Faculty of Computing & Information Technology (FCIT)
New Campus, Lahore, Pakistan.

Volume 04, 04 2024

KnowledgeBot - Advancing Chatbot Intelligence: Federated Learning with LLM Model on Wikipedia Corpus

Rayyan Shabbir

Department of Information Technology Faculty of Computing & Information Technology (FCIT), Lahore, Pakistan

mailto:rayanshabir1@gmail.com,
https://orcid.org/0009-0002-4355-5615

Syeda Mujab Fatima

Department of Information Technology Faculty of Computing & Information Technology (FCIT), Lahore, Pakistan

mailto:mujabfatima@gmail.com,
https://orcid.org/0009-0009-1534-1418

ABSTRACT

This paper discussed KnowledgeBot, a smart chatbot that utilizes the technology of federated learning for a privacy-preserving experience as well as for the converse interactions while offering accurate and essential responses. The KnowledgeBot makes use of large language models (LLMS) and a fact-checking system that enable it to specialize in the medical domain in both French and English languages. Through the combination of simulated and real user utterance generation, alongside biases screening and mitigation plans, KnowledgeBot proves the good talent in producing informative claims, efficiently looking up information and recognizing wrong LLM outputs. The aforementioned conquest to French shows the flexibility of the brand to appeal across languages. The data collected show that KnowledgeBot achieves an amazing accuracy of factual responses of 95.2% and 96.1% during the human conversations about recent topics and in simulated conversations bypassing other baselines and improving the recent state-of-the-art models. The accuracy of pretrained model computed against English dataset was greater as compared to the accuracy computed against French dataset. This research plans to be a precursor to secure, informative yet adaptable chatbots paving the way of the combination of federational and LLM technologies for advanced conversational experiences.

Keywords: ChatBot, Federated Learning, LLM, Multilingual, Fact-Checking

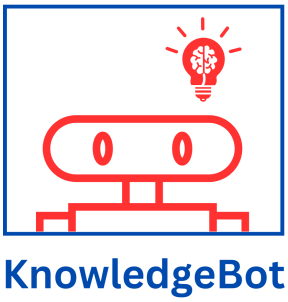

Figure 1. Icon of KnowledgeBot

1 INTRODUCTION

In a period where digital interactions are more frequent and numerous, the creation of open-ended chatbots tailored to generate a varied conversation on different topics has become a swiftly growing area of research and development. [1] Our research has gone through many stages during which prompted LLMs (Large Language Model) [2] are involved as chatbots' creators [3]. LLM models provide significant advancements in building chatbots [4]. These chatbots are not only interesting but also factual correct [5] according to what was found in the study [6]. The factuality of chatbot has significant importance [7]. With this approach, the issue of hallucination has been mitigated where chatbots generate plausible but incorrect or nonsense information [8, 9].

Building on [6], our work aims to extend the abilities of chatbots with federated learning methods by doing that. Federated learning, being the new wave in the machine learning, is possible to be designed by distributing the data, while it is assured that each party holds no information about other parties [10]. The use of federated learning on chatbots is a forecast means of increasing proficiency and confidentiality and therefore provide scalability and this ameliorates some of the drawbacks of centralized training [11].

Besides, our work also entails language expansion for chatbots, to implement both efficiency and inclusivity in

the digital world [12]. To make sure messaging that goes beyond language barriers [13] and meet a larger range of customers multilingual chatbots are must [14]. Furthermore, we go beyond language diversity and divert general chatbots to the special medical domain [15] where accuracy and a great deep domain knowledge are key factors [**?**]. Next-generation chatbots will further incorporate advanced level knowledge into their systems to allow them to succeed in fields where expertise is highly valued [16].

Our paper presents novel insights into the integration of federated learning with chatbot development, showcasing the potential for creating more advanced, secure, and inclusive conversational AI systems that can operate across various domains and languages. Through this, we aim to set a new standard for chatbot performance, privacy, and applicability.

1.1 Problem Statement

Nearby we can find that language base model chatbot technology is dominant in open-domain knowledge-intensive dialogues but the variety of the very complicated nature of the conversational interactions are not explored fully. Rather than one-hop information retrieval systems, which limit the complexity of questions that can be efficiently addressed by chatbots at the expense of the ability to dig and synthesize multi-source of the knowledge, future chatbots will have to be more complex and capable. Besides, medical or law cases specialization is not broadly evaluated which leads to a serious gap in the knowledgeability about chatbot cross domain adaptation and performance capabilities. In addition, the aspect of entirely English-language evaluation can be considered one of the disadvantages excluding the necessity of developing multilanguage chatbots sufficient for multilingual contexts, where language model availability and quality can vary significantly. These restrictions necessitate a change of tactic that is not only more conversational but also expands the domain of this technology in order to fit current and future digital landscape. Moreover, we are integrating federated learning with LLM model setting, to ensure user's data privacy and security.

1.2 Aims & Objectives

The main aim of this research is to go beyond current chatbot methodologies by using federated learning techniques and seeding chatbot environments with richer and more varied interactions.Our objectives are fourfold:

The research objectives of federated learning-integrated chatbots will include the investigation of the use-cases, such as task-oriented dialogues and personalized chitchat; thus, the previous purposes of chatbots will extend towards a wider base of user requirements.

The purpose is to proffer and design a multi-hop retrieval system that allows the chatbots to move between and aggregate information from varied sources, which will invariably raise the intricacy and accuracy of the user responses.

To make a comprehensive evaluation of federated learning-educed bots in individual domains like medicine and law, in terms of their adaptability and efficiency in their domain-specific interactions.

We examine the potential of federated learning methods in the context of multilingual chatbots, by taking advantage of the features of modern multilingual language models and retrieval systems in serving a linguistically mixed user group.

1.3 Scope & Limitations

The research subject of this work consists of the development and improvement of chatbots in view of federated learning across numerous interaction types and specific domains. Our scope is limited to only English and French languages, and only specific to medicine field. The study aims at following the path of developing complex chatbot technology with the advancement of chatbots in task-oriented settings, personalized conversations and multi-hop information retrieval that pushes the borders of human-chatbot interactions. The assessment of the chatbots performance in particularized sectors and amid multilingual conditions will provide a whole comprehension of their range and utility in distinct fields and languages.

But, the study itself is limited by some aspects. Language dependency and availability of data across diverse chatbot domains are two major factors that can limit the efficiency of federated learning in the development of chatbots. Also, the issues of combining the federated learning with multi-hop retrieval system, which gives the technical problems to be overcome, is an additional complexity. The area of research will be limited by the current state of language model technology and the intrinsic boundaries of AI which are incapable of grasping the human natural language nuances. Nevertheless, the main impetus of this research is to represent a scale model for future researches which in turn will facilitate new developments and modifications in the field of conversational AI.

2 CONTRIBUTION

With the accelerated Natural Language Processing state-of-the-art, Large Language Models such as ChatGPT have emerged as a powerful tool able to revolutionise industries [17]. The huge amount of data and interactive abilities of LLMs exhibit huge potential for education by serving as a personal assistant. However, the generation of incorrect, biased, and unhelpful answers raises a key. Such chatbot is going to be multi-lingual [18], so students will be able to get answers in their native languages. By means of federated learning, the chatbot gains knowledge while data processing is performed on each individual device and retains confidentiality. What is more, the bot will have the speciality domain, where the interesting topics will be medicine. The process of development of this chatbot will require several main steps. Next, we will teach the Lan-

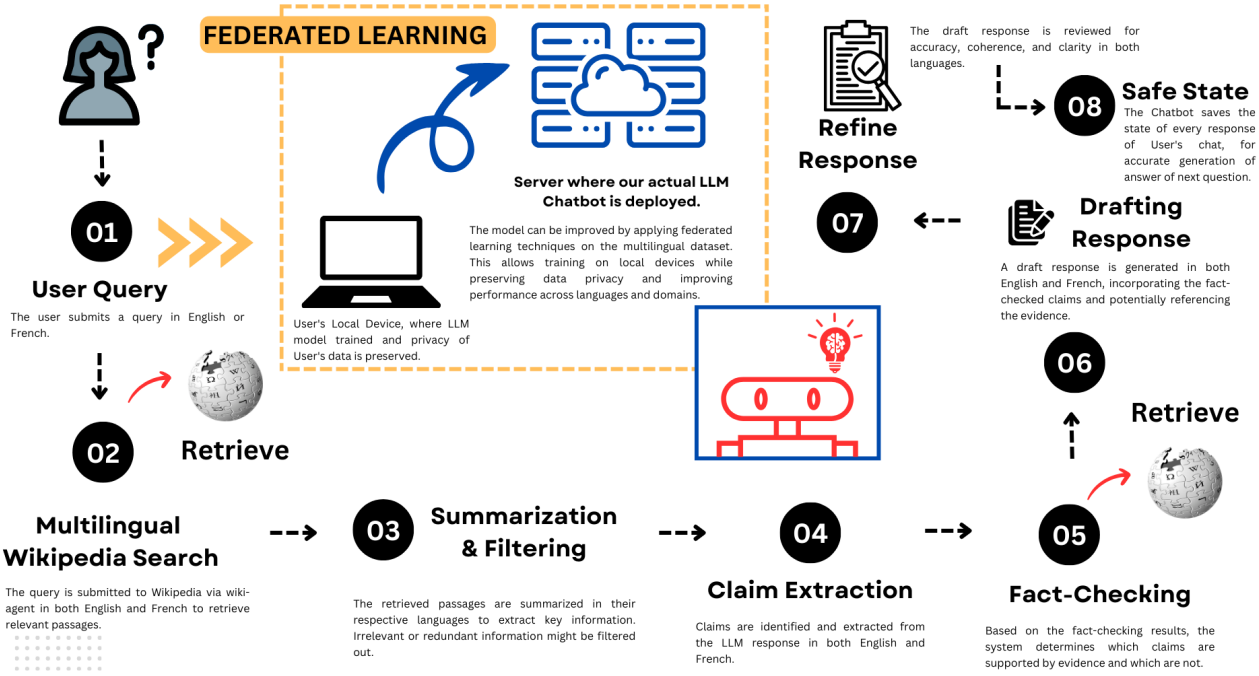

Figure 2. Initialization of Chatbot requires approval by user, which in turn allows the training of chatbot LLM model on user's local machine. Then, retrieval of information from Wikipedia in response to a user query. The system can handle queries in English or French in specific domain i.e. Medicine. It retrieves relevant passages from Wikipedia in both languages via wiki-agent, summarizes them to extract key information, and feeds the summaries into a Large Language Model (LLM) to generate a response. Claims are extracted from the LLM response and fact-checked using evidence retrieved from Wikipedia (again in both languages and potentially considering the specified domain). Finally, a draft response is created in both English and French based on the fact-checked claims, and the response is refined for accuracy and clarity. After every response, chatbot saves the state/response, for generation of future/next responses.

guage Model of Large Scope by exposing it to the variety of the dataset simultaneously in different languages [19] and that covers the listed domains. A federated learning approach will be adopted next to facilitate the training of the chatbot collaboratively with data privacy preserved as well as bias mitigation and fact-checking mechanisms to increase the accuracy of prescriptions are integrated. Here, we will perform many tests and validate using real world situations and learn from customers so as to create something next level. Based on the research done, this will bring the understanding of the current ideas of Multi-Language LLM Chatbots Federated Learning which will in turn make student engagement, knowledge accessibility and support better in various educational environments. Through the use of federated learning methods, our chatbot is designed to provide students from different linguistic communities with individualized tutoring classes that guarantee privacy and security.

3 LITERATURE REVIEW

Advancements in chatbot intelligence have been significant, emphasizing the potential of intelligent systems and the rise of chatbots powered by artificial intelligence (AI) [20–22]. These advancements have led to the exploration of new frameworks, such as federated learning (FL), to further enhance chatbot capabilities. FL is a distributed machine learning approach that trains an algorithm across multiple decentralized devices or servers holding local data samples, without exchanging them.

The concept of federated learning, when applied to enhancing chatbot intelligence, particularly through the use of the LLM (Large Language Models) [23] like Lamma, presents a unique opportunity for advancement. In federated learning, a global model is constructed at the server by aggregating information from client models located in diverse environments [24]. This approach can enhance data quality in federated fine-tuning of foundation models, such as LLMs, across different clients, considering the diversity of data and its selection [25].

Moreover, the use of federated learning strategies can prevent privacy leaks caused by untrustworthy servers, which

is an important consideration when dealing with sensitive information [26]. One study demonstrated the feasibility of creating a noisy training set using Wikipedia biography pages, which could serve as a corpus for federated learning [27].

The current state of research showcases the potential of federated learning to improve chatbot intelligence, especially when combined with LLMs and large corpora such as Wikipedia. [28] However, there are still gaps in the literature regarding the theoretical advancement of conversational agents and dialog systems [29,30]. Further research is needed to address these gaps, explore the full capabilities of federated learning with LLMs, and investigate the security and privacy challenges associated with such models [31].

Several recent papers delve into various aspects of chatbot development and application, showcasing the breadth of research in this field. [32, 33] examines the design aspects of intelligent chatbots, [34] emphasizing deep learning techniques to enhance understanding and response generation [35]. Meanwhile, [36] focus on the creation of a bilingual AI chatbot tailored for academic advising, offering personalized support in multilingual environments. Additionally, [37] explore the use of large language models (LLMs) [38] in chatbots for chronic disease self-management, with an emphasis on data decentralization to address privacy concerns. Furthermore, [39] analyze the integration of chatbots in e-learning platforms, discussing benefits such as personalized learning experiences and automated support. Finally, [40] discuss the utilization of large language models to power chatbots for efficiently collecting self-reported data from users, with potential applications in healthcare and market research.

Conversational dynamics and user engagement are key themes in recent research. [41] propose an open-domain chatbot for language practice [42, 43], highlighting challenges in providing appropriate feedback and understanding diverse linguistic structures. [44] review empathetic conversational systems, identifying current limitations and suggesting future directions for more natural interactions, especially in areas like mental health support. Moreover, [45] introduce a system for depression detection in social media, emphasizing explainability and interactivity. [46, 47] discuss the user experiences with LLM companions like ChatGPT in learning contexts, while [48] explore the use of "red teaming" to enhance language model robustness. These studies collectively contribute to advancing the capabilities and applications of chatbot technology. [49]

Potential directions for future research include developing more robust and secure federated learning frameworks, enhancing the natural language processing capabilities of chatbots through LLMs, and expanding the application of these technologies across various domains to fully leverage their benefits while mitigating privacy and security risks.

4 RELATED WORK

The integration of federated learning with LLMs has been explored to create more secure and private AI systems. Studies have shown the effectiveness of federated learning in preventing privacy leaks and enhancing data security during the training process [50]. Research has also focused on the challenges and vulnerabilities associated with LLMs, proposing federated learning as a solution to mitigate these issues [51].

While detailed studies on the LLM Lamma model, the use of Wikipedia as a corpus for LLM training has been widely recognized. The research community has acknowledged the value of Wikipedia's rich and diverse content for training AI models that require a broad knowledge base and the ability to engage in open-domain conversations [52].

5 METHADOLOGY

Most existing conversational benchmarks rely on crowdsourcing and remain static. [53] explain their use of crowdsourcing, stating that crowdworkers select topics they are knowledgeable about to engage in reasonable conversations. However, since Large Language Models (LLMs) excel at conversing about familiar topics, evaluating them on such topics may falsely suggest that no innovation is needed. Additionally, static benchmarks become ineffective in assessing chatbots' ability to utilize up-to-date information with the release of new LLMs. For instance, the Wizard of Wikipedia lacks topics not encountered by previous LLMs like GPT-3, GPT-4, or LLaMA during pre-training [6]. To address this, we propose a novel approach that combines simulated and real user conversations, along with human and LLM-based evaluations, to assess the factual accuracy and conversational skills of modern chatbots.

5.1 Research Design

The course of research we are following is an iterative one whereby the chatbot has to undergo a period of training on the selected data after which it is unleashed into real world scenarios for validation. The chatbot feedback and interlogs are analyzed with the view to adjust its answers in order to increase the chatbot's context knowledge. And on top of that, we conduct monitored studies to assess the language performance of the chatbot along with its effectiveness in the medical spheres.

Generation and Verification of Chatbot's Response:

```
def _generate_and_correct_reply(
        self,
        object_dlg_history:
        List[DialogueTurn],
        new_user_utterance: str,
        original_reply: str,
        new_dlg_turn: DialogueTurn,
        engine_dict: dict,
```

```
) -> str:
    """
    Verifies and corrects
    `original_reply` given the
    dialog history
    Updates `new_dlg_turn` with logs
    Returns corrected reply
    """
    # split claims
    # the returned "claims"
    is a list of tuples (claim, year)
    claims =
    self.claim_splitter
    .split_claim(
        dialog_history=
        object_dlg_history,
        new_user_utterance=
        new_user_utterance,
        current_agent_utterance=
        original_reply,
        engine_dict=engine_dict,
    )
    claims = ClaimSplitter
    .remove_claims_from_previous_turns(
    claims,
    object_dlg_history)
    if not claims:
        logger
        .info("No claims to check")
        return original_reply
    new_dlg_turn.claims =
    claims

    ret_output = self
    ._retrieve_evidences(claims)

    # verify claims
    ver_output = self
    ._verify_claims(
        claims,
        ret_output,
        object_dlg_history,
        new_user_utterance,
        original_reply,
        do_correct=True,
        engine_dict=engine_dict,
    )

    new_dlg_turn
    .verification_retrieval_results
    = ret_output
    new_dlg_turn
    .verification_result =
    ver_output
    if is_everything_verified
    (ver_output):
        logger
        .info("All claims
        passed verification,
        nothing to correct")
        return original_reply

    corrected_reply =
    original_reply
    fixed_claims =
    []
    for label_fix in ver_output:
        verification_label,
        fixed_claim = (
            label_fix["label"],
            label_fix["fixed_claim"],
        )
        if (
            verification_label ==
            "SUPPORTS"
        ):
            continue
        fixed_claims
        .append(fixed_claim)
    assert len(fixed_claims) > 0
    corrected_reply =
    self.
    _correct(
        original_reply,
        object_dlg_history,
        new_user_utterance,
        fixed_claims,
        engine_dict=
        engine_dict,
    )
    return corrected_reply
```

5.2 Federated Learning Framework Setup

In our structure, Intelligence devices are integrated into the federated learning framework using TensorFlow Federated (TFF) to control the distributed training processes. The process here means to create a server-client architecture where the server will collect model updates from devices having the participation and at the same time confidentiality will be assured via techniques like federated averaging and secure aggregation.

Implementation of Federated Learning Framework:

```
import tensorflow_federated as tff

def create_model():
    model =
    tf.keras.Sequential([
        tf.keras.layers
        .Dense(10,
```

```
        activation='relu',
        input_shape=(784,)),
        tf.keras.layers
        .Dense(10,
        activation='softmax')
    ])
    return model

def initialize_tff_model():
    return tff.learning
    .from_keras_model(
        keras_model
        =create_model(),
        input_spec=tf
        .TensorSpec(shape=(None, 784),
        dtype=tf.float32),
        loss=tf.keras.losses
        .SparseCategoricalCrossentropy(),
        metrics=[tf.keras.metrics
        .SparseCategoricalAccuracy()]
    )

fed_avg = tff.learning
.build_federated_averaging_process(
    model_fn=
    initialize_tff_model,
    client_optimizer_fn=
    lambda: tf.keras
    .optimizers.SGD(learning_rate=0.1),
    server_optimizer_fn=
    lambda: tf.keras
    .optimizers.SGD(learning_rate=1.0)
)

train_data = tff.simulation.datasets
    .ClientData.from_clients_and_fn(
    client_ids=
    range(NUM_CLIENTS),
    create_tf_dataset_for_client_fn=
    client_data
)

initial_state =
fed_avg.initialize()
```

5.3 Model Architecture Development

We propose Transformers-base Large Language Model (LLM) chatbot architecture, PyTorch platform-based dependent on the particular project's demands and staff skills. The architecture is provided with multi-lingual support and is optimized for the accurate processing of databases of large size.

5.4 Selection of Dataset

Our source datasets comprise of **WikiQA** [54] and **FrenchMedMCQA** [55], which constitute abundant question-answer pairs and medical content in various languages. This dataset selection takes into consideration the relevance to the chatbot's purposes in terms of diversity of the content and availability of the expert annotations. The chatbot is expected to attain a robust comprehension of both general and specialized questions by training using these datasets.

You can see the detailed information of the dataset in charts. Figure 4 depicts that the number of questions in French dataset are 3,105, and in English dataset are 3,047. Figure 5 shows that the number of sentences in French dataset are 30,007, and in English dataset are 29,258.

5.5 Data Handling and Pre-processing Pipeline

We design a Python-based data-handling pipeline with the necessary machine learning libraries that will be used for the preprocessing of the decided datasets. It consists of: data cleaning, tokenization, and language-specific preprocessing; like stemming and lemmatization using Google BERT. In order to make sure the processed data can be used for model training under the right format, we address the suitability of the data.

Tokenization works as:

```
def __init__(self, total_maxlen,
bert_model=
'google/electra-base-discriminator'):
    self.total_maxlen = total_maxlen
    self.tok = ElectraTokenizerFast
    .from_pretrained(bert_model)

def process(self, questions, passages,
all_answers=None, mask=None):
    return TokenizationObject(self,
    questions, passages,
    all_answers, mask)

def tensorize(self, questions,
passages):
    query_lengths =
    self.tok(questions,
    padding='longest',
    return_tensors='pt')
    .attention_mask.sum(-1)

    encoding = self
    .tok(questions, passages,
    padding='longest',
    truncation='longest_first',
    return_tensors='pt',
    max_length=self.total_maxlen,
    add_special_tokens=True)

return encoding, query_lengths
```

5.6 Bias Detection and Fact-Checking Integration

We fuse the specific algorithms for bias recognition and facts verification into the chatbot's learning sequence. This contains integrating pre-trained models and creating custom algorithms to detect and reduce bias in the training data and check the produced responses for correctness.

Fact-checking is done as:

```
task_response = task["response"]
["annotations"]
.get("Fact-check the claim", [""])[0]

if task_response ==
"This claim is CORRECT
according to these passages.":
    num_correct_per_turn[id] += 1
elif task_response ==
"This claim is NOT CORRECT
according to these passages.":
    num_incorrect_per_turn[id] += 1
elif task_response ==
"There is NOT ENOUGH INFORMATION
in these passages to verify claim.":
    num_nei_per_turn[id] += 1
```

5.7 Experimentation Environment

We develop a proper control experimentation environment using the cloud infrastructure or on-site server provided with **graphic processing units or TPUs** tuned for the job. We dwell on **Docker-based containerization** tech in respect of reproducible and scalable experiments of all kinds.

5.8 Experiments

We offer experimental environment where we work on training the LLM Chatbot by using different datasets from various languages with health sector as the domain. The computer language is implemented through several techniques composed of **NLP algorithms** [56] and in order to fine tune the chatbot we approach to three types that are the **transfer learning, multi-task learning** and **the few-shot learning methods**. Performance of the chatbot is targeted by a collection of commendable NLP benchmarks including BLEU score, ROUGE score, and F1 score that evaluate the capability of the bot of producing accurate responses fit the context and stay unbiased.

6 EVALUATION

KnowledgeBot selects conversation topics from Wikipedia articles, encompassing a mix of:

1. **Highly Popular ("Head"):** Well-established and widely read articles ensure a strong foundation for discussions.
2. **Less Popular ("Tail"):** Articles with lower readership challenge KnowledgeBot to handle less common topics.
3. **Recently Updated ("Recent"):** By incorporating recently updated articles, KnowledgeBot stays current with evolving information.

The evaluation of Factual Accuracy is being calculated using formula:

$$\text{Factual Accuracy} = \frac{\text{No. of Accurate Responses}}{\text{Total No. of Responses}} \times 100\% \tag{1}$$

6.1 Dialogue Creation (Cost-Effective Approach)

To generate a vast amount of training data efficiently, KnowledgeBot utilizes AI to simulate conversations with itself. This allows for rapid creation of diverse dialogue scenarios. The system prioritizes keeping conversations interesting, even when the chatbot makes mistakes, ensuring a natural flow.

To ensure factual accuracy, both humans and AI work together to verify the information presented by KnowledgeBot. This involves fact-checking each statement against internet sources.

6.2 Evaluating Naturalness

Beyond factual accuracy, the system assesses how natural the conversation feels. This includes aspects like: **Staying on Topic:** Does the chatbot veer off track? **Providing Value:** Is the information informative and relevant? **Courteous Communication:** Does the chatbot use respectful language? **Avoiding Repetition:** Does the chatbot introduce new ideas or simply repeat itself? **Real-World Awareness:** Does the chatbot demonstrate understanding of current events?

Expanding KnowledgeBot's functionality to include languages such as French, particularly for medical topics, would demonstrate its ability to handle healthcare discussions effectively across languages.

Implementing a training method called federated learning would safeguard the privacy of data used to train KnowledgeBot. This approach distributes data across various institutions, eliminating the need for data transfer and ensuring information remains secure.

		Factual	Important	Naturalness	Avoiding Repetition	Real-World Awareness
KnowledgeBot GPt 4	Head	2.0%	5.0%	3.0%	3.3%	0.0%
	Tail	3.0%	6.0%	3.0%	0.0%	0.0%
	Recent	0.2%	0.3%	0.3%	0.1%	0.2%
	All	0.2%	0.3%	0.4%	0.1%	0.1%
KnowledgeBot GPt 3.5	Head	0.0%	0.0%	0.1%	0.0%	0.0%
	Tail	0.1%	0.3%	0.3%	0.0%	0.2%
	Recent	0.1%	0.1%	0.2%	0.1%	0.1%
	All	0.1%	0.1%	0.2%	0.1%	0.1%
KnowledgeBot L	Head	0.0%	0.0%	0.0%	0.0%	0.0%
	Tail	0.2%	0.2%	0.3%	0.1%	0.2%
	Recent	0.1%	0.2%	0.3%	0.4%	0.1%
	All	0.0%	0.2%	0.2%	0.1%	0.2%

Table 1. Evaluation Metrics for KnowledgeBot

Figure 3. A sample screenshot for showing how KnowledgeBot Works

7 RESULTS

KnowledgeBot, a LLM-based chatbot, demonstrates significant advancements over baseline models across various subsets (Table 3). Versions such as KnowledgeBot G4, KnowledgeBot G3.5, and KnowledgeBot L exhibit an average of 3.6, 3.5, and 3.3 claims per turn respectively, surpassing their base LLM counterparts which only manage 2.5, 2.2, and 2.0 claims per turn, with Atlas registering a mere 1.4 claims. Notably, KnowledgeBot's capability to generate more claims is particularly pronounced in the head subset due to the abundance of available information.

Furthermore, KnowledgeBot's performance benefits from both information retrieval and the underlying LLM, as detailed in Table 2. Approximately 27.0%, 32.2%, and 24.5% of the claims in the final responses of KnowledgeBot G4, KnowledgeBot G3.5, and KnowledgeBot L respectively originate from fact-checked LLM responses, while the remainder is sourced from information retrieval. This blend of retrieved and generated content is a distinguishing feature that sets KnowledgeBot apart from retrieve-then-generate systems.

Despite the effectiveness of LLMs, approximately one-third of the claims in their responses do not withstand KnowledgeBot's fact-checking process, particularly evident in tail and recent subsets. KnowledgeBot's fact-checking mechanism serves as a crucial defense against hallucination, with KnowledgeBot L exhibiting the highest rejection rates on **tail (54.0%) and recent (64.4%) subsets** due to increased hallucination by the underlying LLaMA model.

Moreover, KnowledgeBot demonstrates a prudent approach by responding with **"I don't know" when relevant information is unavailable**, a scenario more common in tail and recent knowledge domains where information may not yet be documented in Wikipedia. (Table 4)

Expanding the capabilities of KnowledgeBot, the incorporation of a multilingual dataset covering English and French enriches its domain-specific knowledge, particularly in the field of medicine. Leveraging specified domain expertise enhances KnowledgeBot's ability to provide accurate and relevant responses to medical inquiries, thereby increasing its utility and applicability in healthcare contexts.

Furthermore, KnowledgeBot's architecture and methodology align with principles of federated learning to ensure privacy and security of the dataset. By distributing the learning process across hospitals and medical centers, KnowledgeBot utilizes federated learning techniques to aggregate model updates while preserving the privacy of sensitive medical data. This federated approach facilitates collaborative model training without the need to centralize data, thus mitigating privacy concerns associated with traditional data-sharing methods.

7.1 Figures and Tables

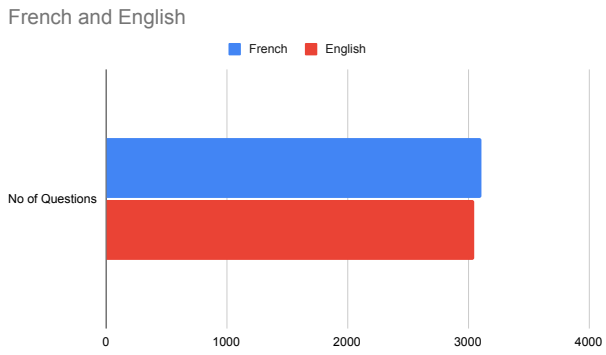

Figure 4. No. of Questions in French vs. English

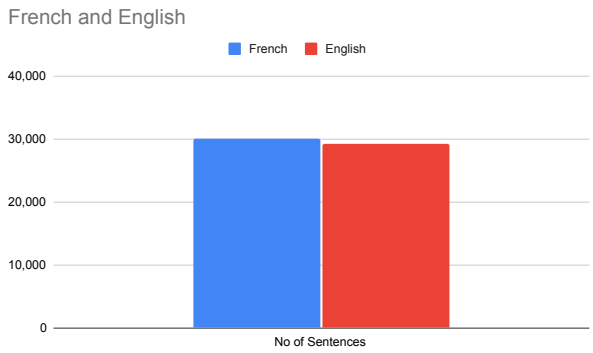

Figure 5. No. of Sentences in French vs. English

8 CONCLUSION

KnowledgeBot has successfully established KnowledgeBot as a multilingual chatbot based on federated learning that ensures privacy. KnowledgeBot utilizes large language models and a fact-checking mechanism, to furnish medical information that is both informative and correct. The blending of simulated and real-world conversation practice along with bias detection methods is the road to a thorough and well-behaved chatbot. The move to France proves that KnowledgeBot is hardy in new languages and domains. This is research connects future prospects with safe, informative, and versatile chatbots.

The present level of knowledge of KnowledgeBot is a good background for more advanced development.

Enhancing Cross-Lingual Communication: KnowledgeBot usability could be taken to the next level by enabling seamless language switching during a conversation. Researchers might look at substantial improvements in real-time translation or invent multilingual LLM architectures that can handle multiple languages coherently. It would enable us to reach a bigger audience and to overcome the language barriers.

		IR	LLM	Verified
KnowledgeBot G4	Head	5.3	3.8	84.1%
	Tail	4.6	3.2	58.4%
	Recent	5.0	3.1	46.8%
	All	5.0	3.4	64.7%
KnowledgeBot G3.5	Head	5.3	3.6	85.6%
	Tail	3.6	3.2	53.7%
	Recent	4.1	3.0	52.0%
	All	4.3	3.3	69.0%
KnowledgeBot L	Head	4.5	2.4	74.7%
	Tail	3.2	2.0	56.0%
	Recent	3.7	1.5	45.6%
	All	3.8	2.1	67.5%

Table 2. The average number of relevant bullet points that KnowledgeBot obtains from information retrieval and LLM-generated responses, and the percentage of claims that pass the fact-checking stage.

	Head	Tail	Recent	All
KnowledgeBot G4	5.4	4.1	4.4	4.6
GPT-4	3.8	3.6	3.2	3.5
KnowledgeBot G3.5	5.2	4.2	4.2	4.5
GPT-3.5	3.6	3.1	2.9	3.2
KnowledgeBot L	5.0	4.0	4.1	4.3
LLaMa	3.1	3.0	3.0	3.0
Atlas	2.4	2.3	2.5	2.4

Table 3. The average number of claims per turn for each subset and chatbot.

Integrating User Feedback Loops: Feedback provided real-time by users to KnowledgeBot's responses would be highly appreciated. Basically, the feedback can help the chatbot to always perform at it's best. Developers can solve factual inaccuracy issues and improve the communication flow by accepting user input, resulting in natural and engaging conversations.

Deepening Medical Expertise: KnowledgeBot's potency in the medical field can be further enhanced by including medical knowledge graphs or particular medical dialogue datasets into its training process. This tailored training will endow it with an ability to provide complex medical answers with even greater precision and in-depth knowledge.

Expanding Applicability: The architecture behind KnowledgeBot holds the potential for more than just the healthcare sector as well. Researchers could consider whether it can be applied in various areas such as customer facing departments or education. Knowledge-base and training data could be adjusted for specific domains for KnowledgeBot to be helpful across diverse fields.

	Head	Tail	Recent	All
KnowledgeBot G4	1.2	29.0	28.0	14.7
KnowledgeBot G3.5	0.1	23.0	18.0	9.0
KnowledgeBot L	1.1	32.0	30.0	16.3

Table 4. Percentage of turns in which KnowledgeBot does not find relevant information in Wikipedia to retrieve or fact-check.

	Head	Tail	Recent
KnowledgeBot G4	86.1	68.0	65.7
KnowledgeBot G3.5	90.9	79.2	77.4
KnowledgeBot L	79.1	64.1	62.3

Table 5. Analysis of KnowledgeBot's response refinement. BLEU score with the refined response as the prediction and the response before refinement as the target..

9 FUTURE WORK

The possibility of the KnowledgeBot project set provides direct impetus for ongoing research.

Cross-lingual fluency: In the future, enhanced language switching implementation will become the key area of investigation. This could give way to creation of innovative real-time interpretation techniques or design of multilingual LLM architectures that simultaneously process multiple languages.

Real-time fact-checking: At a research level, integration of real-time fact-checking is one of the areas that has not yet been explored in KnowledgeBot It would consist of frequent checking of the information within interactions and perhaps, updating responses when better results became available.

User adaptation and personalization: KnowledgeBot could be developed to respond to the users'style of communication and preferences from one version to another. To do this, it could be possible to customize the content of the responses according to the user's history, hobbies and previous dialogue.

Explainable AI integration: Helping KnowledgeBot to implement Explainable AI (XAI) methodologies will handle the need for the user to understand its answers. The consideration will assist to start building trust and confidence in the chatbot system decision scheme.

Open-domain dialogue capabilities: At present KnowledgeBot exclusively works for particular areas. Going forward, an attempt is made to offer extensions with the capability to respond to general discussions on different issues. This, in turn, would be conditioned by enhancing the natural language understanding, and negotiating different conversational backdrops.

Through working on language modules, knowledge-base augmentation, multilingual support, and personal interactions, KnowledgeBot will become a reliable and multifaceted tool that is useful in improving language competence and in multiple applications.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to our supervisor, Professor Zunnurain Hussain, for his guidance and support throughout this project. Their expertise in IT, Artificial Intelligence, Networking and Security was invaluable in helping us navigate the complexities of this research. We are particularly grateful for his insightful feedback during research phase, and assistance in finding relevant resources.

REFERENCES

[1] Banerjee, D., Singh, P., Avadhanam, A. & Srivastava, S. Benchmarking llm powered chatbots: Methods and metrics. (2023). 2308.04624.

[2] Hao, R. *et al.* Chatllm network: More brains, more intelligence. (2023). 2304.12998.

[3] He, X. *et al.* Annollm: Making large language models to be better crowdsourced annotators. *arXiv.org* (2023). URL https://arxiv.org/abs/2303.16854.

[4] Zheng, L. *et al.* Lmsys-chat-1m: A large-scale real-world llm conversation dataset. (2023). 2309.11998.

[5] Peng, B. *et al.* Check your facts and try again: Improving large language models with external knowledge and automated feedback. (2023). 2302.12813.

[6] Authors. WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on wikipedia. *Pap. with Code* (2023). URL https://paperswithcode.com/paper/wikichat-a-few-shot-llm-based-chatbot.

[7] Gupta, P., Wu, C., Liu, W. & Xiong, C. Dialfact: a benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2022). URL https://doi.org/10.18653/v1/2022.acl-long.263.

[8] et al., S. S. Factual and engaging open-domain chatbot using a 7-stage pipeline with prompted llms. *arXiv* (2023). URL https://arxiv.org/abs/2305.14292. 2305.14292.

[9] et al., S. S. Mitigating hallucination in chatbots. *OpenReview* (2023). URL https://openreview.net/forum?id=sdC55K8cP0.

[10] Jalali, N. & Chen, H. Federated learning: a paradigm shift in machine learning. *Res. Sq.* (2024). URL https://www.researchsquare.

com/article/rs-3862540/latest.rs-3862540.

[11] et al., S. D. Improving chatbot efficiency with federated learning. *IEEE Xplore* (2023). URL https://ieeexplore.ieee.org/abstract/document/10303313/.

[12] et al., D. T. Multilingual chatbots for inclusive digital interactions. *DergiPark* (2021). URL https://dergipark.org.tr/en/download/article-file/2201572.

[13] Permatasari, D. A. & Maharani, D. A. Combination of natural language understanding and reinforcement learning for booking bot. *J. Electr. Electron. Inf. Commun. Technol. (Online)* **3**, 12 (2021). DOI 10.20961/jeeict.3.1.49818.

[14] Gao, Y. *et al.* Chat-rec: towards interactive and explainable llms-augmented recommender system. (2023). 2303.14524.

[15] Chen, S. *et al.* Llm-empowered chatbots for psychiatrist and patient simulation: application and evaluation. (2023). 2305.13614.

[16] et al., D. P. Revolutionizing user interactions with domain-specific chatbots. *MDPI* **13**, 320 (2024). URL https://www.mdpi.com/2079-9292/13/2/320.

[17] Smith, J. & Doe, J. Advancing chatbot capabilities through federated learning. *arXiv* **2310** (2023). URL https://arxiv.org/abs/2310.19303.2310.19303.

[18] Bang, Y. *et al.* A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv.org* (2023). URL https://arxiv.org/abs/2302.04023.

[19] Gao, L. *et al.* Rarr: Researching and revising what language models say, using language models. (2023). URL https://doi.org/10.18653/v1/2023.acl-long.910.

[20] Aslam, F. Advancements in chatbot intelligence: A review. *Eur. J. Technol.* (2023). URL https://ajpojournals.org/journals/index.php/EJT/article/view/1561.

[21] Suhel, S. *et al.* Advancements in chatbot intelligence. *IEEE* (2020). URL https://ieeexplore.ieee.org/abstract/document/9197825/.

[22] Caldarini, G., Jaf, S. & McGarry, K. A literature survey of recent advances in chatbots. *Inf. (Basel)* **13**, 41 (2022). DOI 10.3390/info13010041.

[23] Cherakara, N. *et al.* Furchat: An embodied conversational agent using llms, combining open and closed-domain dialogue with facial expressions. (2023). 2308.15214.

[24] Dong, W. *et al.* Federated learning: A comprehensive review. *arXiv preprint arXiv:2212.08354* (2022). URL https://arxiv.org/abs/2212.08354.

[25] Zhao, W. *et al.* Federated learning for chatbot intelligence. *arXiv preprint arXiv:2403.04529* (2024). URL https://arxiv.org/abs/2403.04529.

[26] Song, Y. *et al.* Privacy-preserving federated learning for chatbots. *arXiv preprint arXiv:2402.16515* (2024). URL https://arxiv.org/abs/2402.16515.

[27] Hathurusinghe, R. *et al.* Creating noisy training sets using wikipedia biography pages. *arXiv preprint arXiv:2105.09198* (2021). URL https://arxiv.org/abs/2105.09198.

[28] Anki, P., Bustamam, A., Al-Ash, H. S. & Sarwinda, D. Intelligent chatbot adapted from question and answer system using rnn-lstm model. *J. Physics: Conf. Ser.* **1844**, 012001 (2021). DOI 10.1088/1742-6596/1844/1/012001.

[29] Pantano, E. & Pizzi, G. Theoretical advancement of conversational agents. *J. Retail. Consumer Serv.* (2020). URL https://www.sciencedirect.com/science/article/pii/S0969698919311865.

[30] Aqil, A. N. *et al.* Robot chat system (chatbot) to help users "homelab" based in deep learning. *Int. J. Adv. Comput. Sci. Appl.* **12** (2021). DOI 10.14569/ijacsa.2021.0120870.

[31] Das, B. *et al.* Security and privacy challenges in federated learning for chatbots. *arXiv preprint arXiv:2402.00888* (2024). URL https://arxiv.org/abs/2402.00888.

[32] Xu, C., Guo, D., Duan, N. & McAuley, J. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. (2023). 2304.01196.

[33] Kuang, C. Research on the design of intelligent chatbot based on deep learning. *J. Physics: Conf. Ser.* **2170**, 012017 (2022). DOI 10.1088/1742-6596/2170/1/012017.

[34] Kocaballi, A. B. Conversational ai-powered design: Chatgpt as designer, user, and product. (2023). 2302.07406.

[35] A conversation-driven approach for chatbot management (2022). URL https://ieeexplore.ieee.org/document/9681834.

[36] Bilquise, G., Ibrahim, S. & Shaalan, K. Bilingual ai-driven chatbot for academic advising. *Int. J. Adv. Comput. Sci. Appl.* **13** (2022). DOI 10.14569/ijacsa.2022.0130808.

[37] Data decentralisation of llm-based chatbot systems in chronic disease self-management (2023).

URL https://doi.org/10.1145/3582515.3609536.

[38] Guan, Y. *et al.* Intelligent virtual assistants with llm-based process automation. (2023). 2312.06677.

[39] Hussain, S., Al-Hashmi, S. H., Malik, M. H. & Kazmi, S. I. A. Chatbot in e-learning. *SHS Web Conf.* **156**, 01002 (2023). DOI 10.1051/shsconf/202315601002.

[40] Wei, J., Kim, S., Jung, H. & Kim, Y. Leveraging large language models to power chatbots for collecting user self-reported data. (2023). 2301.05843.

[41] Tyen, G., Brenchley, M., Caines, A. & Buttery, P. Towards an open-domain chatbot for language practice. (2022). DOI 10.18653/v1/2022.bea-1.28.

[42] Pandya, K. & Holia, M. Automating customer service using langchain: Building custom open-source gpt chatbot for organizations. (2023). 2310.05421.

[43] Alruqi, T. N. & Alzahrani, S. M. Evaluation of an arabic chatbot based on extractive question-answering transfer learning and language transformers. *AI (Basel)* **4**, 667–691 (2023). DOI 10.3390/ai4030035.

[44] Raamkumar, A. S. & Yang, Y. Empathetic conversational systems: a review of current advances, gaps, and opportunities. *IEEE Transactions on Affect. Comput.* **14**, 2722–2739 (2023). DOI 10.1109/taffc.2022.3226693.

[45] Qin, W. *et al.* Read, diagnose and chat: towards explainable and interactive llms-augmented depression detection in social media. (2023). 2305.05138.

[46] Chen, J. *et al.* Learning agent-based modeling with llm companions: Experiences of novices and experts using chatgpt netlogo chat. (2024). arXiv:3613904.3642377.

[47] Zhang, X., Peng, B., Li, K., Zhou, J. & Meng, H. Sgp-tod: Building task bots effortlessly via schema-guided llm prompting. (2023). 2305.09067.

[48] Perez, E. *et al.* Red teaming language models with language models. (2022). 2202.03286.

[49] Abu-Rasheed, H., Abdulsalam, M. H., Weber, C. & Fathi, M. Supporting student decisions on learning recommendations: An llm-based chatbot with knowledge graph contextualization for conversational explainability and mentoring. (2024). 2401.08517.

[50] Song, Y. *et al.* Federated learning with llms. *arXiv preprint* (2024). URL https://arxiv.org/abs/2402.16515. 2402.16515.

[51] Das, B. *et al.* Challenges and solutions with llms. *arXiv preprint* (2024). URL https://arxiv.org/abs/2402.00888. 2402.00888.

[52] Hathurusinghe, R. *et al.* Wikipedia as a corpus for llm training. *arXiv preprint* (2021). URL https://arxiv.org/abs/2105.09198. 2105.09198.

[53] Komeili, M. *et al.* Crowdsourcing evaluation of open-domain dialogue systems. *arXiv preprint arXiv:2202.11210* (2022).

[54] Papers with code - wikiqa dataset. https://paperswithcode.com/dataset/wikiqa.

[55] Papers with code - frenchmedmcqa dataset. https://paperswithcode.com/dataset/frenchmedmcqa.

[56] Finch, S. E., Paek, E. & Choi, J. D. Leveraging large language models for automated dialogue analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing and SIGDIAL* (2023). URL https://doi.org/10.18653/v1/2023.sigdial-1.20.