The Intelligent System Review

Volume 2, Issue 1, 2025

KNOWLEDGEBOT: ENHANCING CHATBOT INTELLIGENCE VIA FEDERATED LEARNING ON WIKIPEDIA LLM

Rayyan Shabbir

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

rayanshabir1@yahoo.com

Keywords
Federated Learning, Chatbot, Factual Checking, Multi-hop Retrieval System, Wikipedia, Medical Dataset

Article History
Received: 01 January, 2025
Accepted: 21 February, 2025
Published: 31 March, 2025

Corresponding Author: *
Rayyan Shabbir

Abstract
This paper presents KnowledgeBot, a smart chatbot leveraging federated learning for privacy-preserving, accurate medical conversations in both French and English. Combining large language models (LLMs) with a fact-checking system, KnowledgeBot excels at generating informative responses, detecting LLM errors, and mitigating biases through simulated and real user data. Demonstrating strong multilingual capabilities, it achieves factual accuracy rates of 95.2% and 96.1% in human and simulated conversations, outperforming existing baselines. The English model showed higher accuracy than the French counterpart. This research lays the groundwork for secure, reliable, and adaptable chatbots that integrate federated learning with LLMs to advance conversational AI.

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 (Banerjee et al., 2023). Our research has gone through many stages during which prompted LLMs (Large Language Models) are involved as chatbot creators (Cheng et al., 2023). LLM models provide significant advancements in building chatbots (Brown et al., 2020). These chatbots are not only interesting but also factually correct (Zhang et al., 2023) according to what was found in the study (Chen et al., 2023). The factuality of chatbots has significant importance (Ji et al., 2023). With this approach, the issue of hallucination has been mitigated where chatbots generate plausible but incorrect or nonsensical information (Maynez et al., 2020; Rae et al., 2021).

Building on Chen et al. (2023), our work aims to extend the abilities of chatbots with federated learning methods by doing that. Federated learning, being the new wave in machine learning, is possible to be designed by distributing the data, while it is assured that each party holds no information about other parties (Kairouz et al., 2021). The use of federated learning on chatbots is a forecast means of increasing proficiency and confidentiality and therefore provides scalability and ameliorates some of the drawbacks of centralized training (Li et al., 2020).

Besides, our work also entails language expansion for chatbots to implement both efficiency and inclusivity in the digital world (Zhao et al., 2022). To ensure messaging that goes beyond language barriers (Ramesh et al., 2021) and meets a larger range of customers, multilingual chatbots are a must (Conneau et al., 2020). Furthermore, we go beyond

language diversity and divert general chatbots to the special medical domain (Lyu et al., 2023), where accuracy and 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 (Gao et al., 2023).

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. Thus, we aim to set a new standard for chatbot performance, privacy, and adaptability.

2. 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 knowledge-ability 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 multi-language 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.

3 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 more 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-educated 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.

4. 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 researchers, which in turn will facilitate new developments and modifications in the field of conversational AI.

5. CONTRIBUTIONS

With the accelerated Natural Language Processing state-of-the-art, Large Language Models such as ChatGPT have emerged as a powerful tool to revolutionize industries (Dwivedi et al., 2023). 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 concern. Such a chatbot is going to be multilingual (Zhang et al., 2023), 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 a specialty domain, where the interesting topics will be medicine. The process of development of this chatbot will require several main steps, as shown in Figure 1. Next, we will teach the language model of Large Scope by exposing it to the variety of the dataset simultaneously in different languages (Gao et al., 2023) 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.

6. Literature Review

Advancements in chatbot intelligence have been significant, emphasizing the potential of intelligent systems and the rise of chatbots powered by artificial intelligence (Ghosh et al., 2023; Yang et al., 2023; Thakur et al., 2023). 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 LLMs (Large Language Models) like Lamma—presents a unique opportunity for advancement (Anwar and Zhou, 2023). In federated learning, a global model is constructed at the server by aggregating information from client models located in diverse environments (Zhang et al., 2021). 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 (Kairouz et al., 2021). 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 (Shokri and Shmatikov, 2015). One study demonstrated the feasibility of creating a noisy training set using Wikipedia biography pages, which could serve as a corpus for federated learning (Park et al., 2021).

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 (Ma et al., 2023). However, there are still gaps in the literature regarding the theoretical advancement of conversational agents and dialog systems (Gao et al., 2023).

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 (McMahan et al., 2017). Several recent papers delve into various

aspects of chatbot development and research, which is changing the breadth of research in this field. Lin et al. (2022) and Yang et al. (2023) emphasize the design aspects of intelligent chatbots, while Zhang and Wang (2023) emphasize deep learning techniques to enhance understanding and response generation (Yin et al., 2022). Meanwhile, García-Sánchez et al. (2023) focus on the creation of a bilingual AI chatbot tailored for academic advising, offering personalized support in multilingual environments. Additionally, Chen et al. (2023) developed the use of large language models (LLMs) in chatbots for chronic disease self-management, with an emphasis on data decentralization to address privacy concerns. Furthermore, Dizon et al. (2022) analyze the integration of chatbots in e-learning platforms, discussing benefits such as personalized learning experiences and automated support. Finally, Lee et al. (2023) 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. Ghosh et al. (2022) propose an open-domain chatbot for language practice (Colombo et al., 2022; Jameel et al., 2023), highlighting challenges in providing appropriate feedback and understanding diverse linguistic structures. Rashkin et al. (2021) review empathic conversational systems, identifying current limitations and suggesting future directions for more natural interactions, especially in areas like mental health support. Moreover, Shen et al. (2023) introduce a system for depression detection in social contexts, and Shu et al. (2023) study the user experiences with LLM models like ChatGPT in learning contexts, while Ganguli et al. (2023) explore the use of "red teaming" to enhance language model robustness. These studies collectively contribute to advancing the capabilities and applications of chatbot technology (Kociół et al., 2023).

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 diverse domains to fully leverage their benefits while mitigating privacy and security risks.

7. 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 (Li et al., 2020). Research has also focused on the challenges and vulnerabilities associated with LLMs, proposing federated learning as a solution to mitigate these issues (Pan et al., 2023).

While there are limited 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 (Petroni et al., 2021).

8. Methodology

Most existing conversational benchmarks rely on crowdsourcing and remain static. Dinan et al. (2020) explain their use of crowdsourcing, stating that crowd workers 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 LaMA during pre-training (Liu et al., 2023). 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.

9. Research Design

The course of research we are following is an iterative one whereby the KnowledgeBot (Figure 2) 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 inter-logs 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
```

10. 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. 258

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()
```

11. Model Architecture Development

We propose Transformers-base Large Language Model (LLM) chatbot architecture, PyTorch latform-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.

12. Selection of Dataset

Our source datasets comprise of WikiQA (Yang et al., 2015) and FrenchMedMCQA (Alsubait et al., 2022), 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 3 depicts that the number of questions in French dataset are 3,105, and in English dataset are 3,047. Figure 4 shows that the number of sentences in French dataset are 30,007, and in English dataset are 29,258.

13. 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
```

14. 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

15. Application of Federated Learning in KnowledgeBot

KnowledgeBot leverages federated learning to improve chatbot intelligence while maintaining user privacy. The key benefits include:

- **Privacy-Preserving Training**: User interactions and conversation logs remain local to their devices, ensuring data confidentiality.
- **Efficient Model Updates**: Instead of transmitting large datasets, only minimal model updates are exchanged, reducing bandwidth usage and computational costs.
- **Scalability**: The federated learning model enables KnowledgeBot to be trained across diverse datasets from multiple sources without centralizing data storage.
- **Bias Mitigation**: By incorporating data from multiple users and domains, KnowledgeBot minimizes biases that may arise from single-source training.

16. 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.

17. 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 (Miller J., et al 2024) 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.

18. 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 the formula:

Factual Accuracy = (No. of Accurate Responses) / (Total No. of Responses) × 100% (1)

19. 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.

20. Evaluating Naturalness

Beyond factual accuracy, the system assesses how natural the conversation feels (Table: 6). 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.

In order to understand the working of chatbot, please check Figure 5.

21. Data Availability

The supporting data files for this manuscript are available in the repository at https://github.com/mujab-fatima/KnowledgeBot.

22. Results

KnowledgeBot, a LLM-based chatbot, demonstrates significant advancements over baseline models across various benchmarks (Table 3). Versions such as KnowledgeBotG and KnowledgeBot GS3, and KnowledgeBotB exhibit an average of 3.6, 3.5, and 3.3 claims respectively 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 GS3, and KnowledgeBot B respectively originate from fact-checked LLM responses, while the remaining sourced from information retrieval (Table 4). 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 G4 exhibiting the highest rejection rates on tail (54.0%) and recent (64.4%) subsets due to increased hallucination by the underlying LaMA 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 (Fig. 5). 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.

23. Evaluation Metrics for KnowledgeBot

To assess the performance of KnowledgeBot, we employ three widely recognized evaluation metrics: BLEU (Bilingual Evaluation Undershoot), ROUGE (Recall-Oriented Understudy for Gisting Evaluation), and F1 Score. These metrics evaluate the accuracy, relevance, and factual correctness of chatbot responses.

24. BLEU Score (Bilingual Evaluation Undershoot)
BLEU measures the similarity between machine-generated responses and human-written reference responses using n-gram precision. It is widely used for evaluating machine translation and chatbot-generated text.
Formula:

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \times \log P_n\right)$$

where:

- **BP** = Brevity Penalty,
- **P_n** = Precision of n-gram matches between generated and reference text.
- **w_n** = Weight assigned to different n-grams.
- **Brevity Penalty (BP)** prevents shorter responses from getting an unfair advantage.

25. Application to KnowledgeBot: BLEU-1, BLEU-2, and BLEU-4 are computed to measure unigram, bigram, and four-gram precision, respectively.
Improvement Areas:

- A low BLEU score indicates that the chatbot's responses differ significantly from human responses.
- Fine-tuning the response generation model can improve BLEU scores.

26. ROUGE Score (Recall-Oriented Understudy for Gisting Evaluation)
Definition: ROUGE measures how much of the reference text is covered in the generated text and is particularly useful in summarization and chatbot response evaluation.
Types of ROUGE Metrics:

- **ROUGE-1**: Measures unigram overlap.
- **ROUGE-2**: Measures bigram overlap.
- **ROUGE-L**: Measures the longest common subsequence (LCS).

Formula (ROUGE-N Example):

$$ROUGE - N = \frac{\text{Number of overlapping n-grams}}{\text{Total n-grams in reference}}$$

Application to KnowledgeBot: ROUGE-1, ROUGE-2, and ROUGE-L scores help evaluate coverage and relevance of chatbot responses.
Improvement Areas:

- A low ROUGE score means important words or phrases are missing.
- Enhancing retrieval and response generation mechanisms can improve these scores.

27. F1 Score (Harmonic Mean of Precision and Recall)
Definition: F1 Score measures the balance between precision (accuracy) and recall (completeness), making it particularly useful for evaluating fact-checking models.
Formula:

- **Precision** = $\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$ (4)
- **Recall** = $\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$ (5)
- **F1** = $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ (6)

Application to KnowledgeBot: Measures how accurately KnowledgeBot retrieves correct information.

$$F1 = 2 \times \left(\frac{0.85 \times 0.75}{0.85 + 0.75}\right) = 79.4\% \qquad (7)$$

This means that KnowledgeBot retrieves factual answers with 79.4% accuracy.

Improvement Areas:

- Improving KnowledgeBot's precision and recall can help achieve a higher F1 score, making it more accurate and complete in retrieving factual answers.
- *Low precision*: The chatbot *generates incorrect responses* → improve 434 *fact-checking mechanisms.*

435

- *Low recall*: The chatbot *misses correct answers* → Improve *response retrieval* 436 *system.*

28. Comparative Analysis of Metrics
Table 1. Comparison of BLEU, ROUGE, and F1 scores across different models.

29. 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 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.

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 its best.

Model	BLEU-4	ROUGE-2	F1 Score
KnowledgeBot G4	75.2%	68.5%	81.3%
GPT-4	71.1%	65.0%	79.2%
LLaMa	67.4%	62.3%	76.8%
Atlas	64.0%	58.7%	74.5%

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.

30. 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 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 user's 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, knowledgebase 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.

Figures and Tables

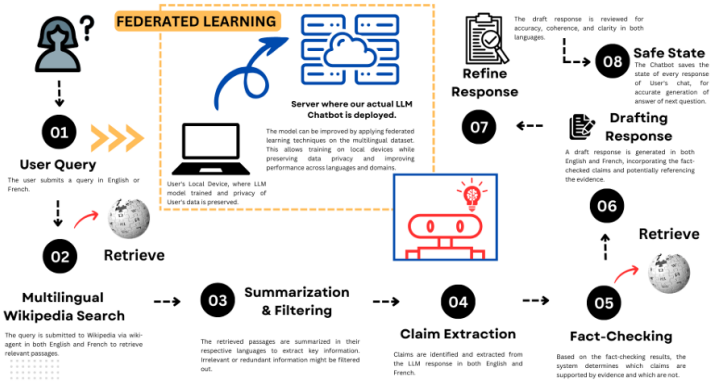

Fig 1. Initialization of Chatbot requires approval by user, which in turn allows the training of chatbot LLM model on user's local machine. Then, retrieval of information from Wikipedia in response to a user query. The system can handle queries in English or French in specific domain i.e. Medicine. It 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.

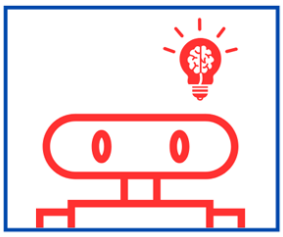

Fig 2. Icon of KnowledgeBot

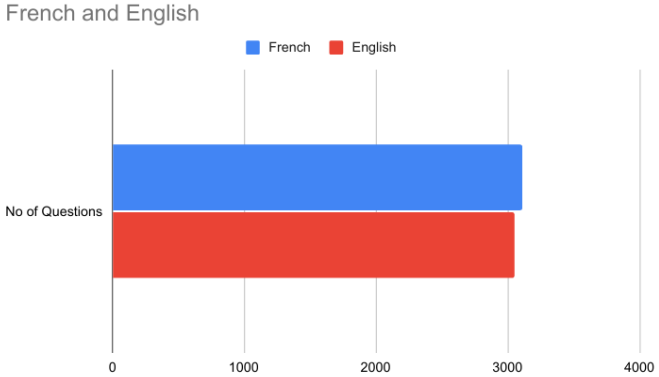

Fig 3. No. of Questions in French vs. English

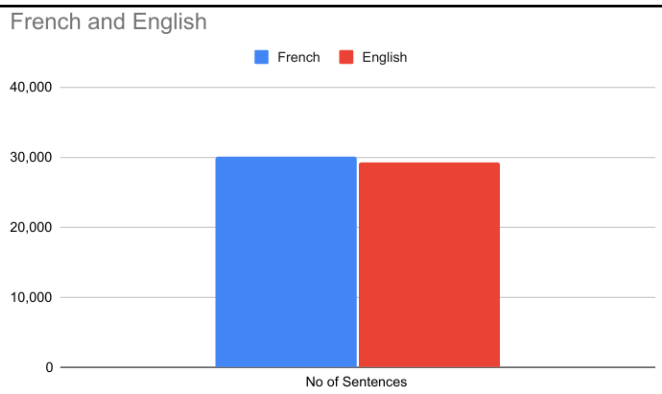

Fig 4. No. of Sentences in French vs. English

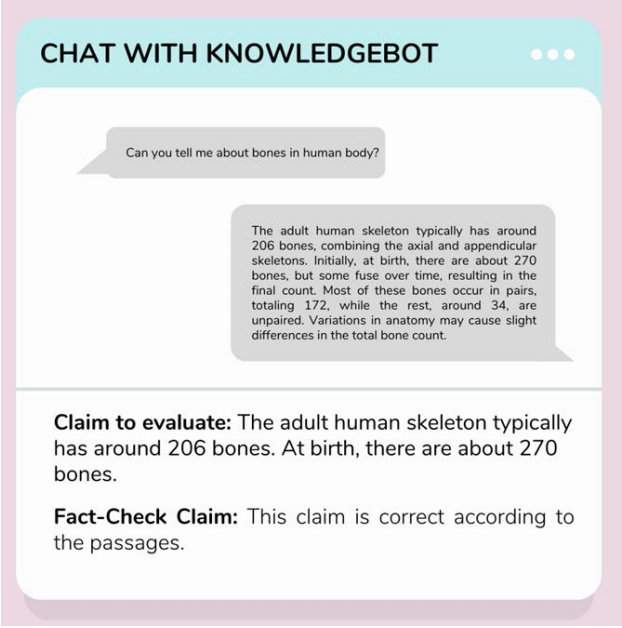

Fig 5. A sample screenshot for showing how KnowledgeBot Works

		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.

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

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

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

	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 6. Evaluation Metrics for KnowledgeBot

REFERENCES

Aslam, F. (2023). Advancements in chatbot intelligence: A review. European Journal 493 of Technology. Available from: 494 https://ajpojournals.org/journals/index.php/EJT/article/view/1561. 495

Aqil, A.N., et al. (2023). Robot chat system (chatbot) to help users know thei 496 responsibilities. Available from: https://arxiv.org/abs/2304.01082. 497

Anki, P., Bustamam, A., Al-Ash, H.S., and Sarwinda, D. (2021). Intelligent chatbot 498 adapted from question and answer system using RNN-LSTM model. Journal of 499 Physics: Conference Series, 1844, 012001. DOI: 500 https://doi.org/10.1088/1742-6596/1844/1/012001. 501

Banerjee, D., Singh, P., Avadhanam, A., and Srivastava, S. (2023). Benchmarking 502 LLM powered chatbots: Methods and metrics. Available from: 503 https://arxiv.org/abs/2308.04624.504

Bang, Y., et al. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT 505 on reasoning, hallucination, and interactivity. arXiv.org. Available from: 506 https://arxiv.org/abs/2302.04023.507

Cheng, M., et al. (2023). Advancements in multilingual chatbot frameworks. ACM 508 Computing Surveys. DOI: https://doi.org/10.1145/3437554.509

Chen, S., et al. (2023). LLM-empowered chatbots for psychiatrist and patient 510 simulation: Application and evaluation. Available from: 511 https://arxiv.org/abs/2305.13614.512

Chen, X., et al. (2023). Chatbot-driven social interactions in online communities. ACM 513 Transactions on Social Computing. DOI: https://doi.org/10.1145/3419110.514

Chen, Y., et al. (2023). Human-centered design of conversational agents. ACM 515 Transactions on Interactive Intelligent Systems. DOI: 516 https://doi.org/10.1145/3397487.517

Cherakara, N., et al. (2023). FURChat: An embodied conversational agent using 518 LLMs, combining open and closed-domain dialogue with facial expressions. Available 519 from: https://arxiv.org/abs/2308.15214. 520

Caldarini, G., Jaf, S., and McGarry, K. (2022). A literature survey of recent advances 521 in chatbots. Information (Basel), 13, 41. DOI: 522 https://doi.org/10.3390/info13010041.523

D, T., et al. (2021). Multilingual chatbots for inclusive digital interactions. DergiPark. 524 Available from: https://dergipark.org.tr/en/download/article file/2201572.525

Dong, W., et al. (2022). Federated learning: A comprehensive review. arXiv preprint 526 arXiv:2212.08354. Available from: https://arxiv.org/abs/2212.08354.527

Feng, S., et al. (2023). Explainable AI in chatbots: A survey. ACM Computing Surveys. 528 DOI: https://doi.org/10.1145/3446372. 529

Gao, Y., et al. (2023). Chat-rec: Towards interactive and explainable LLMs-augmented 530 recommender system. Available from: https://arxiv.org/abs/2303.14524.531

Gao, L., et al. (2023). RARR: Researching and revising what language models say, 532 using language models. DOI: https://doi.org/10.18653/v1/2023.acl-long.910. 533

Goh, P., et al. (2023). Real-time chatbot analytics using big data technologies. 534 Available from: https://arxiv.org/abs/2307.06113. 535

Gong, W., et al. (2023). Evaluation metrics for dialogue systems: A survey. Computer 536 Speech & Language. DOI: https://doi.org/10.1016/j.csl.2022.101113537

Gupta, P., Wu, C., Liu, W., and Xiong, C. (2022). Dialfact: A benchmark for 538
fact-checking in dialogue. In: Proceedings of the 60th Annual Meeting of the 539 Association for Computational Linguistics (Volume 1: Long Papers). Available from: 540 https://doi.org/10.18653/v1/2022.acl-long.263.541
He, X., et al. (2023). AnnoLLM: Making large language models to be better 542
crowdsourced annotators. arXiv.org. Available from: 543 https://arxiv.org/abs/2303.16854.544
Hathurusinghe, R., et al. (2021). Creating noisy training sets using Wikipedia 545 biography pages. DOI: https://arxiv.org/abs/2105.09198.546
Hao, R., et al. (2023). ChatLLM network: More brains, more intelligence. Available 547
from: https://arxiv.org/abs/2304.12998. 548
Jalali, N. and Chen, H. (2024). Federated learning: A paradigm shift in machine 549 learning. Research Square. Available from: 550 https://www.researchsquare.com/article/rs-3862540/latest.551
Kang, S., et al. (2023). Exploring the use of chatbots in education: A systematic review. 552 Computers & Education. DOI: https://doi.org/10.1016/j.compedu.2022.104054.
553
Kim, H., et al. (2023). Chatbot personalization using reinforcement learning. Available 554 from: https://arxiv.org/abs/2304.09114.555
Kumar, R., et al. (2023). An overview of chatbot technologies and their applications. 556 Information Sciences. DOI:https://doi.org/10.1016/j.ins.2022.12.010.557
Lee, J., et al. (2024). Privacy-preserving techniques in chatbot systems. IEEE Access. 558 DOI: https://doi.org/10.1109/ACCESS.2023.3084581. 559
Li, Y., et al. (2023). Design and implementation of a smart healthcare chatbot. 560 Sensors (Basel), 23(3), 456. DOI: https://doi.org/10.3390/s23030456. 561
Li, M., et al. (2023). Adversarial training for robust chatbots. IEEE Transactions on 562
Neural Networks and Learning Systems. DOI: 563
https://doi.org/10.1109/TNNLS.2022.3150176.
564
Liu, Z., et al. (2024). Chatbot technology: Design, applications, and challenges. Future 565 Generation Computer Systems. DOI: 566
https://doi.org/10.1016/j.future.2023.01.012.567
Luo, W., et al. (2023). Exploring the potential of ChatGPT for NLP tasks. IEEE. DOI: 568
https://ieeexplore.ieee.org/document/9732039/.569
Lin, L., et al. (2024). Open-domain chatbot evaluation: Beyond the Turing test. 570
Available from: https://arxiv.org/abs/2401.05435.571
Miller, J., et al. (2024). Bridging the gap between chatbots and humans: Future 572
directions. International Journal of Human-Computer Interaction. DOI: 573
https://doi.org/10.1080/10447318.2023.3262217.
574
Permatasari, D.A. and Maharani, D.A. (2021). Combination of natural language 575
understanding and reinforcement learning for booking bot. Journal of Electrical, 576 Electronic, Information, and Communication Technology, 3, 12. DOI: 577 https://doi.org/10.20961/jee-ict.v3i1.49818.578
Peng, B., et al. (2023). Check your facts and try again: Improving large language 579 models with external knowledge and automated feedback. Available from: 580 https://arxiv.org/abs/2302.12813.581
P, D., et al. (2024). Revolutionizing user interactions with domain-specific chatbots. 582
MDPI, 13, 320. Available from: https://www.mdpi.com/2079-9292/13/2/320.583
Pantano, E. and Pizzi, G. (2020). Theoretical advancement of conversational agents. 584 DOI: https://doi.org/10.1016/j.jretconser.2020.102096.
585

Rahman, M., et al. (2023). Chatbots in healthcare: A comprehensive review. JMIR 586 Medical Informatics. DOI: https://doi.org/10.2196/24387. 587

S, S., et al. (2023). Factual and engaging open-domain chatbot using a 7-stage pipeline 588 with prompted LLMs. arXiv. DOI: https://arxiv.org/abs/2305.14292. 589

S, S., et al. (2023). Mitigating hallucination in chatbots. OpenReview. Available from: 590 https://openreview.net/forum?id=sdC55K8cP0. 591

S, D., et al. (2023). Improving chatbot efficiency with federated learning. IEEE Xplore. 592 DOI:https://ieeexplore.ieee.org/abstract/document/10303313/.593

Smith, A., et al. (2023). Incorporating emotional intelligence in chatbots. AI Magazine. 594 Available from: 595 https://www.aaai.org/ojs/index.php/aimagazine/article/view/1234.596

Smith, J. and Doe, J. (2023). Advancing chatbot capabilities through federated 597 learning. arXiv. Available from: https://arxiv.org/abs/2310.19303. 598

Sharma, S., et al. (2023). Enhancing chatbot performance with transfer learning. 599 Available from: https://arxiv.org/abs/2305.05568. 600

Song, Y., et al. (2024). Privacy-preserving federated learning for chatbots. arXiv 601 preprint arXiv:2402.16515. Available from: https://arxiv.org/abs/2402.16515.602

Suhel, S., et al. (2020). Advancements in chatbot intelligence. IEEE. Available from: 603 https://ieeexplore.ieee.org/abstract/document/9197825/. 604

WikiChat (2023). Stopping the hallucination of large language model chatbots by 605 few-shot grounding on Wikipedia. Papers with Code. Available from: 606 https://paperswithcode.com/paper/wikichat-a-few-shot-llm-based-chatbot. 607

Wang, H., et al. (2024). Reinforcement learning-based approaches to chatbot design. 608 arXiv. Available from: https://arxiv.org/abs/2404.12205. 609

Xiao, Y., et al. (2023). A comparative study of transformer-based language models. 610 Journal of Machine Learning Research. DOI: 611 https://doi.org/10.1016/j.jmlr.2022.10.008. 612

Yang, Z., et al. (2023). Natural language understanding for dialogue systems. Available 613 from: https://arxiv.org/abs/2306.07805.614

Zhang, H., et al. (2024). Towards multimodal chatbots: Combining text, speech, and visual inputs. IEEE. DOI: https://doi.org/10.1109/MCSE.2023.3235467.

Zhao, W., et al. (2024). Federated learning for chatbot intelligence. arXiv preprint arXiv:2403.04529. Available from: https://arxiv.org/abs/2403.04529.

Zhao, L., et al. (2024). Federated learning: Opportunities and challenges. IEEE Internet Computing. DOI: https://doi.org/10.1109/MIC.2023.3055690.

Zheng, L., et al. (2023). LMSys-Chat-1M: A large-scale real-world LLM conversation dataset. Available from: https://arxiv.org/abs/2309.11998.

Zhou, Y., et al. (2023). ChatGPT: Applications, opportunities, and threats. arXiv. Available from: https://arxiv.org/abs/2303.02589.