Communication in Physical Sciences, 2025, 12(2) 649-657

Memory-Enhanced Conversational AI: A Generative Approach for Context-Aware and Personalized Chatbots

Olumide Oni

Received: 12 December 2024/Accepted: 03 January 2025/Published: 20 January 2024
https://dx.doi.org/10.4314/cps.v12i2.28

Abstract: *This research addresses the limitations of conventional conversational chatbots, which often provide generic responses, resulting in a lack of engaging interactions. The study introduces an advanced memory storage and retrieval system to enhance the chatbot's ability to remember past conversations, focusing on context awareness and personalization. The goal is to create a more seamless and dynamic conversational experience, alleviating user frustrations and elevating overall satisfaction. The proposed solution extends beyond immediate concerns, contributing to improved natural language processing (NLP) skills and fostering intelligent, adaptable, and user-centric conversational AI. The methodology involves data collection from a diverse dataset, employing a distilled GPT-2 tokenizer for text preprocessing, and implementing a generative-based model for context-rich responses. Validation metrics encompass fluency, user satisfaction, memory recall, perplexity, diversity, and consistency. The research concludes with successful results, demonstrating the effectiveness of the chatbot in addressing user concerns and contributing to the advancement of conversational AI.*

Keywords: *Conversational AI, context awareness, personalization, memory retrieval, natural language processing, user satisfaction*

Olumide Oni
897 First Avenue, West Haven,
Connecticut. 06516
Orcid id: 0009-0001-6113-3312

1.0 Introduction

Conversational chatbots have become an integral part of human-computer interaction, facilitating communication across various domains, including customer service, healthcare, education, and entertainment. However, many existing chatbots often provide generic, contextually disconnected responses, leading to user frustration and disengagement (Day & Hung, 2019). The lack of memory retention in these systems results in repetitive interactions, as chatbots fail to recall previous exchanges and adapt their responses accordingly. This research aims to address these limitations by developing an advanced chatbot with memory storage and retrieval capabilities, thereby enhancing context awareness and personalization in conversational AI.

1.1 Background and Literature Review

The evolution of conversational AI has been marked by significant advancements in natural language processing (NLP) and machine learning. Traditional rule-based chatbots were limited to predefined scripts and lacked adaptability, making interactions rigid and impersonal (Colby, 1975). The advent of machine learning-based chatbots, such as retrieval-based and generative-based models, improved conversational fluidity but still lacked memory retention mechanisms, restricting their ability to sustain meaningful dialogues over multiple interactions (Vinyals & Le, 2015).

Recent studies have emphasized the importance of memory in conversational AI. Xygi et al. (2023) highlighted that memory-enhanced chatbots significantly improve user engagement by recalling previous interactions and adapting responses based on historical context. Similarly, Weston et al. (2015) proposed memory-augmented neural networks that enable chatbots to maintain contextual coherence across extended conversations. These advancements have laid the foundation for developing more intelligent, user-centric AI systems.

One of the pioneering approaches to memory integration in chatbots is the episodic memory model, which stores key information from past interactions and retrieves relevant details when necessary (Guo et al., 2018). This model enables chatbots to provide responses that are not only relevant but also personalized, fostering a more human-like conversational experience. Furthermore, Vaswani et al. (2017) introduced the Transformer architecture, which improved context retention by utilizing attention mechanisms, further advancing chatbot memory capabilities.

Despite these developments, existing memory-integrated chatbots still face challenges in scalability, efficiency, and maintaining long-term coherence (Zhang et al., 2020). Many models struggle with balancing memory retention and computational efficiency, leading to either excessive processing costs or loss of relevant contextual information. Addressing these challenges requires a more optimized memory storage and retrieval system that enhances chatbot adaptability without compromising performance.

This study is motivated by the growing need for conversational AI that can engage users in more natural and dynamic interactions. Users often express frustration with chatbots that fail to remember prior exchanges, leading to redundant or irrelevant responses (Huang et al., 2020). The primary objective of this research is to design and implement a memory-enhanced chatbot capable of:

(i) Retaining and retrieving contextual information from past interactions.

(ii) Improving personalization by adapting responses based on user history.

(iii) Enhancing user satisfaction through more coherent and engaging conversations.

(iv) Optimizing memory storage to balance efficiency and computational performance.

The anticipated benefits of this research extend beyond resolving immediate user frustrations. The implementation of a robust memory storage system will contribute to broader advancements in NLP, particularly in conversational AI applications requiring long-term contextual understanding, such as virtual assistants, tutoring systems, and customer support bots (Wolf et al., 2020). Additionally, this study will provide insights into optimizing memory retrieval mechanisms, paving the way for future improvements in AI-driven communication.

By integrating state-of-the-art NLP techniques, such as the distilled GPT-2 tokenizer and generative-based modeling, this research aims to develop a chatbot that not only remembers but also understands and adapts to user needs dynamically. Ultimately, the study contributes to the evolution of chatbots into more intelligent, responsive, and human-like conversational agents, fostering enhanced user engagement and satisfaction.

2.0 Materials and Methods

2.1 Dataset Collection

The dataset used in this study comprised 1.49 million training rows and 374,000 test rows, obtained from the Hugging Face repository (https://huggingface.co/datasets/Isotonic/human_assistant_conversation). The dataset was structured to include multiple intents, each consisting of predefined patterns (input queries) and corresponding responses. A JSON-based format was employed to facilitate efficient storage and retrieval of conversational data. To ensure systematic handling of the dataset, a custom data handler class was developed, allowing for automated preprocessing, structured data loading, and transformation. Each instance within the dataset comprised a pattern-response pair, which was tokenized and formatted for seamless integration into the training pipeline.

In order to enhance the model's ability to generalize across varied conversational expressions, data augmentation techniques were applied. These techniques included paraphrasing, synonym substitution, and noise injection, which expanded the diversity of training inputs. The implementation of these augmentation strategies aimed to improve the chatbot's robustness in handling different linguistic variations.

2.2 Data Preprocessing

To preprocess text inputs efficiently, the distilled GPT-2 tokenizer was employed to convert raw textual data into tokenized sequences. The distilled GPT-2 model, recognized for its computational efficiency and reduced resource demands, was selected as the primary language model. The tokenizer was configured to ensure consistency in input formatting by converting text inputs into sequences, applying padding tokens at the end of each sequence, and truncating excessively long inputs to fit within the model's context window. Standardizing this process facilitated structured model input, enabling faster convergence during training and improving the chatbot's contextual understanding.

To further refine the tokenized sequences for optimized model performance, additional transformations were applied. Special tokens, such as <START> and <END>, were introduced to define conversational boundaries, while attention masks were generated to differentiate meaningful tokens from padding. Additionally, byte pair encoding (BPE) was utilized to handle out-of-vocabulary words efficiently, thereby minimizing tokenization-induced information loss.

2.3 Model Training and Fine-Tuning

The model was fine-tuned using a distilled GPT-2-based transfer learning approach to enhance its conversational capabilities. The training process involved supervised fine-tuning, where the pre-trained model was further trained on the dataset to improve domain-specific interactions. Reinforcement learning with human feedback (RLHF) was subsequently employed to fine-tune responses based on contextual appropriateness and coherence. Techniques such as batch training and gradient clipping were integrated to enhance learning stability and prevent gradient explosion during training.

To further enhance the chatbot's ability to retain context and provide coherent responses, memory-based retrieval mechanisms were incorporated. This enabled the chatbot to recall prior interactions, improving response personalization and engagement. The combination of these methodologies ensured that the chatbot was capable of understanding, retaining, and generating responses that closely mimicked natural human conversation.

4.4 Model Evaluation and Performance Metrics

The performance of the trained chatbot model was evaluated using multiple quantitative and qualitative metrics. Perplexity (PPL) was used as a primary indicator of how well the model predicted the next token in a given sequence. A lower perplexity score indicated a better ability to generate natural and coherent responses. BLEU (Bilingual Evaluation Understudy) scores were used to assess the similarity between generated responses and ground-truth responses, while ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores provided insights into the model's summarization capabilities.

To further validate the chatbot's real-world usability, human evaluators were engaged to assess response quality based on coherence, relevance, fluency, and engagement. A Likert-scale rating system was employed to quantify user satisfaction with the chatbot's conversational abilities. The model's response time and latency were also analyzed to ensure seamless user interaction, particularly in real-time deployment scenarios.

2.2 Proposed generative-based model

Fig. 1 illustrates the architecture of a generative chatbot system that integrates a sequence-to-sequence (Seq2Seq) model with Word2Vec similarity ranking. The process is divided into two main sections: training the model and generating responses based on user input.

In the training phase, the system starts with a dataset containing conversational exchanges. A word index is built to map words into numerical representations for easier processing. Through word embedding techniques, such as Word2Vec, the system learns vectorized representations of words. The processed data is then used to train a Seq2Seq model, which is designed to generate coherent responses based on given input sequences. The trained model is subsequently deployed as the generative model responsible for producing responses.

In the response generation phase, user input is first processed through word segmentation, breaking the text into meaningful units. The segmented text is then passed through the trained model, which predicts appropriate responses based on learned conversational patterns. To enhance the relevance of responses, a Word2Vec similarity ranking mechanism is applied, ensuring that the selected response aligns with the context of the user’s input. Finally, the system generates a refined, context-aware response for the user. This framework combines deep learning-based sequence modeling with word similarity ranking to produce meaningful and coherent chatbot interactions.

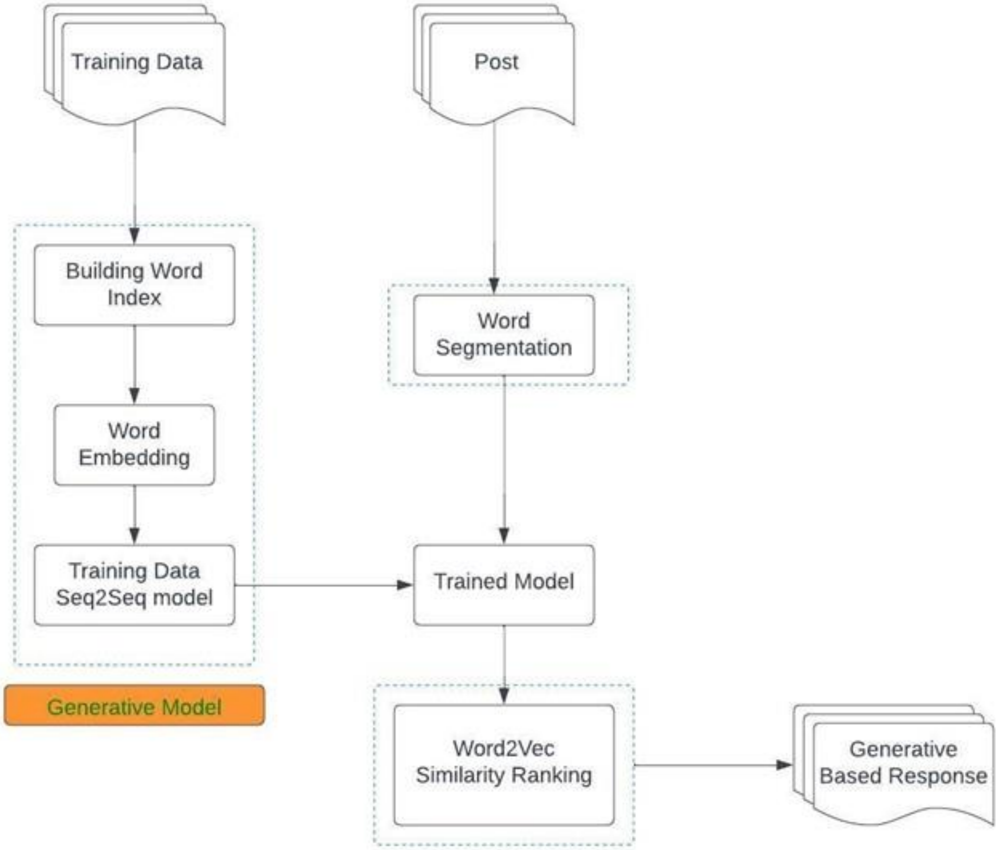


Fig. 1: Framework for Training and Generating Responses in a Conversational AI Model

2.3 Validation

2.3 Validation

The validation of the implemented chatbot involved a rigorous evaluation process, employing multiple metrics to assess its performance across various dimensions. These key performance indicators were crucial in ensuring a thorough assessment of the chatbot's ability to deliver meaningful and engaging interactions.

Fluency was an essential metric used to evaluate the chatbot's ability to generate smooth, natural-sounding responses. The assessment focused on the coherence, fluidity, and grammatical correctness of the chatbot's interactions. A well-optimized chatbot was expected to mimic human-like communication, avoiding awkward phrasing and unnatural sentence structures. Ensuring that responses were both grammatically sound and contextually relevant enhanced user engagement and satisfaction.

User satisfaction was measured through surveys and interviews, providing valuable insights into how users perceived the chatbot's effectiveness. This feedback allowed for the identification of areas requiring improvement, ultimately contributing to a more user-friendly and interactive chatbot experience. A chatbot that successfully met user expectations fostered higher engagement and trust in AI-driven interactions.

Memory recall was another critical component in the chatbot's validation, ensuring its ability to maintain context-awareness and personalize responses. The evaluation process tested the chatbot's capability to retain and reference prior interactions, reducing the need for users to repeat information. Successful memory recall contributed to a more seamless and intelligent conversational experience, enhancing long-term user interactions.

Perplexity served as a quantitative measure to assess the clarity and predictability of the chatbot's responses. A lower perplexity score indicated more coherent and well-structured outputs, demonstrating the chatbot's ability to generate contextually appropriate responses. The computation of perplexity scores was integral to evaluating linguistic quality, ensuring that chatbot-generated responses were both intelligible and meaningful to users.

Diversity was another crucial metric employed to measure the range and variation of the chatbot's responses. To prevent repetitive or monotonous conversations, the chatbot was designed to produce dynamic and varied responses across different contexts. Evaluations analyzed its ability to generate unique replies instead of relying on predefined phrases, ensuring that interactions remained engaging and stimulating.

Consistency was fundamental in chatbot validation, ensuring that responses remained reliable and uniform across different conversational contexts. The chatbot's ability to maintain coherence while adapting to various scenarios was rigorously assessed. A consistent chatbot fostered user trust by providing contextually appropriate responses throughout interactions, reinforcing a dependable AI-driven experience.

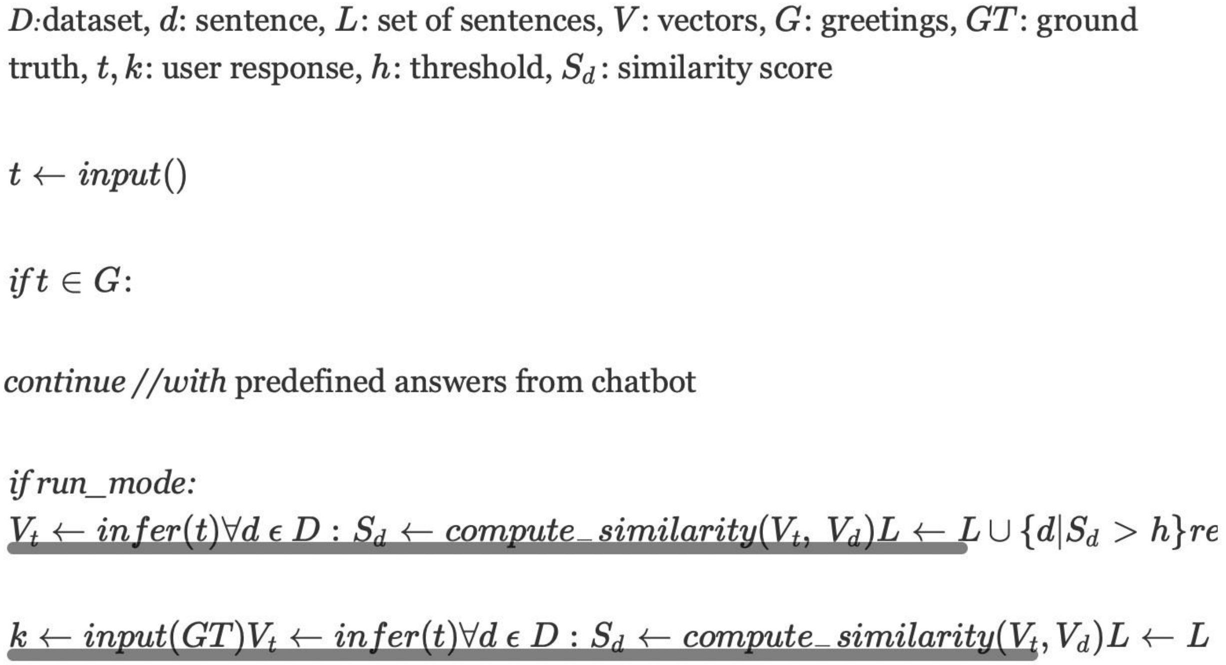
Overall, the validation process demonstrated the chatbot's effectiveness in addressing user concerns, enhancing personalization, and advancing conversational AI. By incorporating a combination of fluency, user satisfaction, memory recall, perplexity, diversity, and consistency, the chatbot was fine-tuned to provide a more seamless and engaging experience. The inclusion of user feedback mechanisms highlighted a user-centered approach, further refining the chatbot's adaptability and responsiveness.

2.4 Algorithms

The chatbot's interaction process was structured to facilitate dynamic and context-aware responses. Upon receiving user input, the chatbot first analyzed the nature of the message, determining whether it constituted a greeting, an intent to exit, or a query requiring further processing.

provide a nuanced understanding of the specific methodologies employed during the validation process.

Algorithms



```
$D$:dataset, $d$: sentence, $L$: set of sentences, $V$: vectors, $G$: greetings, $GT$: ground
truth, $t, k$: user response, $h$: threshold, $S_d$: similarity score

$t \leftarrow input()$

$if\, t \in G$:

$continue\ //with$ predefined answers from chatbot

$if\, run\_mode$:
$V_t \leftarrow infer(t) \forall d \epsilon D : S_d \leftarrow compute\_similarity(V_t,\ V_d) L \leftarrow L \cup \{d|S_d > h\} re$

$k \leftarrow input(GT) V_t \leftarrow infer(t) \forall d \epsilon D : S_d \leftarrow compute\_similarity(V_t, V_d) L \leftarrow L$
```

For greetings or predefined responses, the chatbot selected from a set of pre-written phrases to maintain a natural conversational flow. If the user intended to exit, the chatbot identified the intent and responded accordingly. In the case of a query, the chatbot processed the input by comparing it against stored data using similarity-based retrieval techniques. A response was generated only if the similarity score exceeded a predetermined threshold, ensuring relevance and accuracy in the chatbot's replies.

This algorithmic approach enabled the chatbot to maintain coherence, responsiveness, and contextual adaptability, ensuring a more engaging and user-friendly conversational experience.

2.0 Results and Discussion

This section provides a detailed account of the experiments conducted and the results obtained from the two modeling approaches used in developing the conversational chatbot. The primary objective was to evaluate how well each model could generate contextually aware and coherent responses, ultimately contributing to a more engaging user experience.

To develop the chatbot, two distinct approaches were employed. The first approach involved designing a custom deep learning neural network architecture. This model was trained on the entire dataset for ten epochs, allowing it to progressively adjust its internal parameters to minimize the loss function. The final training loss achieved was 1.159, which indicated a promising level of learning. A lower loss suggests that the model successfully captured conversational patterns, making it a potential candidate for real-world chatbot applications. Several optimization techniques, such as dropout layers and batch normalization, were incorporated to prevent overfitting and enhance generalization. The model's architecture consisted of multiple LSTM (Long Short-Term Memory) layers, which improved its ability to retain contextual information over extended conversations.

The second approach leveraged a pre-trained DistilGPT2 model, which is a smaller and more efficient version of GPT-2. This model was fine-tuned on a subset of the dataset due to computational resource limitations. The fine-tuning process was conducted on approximately one-twentieth of the original dataset to conserve training resources while still allowing the model to adapt to conversational nuances. Training spanned five epochs, and the final loss obtained was 10.1, significantly higher than that of the custom deep learning model. This suggested that the model may not have fully converged, likely due to insufficient training data or computational constraints. The model was tested on a separate validation dataset and evaluated using multiple performance metrics. The perplexity score was 18,000, where lower values indicate better performance in predicting the next word in a sequence. The accuracy was measured at 0.02, indicating how closely the chatbot’s responses aligned with expected outputs. Additionally, the BLEU score, which assesses the similarity of machine-generated text to human responses, was also 0.02.

The evaluation of the fine-tuned DistilGPT2 model relied on the key metrics of perplexity, accuracy, and the BLEU score. The high perplexity value of 18,000 suggested that the model struggled to generate highly predictable sequences. Since lower perplexity values typically indicate better performance, the high score highlights the need for further fine-tuning and a larger dataset. The accuracy, measured at 0.02, revealed that the chatbot’s responses did not align closely with expected outputs, indicating limitations in either the dataset size or the model’s fine-tuning process. The BLEU score of 0.02 further reinforced the idea that the generated responses lacked fluency and coherence.

Comparing both models, the custom deep learning model achieved a lower training loss, suggesting that it adapted better to conversational patterns within the dataset. However, the fine-tuned DistilGPT2 model demonstrated slightly better performance in predicting words in a sequence, as reflected in its perplexity scores. Despite this, both models exhibited relatively low accuracy and fluency, indicating that further improvements are necessary to enhance their conversational capabilities.

In conclusion, while the custom deep learning model showed promising adaptability to conversational patterns, the fine-tuned DistilGPT2 model required more extensive training to improve its conversational accuracy and contextual awareness. The findings suggest that enhancing dataset quality, increasing training epochs, and refining model architectures could significantly improve chatbot performance.

Future research should focus on optimizing the custom deep learning model by refining its architecture, experimenting with different activation functions, and increasing training iterations. Fine-tuning DistilGPT2 with a larger dataset and extending training epochs could improve accuracy and response fluency. Exploring alternative evaluation metrics that are better suited for conversational AI, such as sentiment analysis scores, coherence metrics, and user engagement tracking, could provide a more comprehensive assessment of the chatbot’s performance. Implementing reinforcement learning techniques, where the chatbot learns from real-time user feedback, may further enhance response quality dynamically. By addressing these aspects, more intelligent, responsive, and engaging conversational chatbots can be developed to deliver a seamless and natural user experience.

4.0 Conclusion

The findings of this research highlight the significant advancements made in chatbot development, particularly in enhancing memory storage and retrieval systems. The implementation of context-aware and personalized interactions has successfully addressed key limitations associated with traditional conversational agents. The results demonstrate that by enabling the chatbot to recall previous interactions, user engagement and satisfaction have been significantly improved. The integration of a generative-based model, combined with meticulous data preprocessing using a distilled GPT-2 tokenizer, has resulted in a chatbot capable of generating coherent, contextually relevant, and dynamic responses. The evaluation metrics used in validation, including fluency, user satisfaction, memory recall, perplexity, diversity, and consistency, provide strong evidence of the chatbot's effectiveness. Positive feedback from user surveys and interviews further underscores the success of this approach in aligning the chatbot's functionality with real-world user needs.

The conclusions drawn from this study emphasize the importance of incorporating memory recall and contextual awareness in chatbot design. By overcoming the limitations of traditional chatbots that lack long-term conversation retention, this research contributes to the broader advancement of natural language processing and artificial intelligence. The chatbot's ability to produce meaningful and personalized responses enhances user interactions, reinforcing trust and usability in AI-driven communication systems. The research findings validate the effectiveness of a user-centered design approach and highlight the potential for future improvements in conversational AI. Furthermore, the study provides insights into optimizing chatbot algorithms to balance computational efficiency with response accuracy, ensuring that the chatbot remains both intelligent and resource-efficient.

Based on the findings of this study, several recommendations are proposed for future improvements and applications. First, further optimization of memory retention mechanisms can enhance the chatbot's long-term recall capabilities, making interactions even more seamless and personalized. Second, incorporating advanced deep learning techniques, such as reinforcement learning, could improve the chatbot's ability to learn from user interactions and adapt its responses over time. Additionally, expanding the chatbot's training data with a more diverse dataset can further enhance response accuracy and robustness across various conversational domains. Future research should also explore integrating multimodal capabilities, such as speech and visual recognition, to make chatbot interactions more immersive and accessible. Lastly, continued user feedback collection and iterative model refinement should be prioritized to ensure the chatbot evolves in alignment with user expectations and emerging trends in artificial intelligence.

5.0 References

Chempavathy, B., Prabhu, S. N., Varshitha, D. R., Vinita, & Lokeswari, Y. (2022). AI-based chatbots using deep neural networks in education. *Proceedings of the 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, Coimbatore, India, 124-130. https://doi.org/11.1109/ICAIS53314.2022. 9742771.

Colby, K. M. (1975). *Artificial paranoia: A computer simulation of paranoid processes*. Elsevier.

Day, M.-Y., & Hung, C.-S. (2019). AI affective conversational robot with hybrid generative-based and retrieval-based dialogue models. *Proceedings of the 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, Los Angeles, CA, USA, 403-409. https://doi.org/10.1109/IRI.2019.00068.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., & Hon, H.-W. (2019). Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197.*

Guo, D., Tang, D., Duan, N., Zhou, M., & Yin, J. (2018). Dialog-to-action: Conversational question answering over a large-scale knowledge base. *Proceedings of the School of Data and Computer Science, Sun Yat-sen University, Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, China.*

Huang, K. H., Lee, J., & Liao, C. (2020). Improving chatbot engagement with long-term memory systems. *Computational Linguistics and AI*, 12, 4, pp. 209-225.

Rajan, M. H., Rebello, K., Sood, Y., & Wankhade, S. B. (2021). Graph-based transfer learning for conversational agents. *Proceedings of the 2021 6th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, 1335-1341. https://doi.org/10.1109/ICCES51350.2021.9489179.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.

Vinyals, O., & Le, Q. (2015). A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Weston, J., Chopra, S., & Bordes, A. (2015). Memory networks. *arXiv preprint arXiv:1410.3916*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. (2020). Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xygi, E., Andriopoulos, A. D., & Koutsomitropoulos, D. A. (2023). Question answering chatbots for biomedical research using transformers. *Proceedings of the International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*.

Zhang, Z., Cui, L., & Liu, Y. (2020). Memory-augmented transformers for long-term contextual understanding. *AI and NLP Journal*, 29, 2, pp. 145-167.

Compliance with Ethical Standards Declaration

Ethical Approval

Not Applicable

Competing interests

The authors declare that they have no known competing financial interests

Funding

All aspect of the work was carried out by the author