News Detection Based on Machine Learning Approaches. *Passer J. Basic Appl. Sci*. **5**, 262–271, https://doi.org/10.24271/psr.2023.380132.1226 (2023).
[31] Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *arXiv*. 1301.3781, https://doi.org/10.48550/arXiv.1301.3781 (2013).
[32] Carpineto, C. & Romano, G. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Comput. Surv.* **44**, 1:1-1:50, https://doi.org/10.1145/2071389.2071390 (2012).
[33] Rafa, S. A. *et al.* A Birds Species Detection Utilizing an Effective Hybrid Model. in *2024 21st International Multi-Conference on Systems, Signals & Devices (SSD)* 705–710, doi: 10.1109/SSD61670.2024.10549480 (2024).
[34] Hussein, D. J., Rashad, M. N., Mirza, K. I. & Hussein, D. L. Machine Learning Approach to Sentiment Analysis in Data Mining. *Passer J. Basic Appl. Sci*. **4**, 71–77, https://doi.org/10.24271/psr.2022.312664.1101 (2022).
[35] Kadhim, O. J., Nafea, A. A., Aliesawi, S. A. & Al-Ani, M. M. Ensemble Model for Prostate Cancer Detection Using MRI Images. in *2023 16th International Conference on Developments in eSystems Engineering (DeSE).* 492–497, https://doi.org/10.1109/DeSE60595.2023.10468866 (2023).
[36] Amer, E. Hazem, A., Farouk, O., Louca, A., Mohamed, Y., Ashraf, M. A proposed chatbot framework for COVID-19. in *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC).* 263–268, https://doi.org/10.1109/MIUCC52538.2021.9447652 (2021).
[37] Pandey, A. K. & Roy, S. S. Extractive question answering over ancient scriptures texts using generative AI and natural language processing techniques. *IEEE Access.* **12**, 101197 - 101209, https://doi.org/10.1109/ACCESS.2024.3431282 (2024).
[38] Gunnam, G. R., Inupakutika, D., Mundlamuri, R., Kaghyan, S. & Akopian, D. Hybrid Machine Learning Approach for Task-Oriented Dialog Systems. *Int. J. Comput. Applications.* **186**, 35-42, http://dx.doi.org/10.5120/ijca2024923679 (2024).
[39] Syed, Z. H., Trabelsi, A., Helbert, E., Bailleau, V. & Muths, C. Question answering chatbot for troubleshooting queries based on transfer learning. *Procedia Comput. Sci*. **192**, 941–950, http://dx.doi.org/10.1016/j.procs.2021.08.097 (2021).
[40] Arora, S. & Chatterjee, N. Automated Evaluation Approach for Time-Dependent QA Pairs on Web Crawler-Based QA System. in *2022 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI).* 1–6, http://dx.doi.org/10.1109/SOLI57430.2022.10294578 (2022).
[41] Vakayil, S., Juliet, D. S. & Vakayil, S. RAG-Based LLM Chatbot Using Llama-2. in *2024 7th International Conference on Devices, Circuits and Systems (ICDCS).* 1–5, http://dx.doi.org/10.1109/ICDCS59278.2024.10561020 (2024).
[42] Hung, C. & Wu, M.-H. A Classification Intelligent Question Answering Model for Retrieval-Based Chatbots. *Adv. Manag. Appl. Econ*. **15**, 1–3, http://dx.doi.org/10.47260/amae/1513 (2025).