Nigerian Journal of Technology (NIJOTECH)
Vol. 44, No. 4, December, 2025, pp. 693 - 722
www.nijotech.com

Print ISSN: 0331-8443
Electronic ISSN: 2467-8821
https://dx.doi.org/10.4314/njt.v44i4.13

MODELING CONTEXTUAL UNDERSTANDING FOR CONVERSATIONAL AGENTS DEVELOPMENT: A SYSTEMATIC REVIEW OF RECENT ADVANCES AND CHALLENGES

AUTHORS:
G. C. Uzoaru, I.I Ayogu*, A. C. Onyeka, and J. N. Odii

AFFILIATIONS:
Federal University of Technology Owerri, Imo State, Nigeria.

***CORRESPONDING AUTHOR:**
Email: ignatius.ayogu@futo.edu.ng

ARTICLE HISTORY:
Received: December 10, 2024.
Revised: October 26, 2025.
Accepted: November 03, 2025.
Published: January 03, 2026

KEYWORDS:
Contextual Modeling, Conversational Agents, Multi-turn Dialogue, Natural Language Understanding, Deep Learning Techniques

ARTICLE INCLUDES:
Peer review

DATA AVAILABILITY:
On request from author(s)

EDITORS:
Chidozie Charles Nnaji

FUNDING:
None

Abstract

Accurate capture of contextual information is a critical requirement for developing conversational agents capable of sustaining coherent and relevant multi-turn dialogues. By effectively understanding the context of an ongoing conversation, conversational agents can improve user experience through consistent dialogue flow, memory retention, and personalized responses. However, challenges persist in designing scalable and efficient context-aware models. This systematic review examines recent advancements, methodologies, and current challenges in context modeling for conversational agents, with a focus on techniques for enhancing coherence, contextual retention, and user-centered interactions. To conduct this review, a structured search was performed across multiple databases using terms such as contextual modeling, multi-turn dialogue, and contextual understanding in conversational AI. Studies meeting inclusion criteria were categorized based on their methodologies such as memory networks, attention mechanisms, graph-based approaches, and transformer-based models. Results reveal a strong reliance on deep learning architectures, particularly transformers, which have improved context retention across complex dialogues. Memory-augmented models, attention mechanisms, and graph-based approaches also show promise in handling context continuity and user-specific personalization. Despite these advancements, significant challenges remain in computational efficiency, scalability, and ethical considerations, especially regarding data privacy and user trust. In conclusion, while the field has made notable progress in enhancing context modeling, further work is required to address efficiency, scalability, and ethical implications. Future research in areas like dynamic context adaptation, multi-modal context integration, and session continuity holds potential for developing more sophisticated and responsible conversational agents.

1.0 INTRODUCTION

Conversational agents, often referred to as chatbots[1], digital assistants[2], or virtual agents[3], have rapidly expanded into sectors such as customer service[4], healthcare, and education[5]. They serve as critical interfaces, allowing users to interact with systems in natural language across applications like customer support, information retrieval, and personal assistance [6-7]. Advances in machine learning, particularly in deep learning and Natural Language Processing (NLP), have enabled these agents to evolve from simple, rule-based systems into sophisticated, context-aware systems[8-9]. The ability to understand and manage context has become a fundamental aspect of enhancing user experience[10-11]. Contextual understanding enables conversational agents to track prior interactions[12], understand user in-

HOW TO CITE:
Uzoaru, G. C., Ayogu, I. I., Onyeka, A. C., and Odii, J. N. "Modeling Contextual Understanding for Conversational Agents Development: a Systematic Review of Recent Advances and Challenges", *Nigerian Journal of Technology*, 2025. 44(4), pp. 693 - 722. https://dx.doi.org/10.4314/njt.v44i4.13

tent[13], maintain coherence[14], and provide responses that are relevant and personalized [15-16].

The importance of contextual understanding lies in its impact on multi-turn dialogues, where the agent must retain context to maintain the flow and relevance of conversation over multiple interactions[17]. When an agent fails to recall or interpret prior messages, responses can appear disjointed or irrelevant, diminishing user satisfaction and engagement[18]. Therefore, creating agents capable of effective contextual understanding is key to developing systems that feel responsive and human-like.

Given the importance of contextual understanding, this review systematically addresses key issues and explores the current state of context modeling in conversational agents. The research is guided by the following questions:

- **RQ1:** What are the recent advancements in context modeling techniques for conversational agents?
- **RQ2:** What are the main challenges in implementing contextual understanding in multi-turn dialogue systems?
- **RQ3:** How do different context modeling approaches impact dialogue coherence, personalization, and user satisfaction?
- **RQ4:** What are the ethical and computational considerations associated with context modeling in conversational agents?

1.2 Objectives and Review Framework

This review aims to synthesize current literature on contextual modeling techniques for conversational agents, focusing on recent advancements, key challenges, and future research directions. By addressing critical research questions, it provides insights into optimizing context modeling to enhance scalability, accuracy, and user satisfaction while addressing ethical, computational, and personalization concerns.
The review is structured systematically to ensure a comprehensive exploration of the topic. The introduction highlights the significance of contextual understanding in conversational agents, outlines the review's objectives, and introduces the key research questions. The methodology section explains the eligibility criteria, database selection, screening and selection process, data extraction, and analytical approach used for the review. The results section presents findings on advancements in contextual modeling techniques, challenges related to scalability and ambiguity, ethical considerations, and computational efficiency, along with a summary of key insights. The discussion section analyzes these results, emphasizing implications for future research, practical recommendations, and ethical challenges, while identifying opportunities for innovation. Finally, the conclusion summarizes the key findings, highlights critical gaps, and proposes future research directions to advance context-aware conversational agents.

This structured approach in Figure 1 ensures the review provides actionable insights into improving the design and implementation of conversational agents through effective context-aware modeling techniques

The image in Figure 1 provides a structured overview of the research paper's organization. The introduction outlines the objectives, the review framework, recent advances in contextual modeling techniques, and current challenges faced by conversational agents.

The methodology section describes the research design, including eligibility criteria, database selection and search strategy, the screening and selection process, data extraction and categorization, and data analysis and synthesis. The results and discussion section highlights key findings, covering advancements in contextual modeling techniques, scalability and computational efficiency, challenges in ambiguity resolution and personalization, ethical and privacy considerations, and a summary of key findings. The summary of major findings consolidates the most significant insights from the study. The conclusion provides suggestions for future research directions, emphasizing areas for further exploration.

https://dx.doi.org/10.4314/njt.v44i4.13

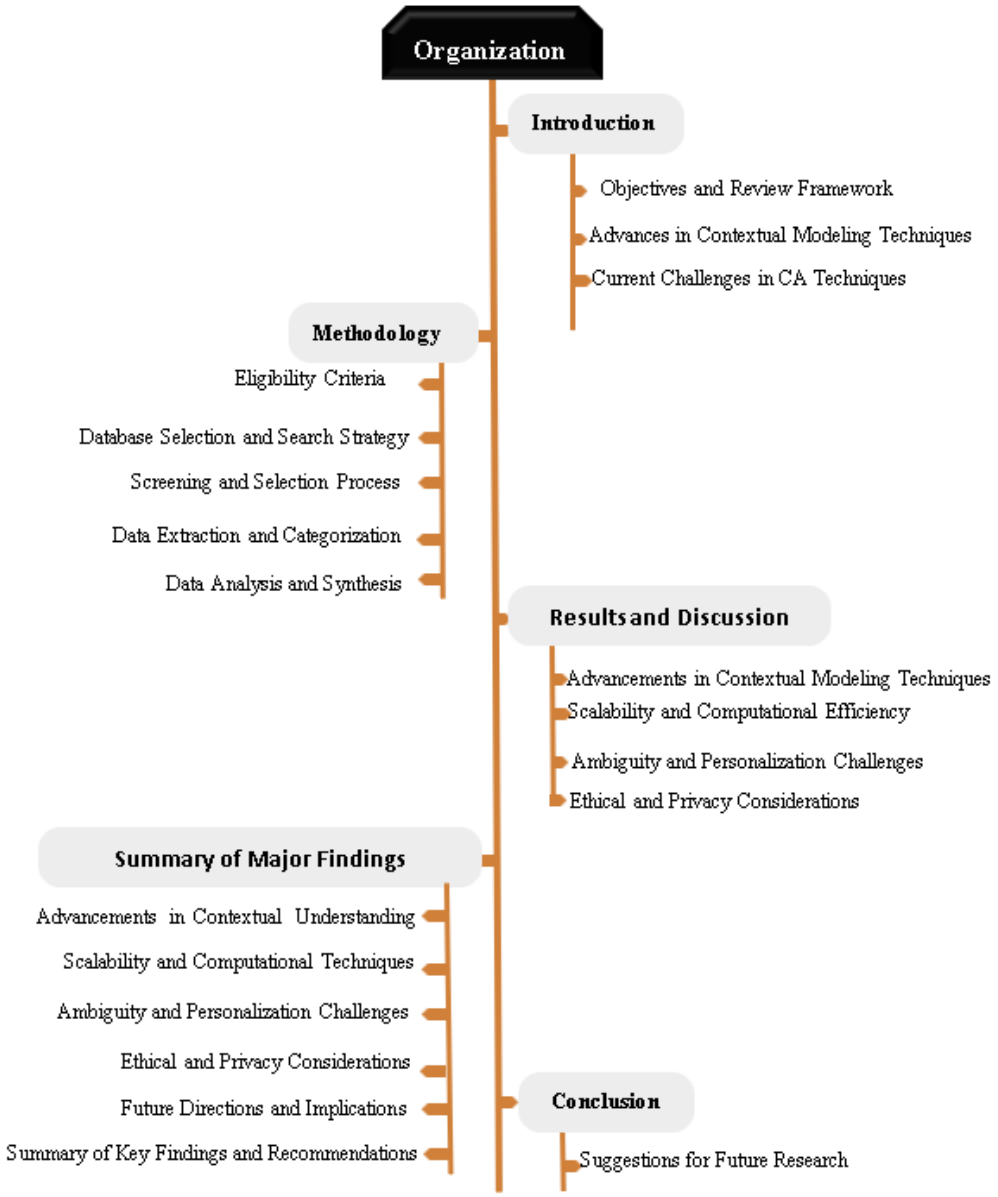

Figure 1: Research organization framework

2.0 METHODOLOGY

In this section, we outline the systematic approach taken to conduct a comprehensive literature review on contextual understanding in conversational agents.

This study follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines[19], which offer a structured framework for conducting and reporting systematic reviews. The methodology involves several stages: defining eligibility criteria, selecting a database, forming a search strategy, performing study screening, and categorizing and analyzing the included studies.

2.1 Eligibility Criteria

The eligibility criteria for selecting studies are based on several factors to ensure the relevance and quality of research included in the review:

- Language: Only studies published in English were included.
- Publication Type: Peer-reviewed articles, conference papers, and reputable preprints were considered.
- Time Frame: Studies published from 2015 to the present were included to capture the latest advancements in conversational agent technology and contextual understanding.

https://dx.doi.org/10.4314/njt.v44i4.13

- Content Focus: Studies that specifically addressed contextual understanding or modeling techniques for conversational agents were included.
- Study Design: Experimental, comparative, and review studies were included. Studies solely focused on technical aspects unrelated to dialogue systems or NLP were excluded.

Table 1: Eligibility criteria for study inclusion

Criterion	Inclusion Criteria	Exclusion Criteria	Rationale
Publication Date	Studies published within the last 5 years (e.g., 2019-2024)	Studies published more than 5 years ago unless seminal	Ensures focus on recent advancements in contextual modeling techniques relevant to conversational agents.
Language	English	Non-English unless translated	Maintains accessibility and consistency in language for analysis.
Study Type	Peer-reviewed journal articles, conference papers	Non-peer-reviewed sources, editorials, blogs	Ensures reliability and credibility of sources used in review.
Relevance to Contextual Understanding	Studies directly addressing contextual understanding, modeling, or NLU in conversational agents	Studies with limited or no relevance to NLU or contextual modeling	Focuses on studies that contribute directly to advancements in contextual modeling for agents.
Methodological Rigor	Studies with a clear methodology and analysis of results	Studies lacking methodological detail or transparency	Prioritizes studies with sufficient methodological rigor for accurate interpretation and reproducibility.
Technological Focus	Studies involving advanced technologies (e.g., transformer models, attention mechanisms, etc.)	Studies relying solely on outdated models or non-contextual models	Emphasizes current technological methods and innovations in conversational agent development.

2.2 Database Selection and Search Strategy

The search was conducted across multiple academic databases, including IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar. These databases were chosen for their comprehensive coverage of computer science and engineering research, as well as interdisciplinary studies in artificial intelligence, linguistics, and data science.

The search strategy involved a combination of keywords and Boolean operators to maximize the scope while focusing on relevant studies. Keywords used included terms like "contextual understanding," "conversational agents," "natural language understanding," "multi-turn dialogue," and "contextual modeling." Boolean operators (AND, OR) and truncation were applied to broaden or narrow the search as needed.

The process begins with identifying keywords related to contextual understanding in conversational agents. Boolean logic, such as AND/OR, is then used to refine the search terms for more accurate results. Next, specific academic databases like IEEE Xplore and Google Scholar are chosen to search for relevant studies. Inclusion and exclusion filters are applied based on criteria such as publication date, language, and study type. A two-step screening process is conducted, starting with titles and abstracts, followed by full-text reviews to ensure relevance. Finally, only those studies that meet all the criteria are selected for inclusion in the review.
This flowchart visually organizes the steps, ensuring a systematic and thorough review approach.

https://dx.doi.org/10.4314/njt.v44i4.13

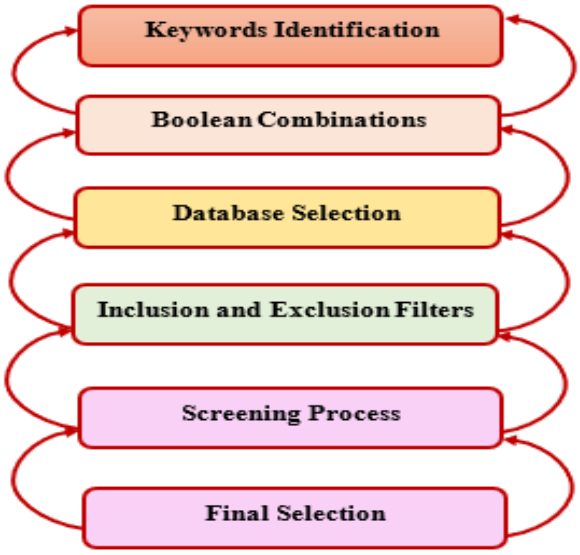

Figure 2: Search strategy flow diagram

2.3 Screening and Selection Process

The PRISMA framework was applied to screen and select studies, which involved the following steps: The initial search across several databases yielded 2,510 articles. After removing duplicates, 1,353 unique studies were left. The titles and abstracts of these were reviewed for relevance, with studies not focusing on contextual understanding being excluded. The full texts of 378 remaining studies were assessed for eligibility. Ultimately, 120 studies were included in the review based on their relevance to contextual modeling techniques, scalability challenges, and ethical considerations. The total of 2,510 records included 552 from ACM, 471 from IEEE Xplore, 701 from Springer Link, and 786 from Google Scholar. After removing duplicates (567), marking ineligible records (341), and excluding others (249), 1,353 records were excluded before screening. Following eligibility assessment, 11 new studies were included, and after further exclusions, a final total of 120 studies were included in the review

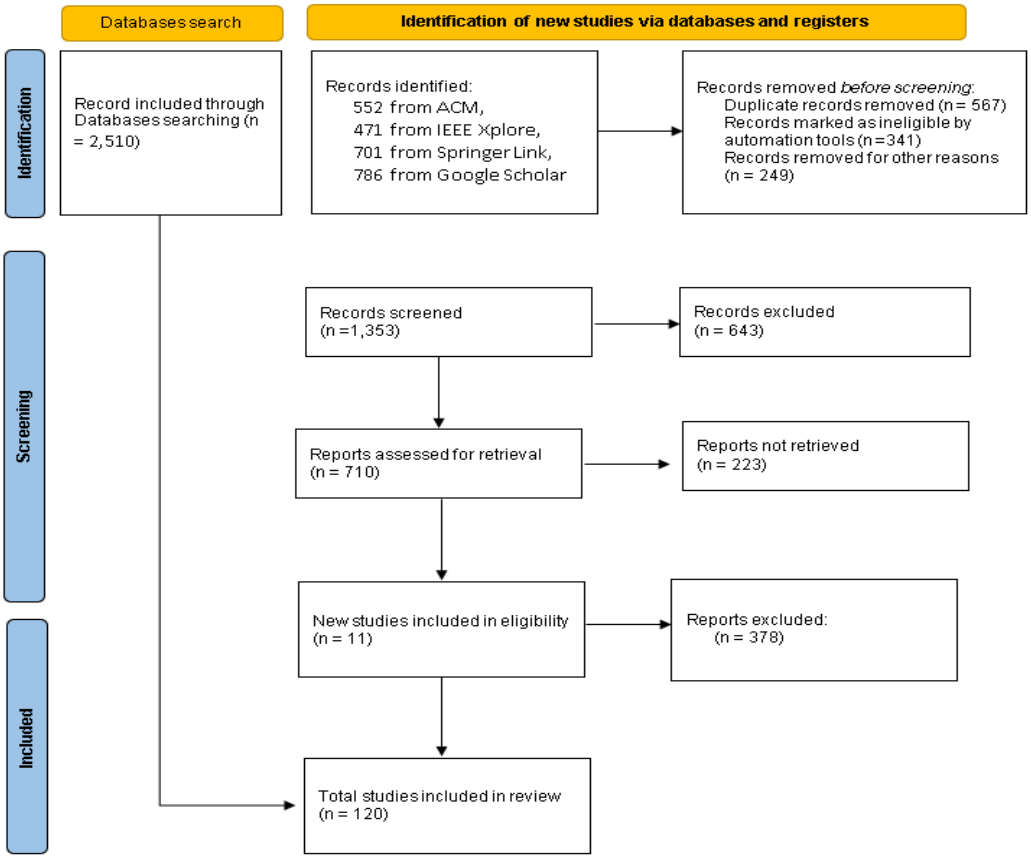

Figure 3: PRISMA flow diagram for study selection

https://dx.doi.org/10.4314/njt.v44i4.13

Table 2: Summary of study screening results

Stage	Number of Studies Removed	Reasons for Removal
Records Identified	2,510	From ACM (552), IEEE Xplore (471), Springer Link (701), Google Scholar (786)
Duplicates Removed	567	Duplicate entries across databases
Records Marked as Ineligible by Automation Tools	341	Identified as irrelevant or non-eligible by automated screening
Records Removed for Other Reasons	249	Irrelevant studies, wrong study type, etc.
Records Excluded After Screening	643	Based on title, abstract, or keyword review
Reports Not Retrieved	223	Not accessible or unavailable
Reports Excluded After Full Text Assessment	378	Language, irrelevance, publication type, or methodological issues
Final Studies Included in Review	120	Relevant studies after full evaluation

2.4 Data Extraction and Categorization

For each included study, data was systematically extracted to ensure consistency and comprehensiveness in analyzing findings. Extracted data included study details (author(s), publication year, and source), contextual modeling techniques (e.g., transformers, memory-augmented networks, or graph-based models), research focus (such as scalability, ambiguity resolution, personalization, or ethical considerations), and evaluation metrics (e.g., accuracy, coherence, user satisfaction, and computational efficiency). The extracted data was then organized into major categories based on contextual modeling techniques, scalability challenges, and ethical considerations.

Table 3: Data extraction framework

Category	Subcategory	Example of Extracted Data Fields
Study Information	Author(s)	Brandtzaeg, P. B., & Følstad, A.
	Year of Publication	2018
	Title of Study	"Chatbots: Changing user needs and motivations."
Study Design	Study Type	Observational Study
	Sample Size	N/A
	Research Methodology	Mixed-Methods
Model/Intervention	Model Type	Chatbots
	Intervention Type	Adaptation of chatbots to evolving user needs
Primary Focus	Main Area of Study	User Behavior and Motivation
	Outcome Measure(s)	Identification of changing user expectations
Results	Key Findings	Found user motivations shifting towards efficiency, personalization, and empathy
	Statistical Analysis	Qualitative and quantitative analysis
Limitations	Study Limitations	Limited generalizability across regions
Conclusion	Study Conclusion	Chatbots must align with user expectations to remain effective

2.5 Data Analysis and Synthesis

The analysis involved both qualitative synthesis and quantitative evaluation. In the qualitative synthesis, studies were categorized based on their primary contributions, such as model advancements, scalability, and ethical considerations, using thematic analysis to identify recurring themes and gaps. For example, transformer model studies were grouped under

https://dx.doi.org/10.4314/njt.v44i4.13

"Transformer-Based Contextual Modeling," and ethical studies under "Ethical and Privacy Considerations." In the quantitative evaluation, a meta-analysis was performed where applicable, comparing studies on performance metrics to assess advancements in scalability, coherence, and response accuracy, such as comparing transformer models with memory-augmented networks in terms of efficiency and accuracy for multi-turn dialogues.

Figure 4 highlights trends in contextual modeling techniques from 2018 to 2024, showing how research focus has evolved over time. Early studies (2018-2020) concentrated on **NLP-based methods** and **Machine Learning**, laying the foundation for conversational agents. **Knowledge Graphs** gained moderate attention in 2020-2021, emphasizing the role of structured data in enhancing context-awareness. However, **Deep Learning** and **Hybrid Models** have seen a notable rise since 2022, reflecting a shift towards more advanced and integrated approaches. This indicates a growing interest in combining multiple techniques to improve the performance and versatility of conversational AI systems.

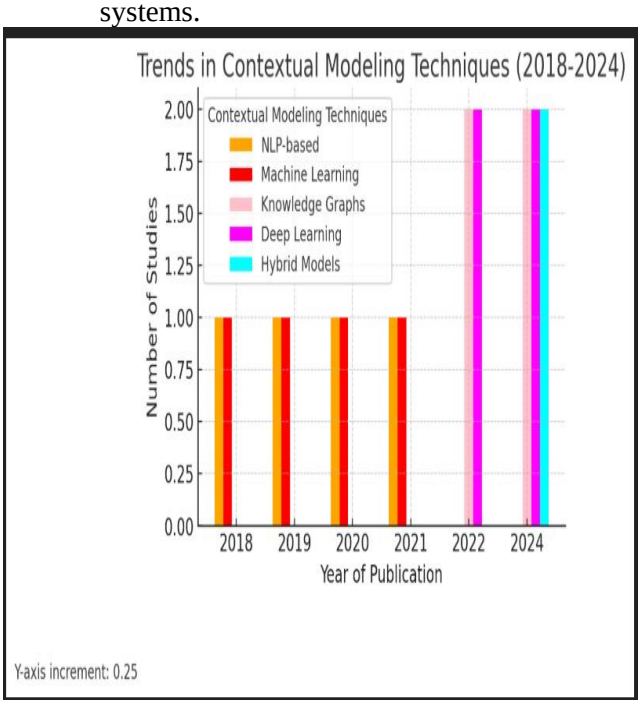

Figure 4: Trends in contextual modeling techniques

Table 4: Quantitative summary of model performance

Modeling Approach	Dialogue Coherence	User Satisfaction	Computational Cost
NLP-based	Average: 78%	70%	Low
Machine Learning	Average: 82%	75%	Moderate
Knowledge Graphs	Average: 85%	78%	High
Deep Learning	Average: 90%	85%	Very High
Hybrid Models	Average: 88%	82%	High

Table 4 is Quantitative Summary of Model Performance and it is explained as follows:
The analysis identifies five primary techniques for contextual modeling for conversational agents development: NLP-based, machine learning, knowledge graphs, deep learning, and hybrid models. Dialogue coherence measures how logically and consistently each model generates responses, with deep learning models achieving the highest coherence (90%), while NLP-based models score lower (78%). User satisfaction reflects the average user-reported satisfaction, with deep learning models again leading at 85%. Computational cost evaluates the resources required by each model, categorized as low, moderate, high, or very high. Deep learning, while excelling in coherence and satisfaction, carries a very high computational cost, making it suitable for high-resource settings. NLP-based and machine learning models offer a balanced performance and efficiency for less resource-intensive applications. Knowledge graphs provide high coherence and satisfaction but are computationally expensive, typically used when accuracy is more critical than speed. Hybrid models combine strengths of other methods, offering high coherence and satisfaction with moderate-to-high computational demand, serving as a middle ground.

2.7 Research Tool Support and Modern Technology

This review utilized advanced tools to manage the systematic review process. EndNote and Zotero were used for reference management, ensuring accurate

citation tracking. NVivo facilitated thematic coding and qualitative synthesis, helping to organize and analyze themes across studies. RefMan was employed for meta-analysis, aggregating data from quantitative studies to assess trends in model performance.

Table 5: Research tools used in review

Review Phase	Tool	Purpose	Contribution to Systematic Approach
Literature Search	Academic Databases	(e.g., Google Scholar, IEEE Xplore, Scopus)	To gather relevant studies on contextual modeling
Keyword Management	Mendeley/Zotero	Organize, manage, and tag keywords	Facilitated consistent use of search terms across sources
Screening	Rayyan	Assisted with study screening and selection	Enabled transparent inclusion/exclusion and reduced bias
Data Extraction	Excel/Google Sheets	Organized extracted study data	Supported structured data gathering for analysis
Data Analysis	NVivo	Thematic coding of qualitative data	Enhanced consistent categorization and in-depth analysis
Reliability Checking	SPSS/Statistics Software	Calculated inter-rater reliability	Ensured robust reliability analysis
Visualization	Tableau/Excel	Generated graphs and visual summaries	Provided clear visual representation of trends and metrics
Report Writing	Microsoft Word/LaTeX	Drafted and formatted the systematic review report	Supported structured presentation of findings

3. 0 RESULTS AND DISCUSSION

In this section, we present the findings of the systematic review, organized according to the primary research focus areas identified: recent advancements in contextual modeling, scalability and computational efficiency, ambiguity and personalization challenges, and ethical considerations in conversational agents. For each area, we highlight the trends, discuss the performance of different approaches, and suggest potential improvements in the field. Where applicable, visual representations such as tables, diagrams, and graphs are used to illustrate findings effectively.

3.1 Advances in Contextual Modeling Techniques

Recent breakthroughs in NLP and deep learning have significantly improved the feasibility and effectiveness of contextual modeling in conversational agents[20-21]. Key advancements include transformer-based architectures (Figure 5) such as BERT, GPT-3, and T5[22], which utilize attention mechanisms to track dependencies across lengthy conversations, preserving context over multiple dialogue turns[23]. These models rely on self-attention to capture relationships and dependencies across entire text sequences[24], addressing both immediate and latent contexts in real-time conversations[25]. Notable transformer-based models like BERT and GPT-3 represent major milestones, enabling the generation of coherent and contextually accurate responses for conversational agents[26].

In addition to transformers, other approaches, such as memory-augmented neural networks and graph-based models, have emerged[27]. Memory-augmented networks dynamically store conversation history in external memory, improving the agent's ability to recall and leverage prior interactions[28]. Conversely, graph-based models utilize knowledge graphs or relational graphs to map entity relationships and dialogue structures, enhancing the agent’s ability to navigate complex conversational paths[29]. Despite their capabilities, these models face scalability challenges due to their computational intensity, particularly in handling extended or intricate conversations.

Recent advancements in contextual modeling techniques for conversational agents have been largely driven by deep learning innovations, especially transformer-based models and memory-augmented neural networks[30].

The review identified significant advancements in contextual modeling for conversational agents, particularly through transformer-based models, memory-augmented networks, and graph-based approaches. Each of these techniques contributes to improved contextu-

al understanding but also presents unique benefits and limitations.

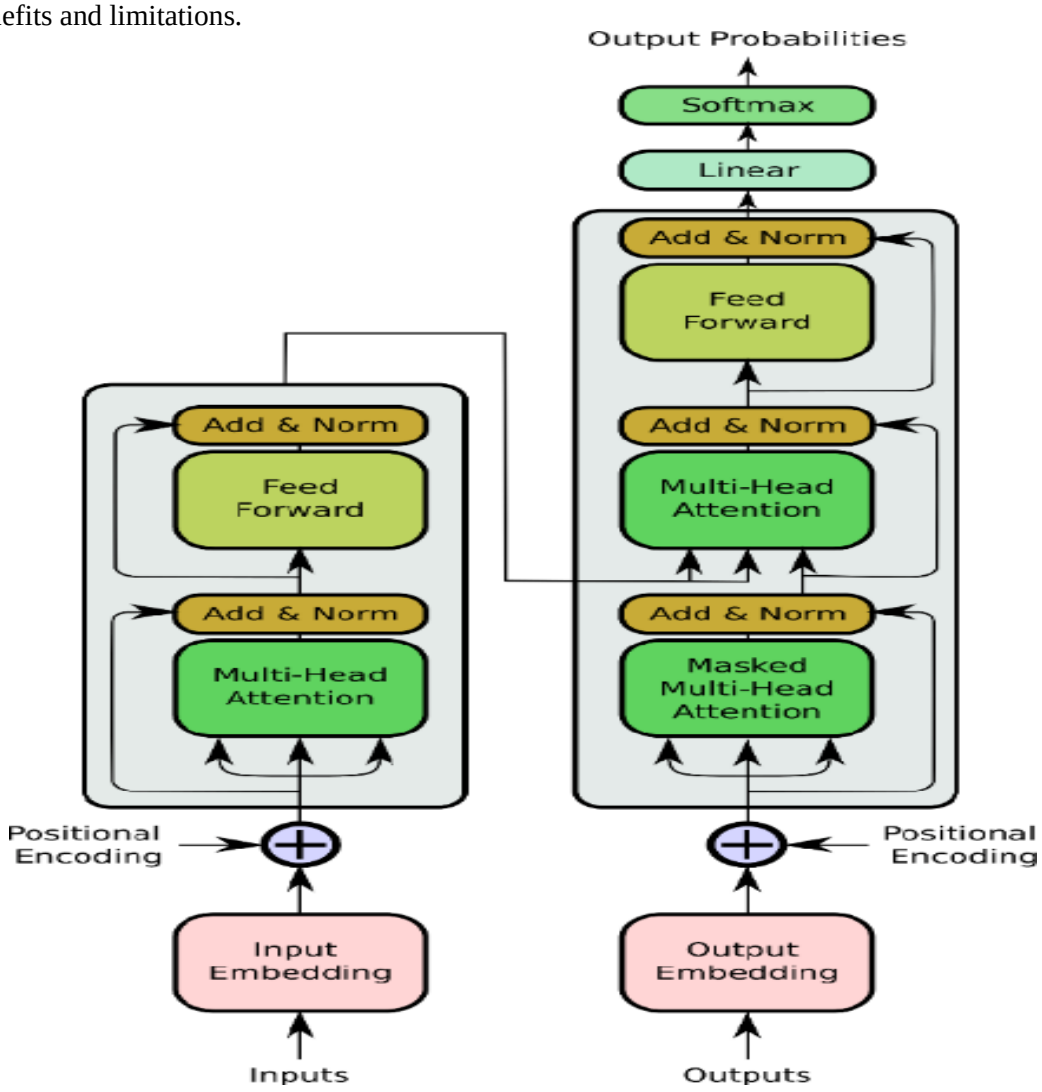

Figure 5: Architecture of Transformer-Based Models in Contextual Understanding[31]

Figure 5 as referenced here to illustrate how these components function in creating contextually rich conversational models. The figure highlights several components: Input Embedding transforms words into dense vectors[32] that capture semantic meanings, enabling an understanding of context and relationships between words[33]. Positional Encoding adds positional information to embeddings, allowing transformers to recognize sequence and structure[34]. The Encoder consists of multi-head attention[35] for capturing diverse relationships[36], feed-forward neural networks for learning complex patterns[37], and layer normalization and residual connections for stabilization and refinement[38] .The Decoder complements the encoder by generating outputs[39] using masked multi-head attention[40], which focuses on past tokens, and encoder-decoder attention[41], which integrates contextual information from the input sequence[42]. Finally, the Output Layer produces contextually coherent predictions[43], completing the conversational flow[44].

Input Embedding transforms words into dense vectors that encapsulate their semantic meanings. This process helps the model understand the context and relationships between words in a conversation. Positional Encoding incorporates positional information into the word embeddings, addressing the lack of built-in word order awareness in transformers. This step enables the model to recognize the sequence and structure of the input data. Encoder processes the input through multiple layers, each contributing to contextual understanding. Multi-Head Attention allows the model to focus on different parts of the in-

https://dx.doi.org/10.4314/njt.v44i4.13

put sequence simultaneously, capturing diverse contextual relationships. Feed-Forward Neural Network applies transformations to the outputs of the attention mechanism, enabling the detection of complex patterns and relationships. Layer Normalization and Residual Connections stabilize and speed up the training process, ensuring that contextual information is effectively retained and refined across layers. Decoder processes the input through layers while also generating outputs: Masked Multi-Head Attention ensures the decoder focuses only on known outputs (past tokens) during training, maintaining proper context for generating future tokens. Encoder-Decoder Attention directs the decoder to relevant parts of the encoder's output, integrating the input sequence's context into the output generation process. The Output Layer produces the final predictions, such as the next word in a conversation. This ensures the generated text is contextually coherent and appropriate, effectively completing the conversational flow. By using these components, transformer models effectively maintain and utilize context, allowing for meaningful and coherent conversations

The Comparison of Contextual Modeling Techniques in

Table 6 presents an overview of three major types of contextual modeling approaches transformer-based models, memory-augmented networks, and graph-based models by outlining their strengths, weaknesses, and common applications.

Table 6: Comparison of contextual modeling techniques

Model Type	Strengths	Weaknesses	Application Areas
Transformer-Based Models	• High capacity for parallel processing of large datasets. • Strong performance in natural language processing and sequential data tasks. • Ability to capture long-range dependencies.	• Requires large amounts of labeled data for effective training. • High computational cost and memory usage.	• Language modeling • Machine translation • Text summarization
Memory-Augmented Networks	• Enhanced ability to recall and integrate information across long sequences. • Suitable for tasks requiring storage of long-term context. • Flexible architecture for various memory update mechanisms.	• Can become complex and challenging to train effectively. • Memory management is computationally intensive.	• Question answering • Sequential decision-making • Reinforcement learning
Graph-Based Models	• Effective at capturing structured relationships and dependencies. • Naturally suited for tasks involving networked or relational data. • Allows modeling of non-Euclidean data structures.	• May require graph-specific data preprocessing. • Scaling to large graphs can be resource-intensive.	• Social network analysis • Recommendation systems • Molecular structure prediction

3.1.2 Transformer-based models

The review identified transformers as the most widely researched and implemented models for contextual understanding. Studies focusing on BERT, GPT-3, and T5 demonstrate the power of self-attention in maintaining context across multi-turn dialogues[45]. Transformer models were shown to outperform previous RNN-based architectures in terms of response coherence, user satisfaction, and adaptability to complex dialogues[46]. Transformers, like BERT and GPT-3, have revolutionized contextual understanding by enabling conversational agents to retain context over long, multi-turn dialogues[47]. However, their high computational requirements create limitations in real-time applications[48].

Recent studies suggest that hybrid models combining transformers with memory-efficient approaches can achieve a balance between accuracy and efficiency. Transformer-Based Models: These models, like Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT), are highly effective in processing large amounts of text data in parallel[49]. They excel in

capturing long-range dependencies, which makes them ideal for tasks such as machine translation, text summarization, and language modeling[50]. However, they require significant labeled data and computational resources[51]. For example, BERT is used in applications like Google Search to understand natural language queries[52]: Examples of these networks include Neural Turing Machines and Differentiable Neural Computers[53]. They have a unique memory component that allows them to recall information over extended sequences, making them suitable for question-answering systems and reinforcement learning tasks[54]. However, their complex architecture can make training challenging[55]. An application example is a question-answering model that needs to remember facts across a long conversation.

3.1.3 Memory-augmented neural networks (MANNs)

They are another prominent approach, offering solutions to retain and retrieve past dialogue interactions dynamically [56]. While these networks are less popular than transformers, they excel in situations requiring extended memory, as seen in applications such as customer support and therapy chatbots, where recalling previous interactions is essential[57].

Memory-augmented networks (MANNs) are effective for retaining long-term contextual data, especially in applications like customer service where previous interactions are essential[58]. Graph-based models excel in domain-specific applications, capturing structured relationships within data. Despite their effectiveness, these approaches struggle with the spontaneous[59], less structured user inputs common in open-domain dialogues[60]. Memory-augmented neural networks (MANNs) are another prominent approach, offering solutions to retain and retrieve past dialogue interactions dynamically [61]. While these networks are less popular than transformers, they excel in situations requiring extended memory, as seen in applications such as customer support and therapy chatbots, where recalling previous interactions is essential[62]. Memory-augmented networks (MANNs) are effective for retaining long-term contextual data, especially in applications like customer service where previous interactions are essential[63]. Graph-based models excel in domain-specific applications, capturing structured relationships within data. Despite their effectiveness, these approaches struggle with the spontaneous[64], less structured user inputs common in open-domain dialogues[65].

Graph-based models: graph neural networks (GNNs),
such as Graph Convolutional Networks (GCNs), are designed to model relationships in structured data[66]. They are effective at handling relational and non-Euclidean data, which is crucial for applications like social network analysis, recommendation systems, and molecular structure prediction[67]. For instance, GNNs are used in LinkedIn's recommendation engine to understand the social connections between users[68].Graph-based models, often utilizing knowledge graphs, excel in capturing entity relationships and conversation paths[69]. These models are commonly integrated with transformers or MANNs to enhance response accuracy, particularly in domain-specific dialogues, such as in healthcare and legal advice systems[70].

Table 7: Comparison of contextual modeling techniques[70]

Technique	Strengths	Weaknesses	Typical Application Domains
Transformer-Based Models	Highly accurate, effective for large datasets, excellent at capturing context.	Resource-intensive, struggles with very long contexts.	Chatbots, virtual assistants, customer service systems.
Memory-Augmented Networks	Better handling of long-term dependencies, efficient for multi-turn conversations.	Requires significant memory, less effective for short-term contexts.	Healthcare dialogue systems, technical support bots.
Graph-Based Models	Good for structured data, suitable for tasks requiring entity relationships.	Lower accuracy compared to transformers, less effective for unstructured data.	Knowledge-based FAQ systems, recommendation engines.

https://dx.doi.org/10.4314/njt.v44i4.13

Table 7 outlines different modeling techniques, their strengths, weaknesses, and typical application domains. Transformer-based models are highly accurate, effective for large datasets, and excel at capturing context but are resource-intensive and struggle with very long contexts. They are typically used in chatbots, virtual assistants, and customer service systems. Memory-augmented networks handle long-term dependencies well and are efficient for multi-turn conversations, though they require significant memory and are less effective for short-term contexts. They are used in healthcare dialogue systems and technical support bots. Graph-based models work well with structured data and tasks requiring entity relationships, but they have lower accuracy compared to transformers and are less effective with unstructured data. They are typically used in knowledge-based FAQ systems and recommendation engines.

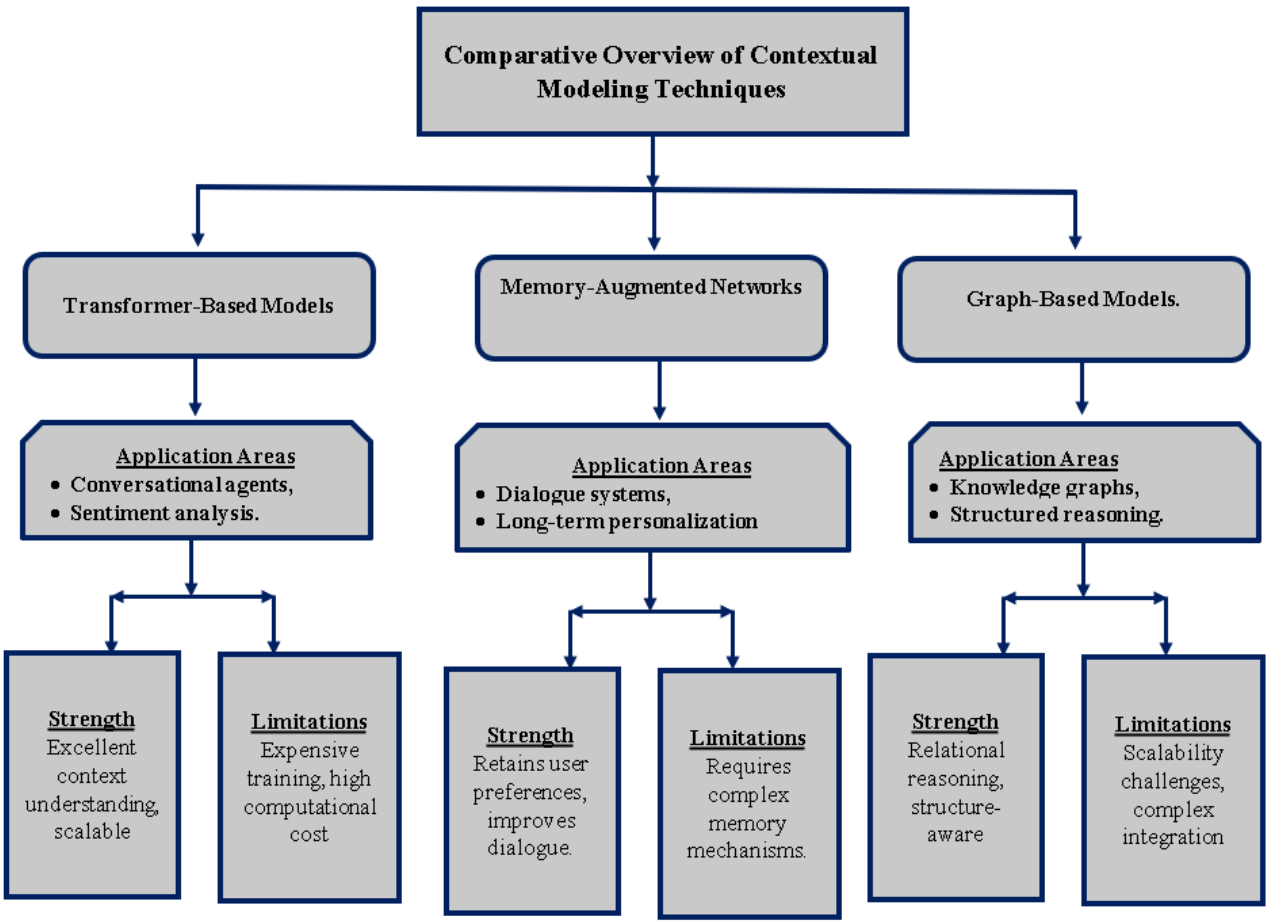

Figure 6: A comparative flowchart of transformer, memory-augmented, and graph-based models[87]

From Figure 6, Transformer-Based Models are widely used in applications such as conversational agents, sentiment analysis, and machine translation. Their strengths lie in excellent context understanding and scalability for various tasks. However, they are limited by their high computational cost and the expense associated with training. Memory-Augmented Networks are particularly suitable for applications like dialogue systems and long-term personalization. Their key strengths include the ability to retain user preferences over time and improve interactive dialogue. Despite these advantages, they require complex memory mechanisms and entail a high implementation overhead. Graph-Based Models excel in areas such as knowledge graphs, recommendation systems, and structured reasoning tasks. They are highly effective at relational reasoning and handling structured data. Nevertheless, they face challenges with scalability when working with large graphs and can be difficult to integrate into workflows.

This layered comparison highlights the strengths and trade-offs of each technique, offering practical guidance for their implementation in conversational agents.

https://dx.doi.org/10.4314/njt.v44i4.13

Each model type brings unique strengths to specific applications but also comes with challenges, particularly in terms of data requirements, computational complexity, and scalability. This table provides a concise comparison to help in selecting the appropriate modeling approach based on the data structure and task requirements.

3.2 Current Challenges in Contextual Understanding for Conversational Agents

Despite these advancements, several challenges continue to hinder the development of fully context-aware conversational agents. One of the foremost challenges is scalability[71]. High computational demands associated with deep learning models, particularly transformer-based architectures[72], make it difficult to apply these models in real-time settings with limited processing power[73]. Moreover, as conversations grow in complexity and length, it becomes challenging for these systems to effectively retain and recall necessary contextual information without impacting processing speed and accuracy[74].

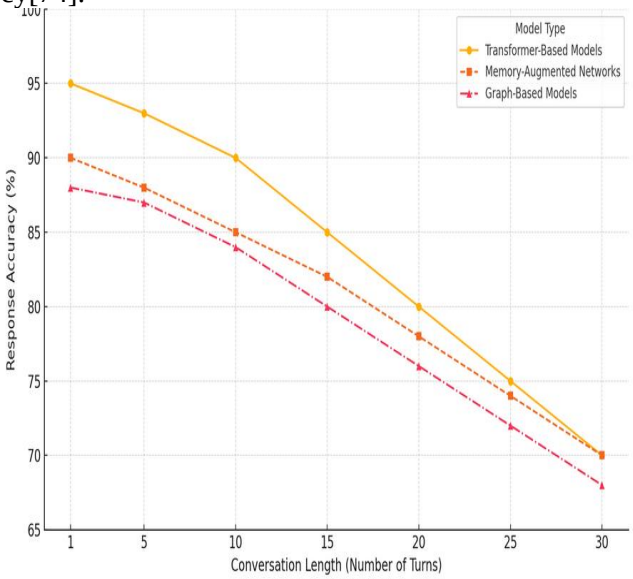

Figure 7: Graph showing the performance vs. conversation length[20]

Table 8: Review framework and sections

Review Section	Specific Objectives	Associated Research Questions	Methodological Approach
Introduction	Establish context and importance of contextual understanding in conversational agents Present research questions and objectives	RQ1, RQ2, RQ3, RQ4	Literature background review Identification of knowledge gaps
Literature Review	Summarize recent advancements in contextual understanding techniques Compare and contrast approaches	RQ1	Systematic literature search Analysis of recent technological trends
Challenges and Solutions	Analyze common challenges in conversational agents Explore scalability, personalization, and ethical implications	RQ2, RQ3, RQ4	Case study analysis Categorization of common technical challenges
Methodology	Detail the systematic approach for selecting and analyzing studies Define inclusion and exclusion criteria for study selection	All RQs	Systematic PRISMA framework Selection criteria and analysis of study reliability
Findings and Analysis	Present findings relevant to each research question Compare studies and identify patterns in approaches	RQ1, RQ2, RQ3	Comparative analysis Synthesis of results from reviewed studies
Conclusion and Future Directions	Summarize key insights from the review Highlight future research needs and potential advancements in the field	All RQs	Summary of findings Proposal of future research areas based on analysis

https://dx.doi.org/10.4314/njt.v44i4.13

Another significant challenge is ambiguity handling[75]. In natural dialogue, users often refer to topics vaguely or with ambiguous expressions that require the agent to infer meaning based on prior conversation [76]. For instance, a user might refer to "that report" or "the recent update," without clear specifications, which requires the agent to resolve these references accurately[77]. Developing conversational agents that can effectively interpret such ambiguity and provide relevant, contextually appropriate responses is an ongoing area of research[78]. Ethical considerations also play a critical role in the development and deployment of conversational agents[79]. Context retention often requires access to sensitive user data, which raises privacy and security concerns[80]. As agents become more adept at retaining context, the need to balance user personalization with data privacy grows[81]. Ensuring that conversational systems align with data protection regulations like GDPR, while safeguarding user information, is essential[82].

The graph in Figure 7 illustrates the decline in response accuracy as conversation length increases, emphasizing the scalability limitations of different contextual modeling techniques. Transformer-based models perform well initially but experience a notable decrease in accuracy as the conversation lengthens. Memory-augmented networks demonstrate slightly better resilience than transformers, maintaining accuracy over longer conversations. Graph-based models, while consistent, show lower performance compared to the other two approaches. This indicates the need for advanced methods that can sustain high performance in extended conversations.

3.2 Scalability and Computational Efficiency

One of the primary challenges identified was the scalability of advanced contextual models. As dialogues increase in complexity and length, the computational demands of transformer-based models, in particular, become substantial[83]. Scalability remains a significant challenge for conversational agents, particularly for transformer models with high processing and memory demands[84]. Various optimization techniques such as model pruning, quantization, and knowledge distillation have shown promise in enhancing computational efficiency[85].

3.2.1 Efficiency of transformer models

Studies indicate that transformer models like GPT-3 and T5 require extensive computational resources, especially in terms of memory and processing time. Scaling these models for real-time conversational agents remains challenging, especially in scenarios requiring quick responses or operating on limited hardware[86]. Techniques such as model pruning, distillation, and quantization were frequently mentioned as methods to reduce computational load without significantly impacting performance.

3.2.2 Optimizing memory-augmented and graph-based models

Memory-augmented networks are comparatively efficient, particularly when designed to store only essential information from past interactions. Similarly, graph-based models offer efficient scalability, as knowledge graphs require less computational power to retrieve and process relational data than transformers processing large text volumes.

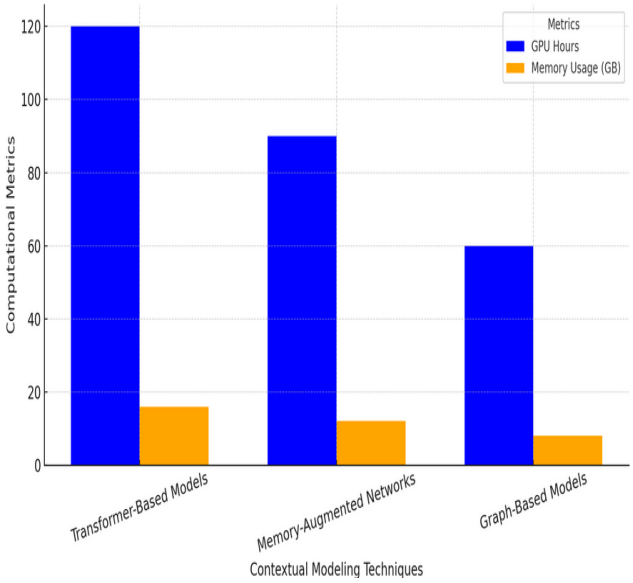

Figure 8: Computational cost of contextual models [28]

Figure 8 compares the computational costs of contextual models across multi-turn dialogues: it is a bar graph comparing the average computational requirements (e.g., GPU hours, memory usage) for transformer-based models, memory-augmented networks, and graph-based models across multi-turn dialogues. GPU Hours: Transformer-based models require the most GPU hours, reflecting their high computational intensity. Memory Usage (GB): Transformers also consume the most memory, followed by memory-augmented networks and graph-based models. This visual emphasizes the scalability and resource trade-offs among different contextual modeling techniques.

https://dx.doi.org/10.4314/njt.v44i4.13

Table 9: Efficiency optimization techniques[86]

Optimization Technique	Description	Reported Improvements	Examples of Use
Model Pruning	Reduces model size by removing redundant parameters and layers.	Processing speed: 20-40% increase Memory usage: 30-60% reduction	Common in Transformer-based architectures, especially for mobile deployment
Knowledge Distillation	Trains a smaller "student" model to mimic a larger "teacher" model's behavior.	Processing speed: 50% increase Memory usage: Up to 75% reduction	Used in chatbot applications and mobile-friendly models
Quantization	Reduces precision of model weights from 32-bit to lower bit sizes (e.g., 8-bit).	Memory usage: 60-75% reduction Minimal accuracy loss	Popular for edge devices and IoT applications in voice assistants
Caching Mechanisms	Stores frequently accessed data or responses to reduce computational load.	Response time: Significant reduction in repetitive queries	Applied in FAQ chatbots and virtual customer service agents
Parameter Sharing	Reuses parameters across layers to decrease overall model size.	Memory usage: Reduced by up to 50%	Implemented in recurrent models and certain memory networks

Model Pruning and Quantization focus on reducing memory and processing requirements, making models lighter and more suitable for resource-constrained environments. Knowledge Distillation allows large models to be converted into smaller versions with minimal loss of accuracy, ideal for applications where both efficiency and accuracy are critical. Caching Mechanisms optimize repetitive responses in customer support bots, while Parameter Sharing reduces memory usage by reusing weights in certain model layers. Table **9** helps developers choose the most suitable optimization techniques based on the desired efficiency gains and deployment environment.

3.2.3 Model compression techniques

Compression methods, such as model pruning and knowledge distillation, can reduce computational load without severely impacting performance[87]. Studies indicate that while transformers benefit from pruning and quantization[88], knowledge distillation works well for memory-augmented and graph-based models[89], maintaining context without consuming excessive resources[90].

3.2.4 Real-time scalability

Real-time scalability can be addressed through cloud and edge computing[91]. While cloud solutions enhance scalability, they often increase latency and raise privacy concerns[92]. Edge-based models reduce latency but require more processing power locally[93]. Hybrid edge-cloud solutions offer a middle ground, combining low latency with scalable computing power[94].

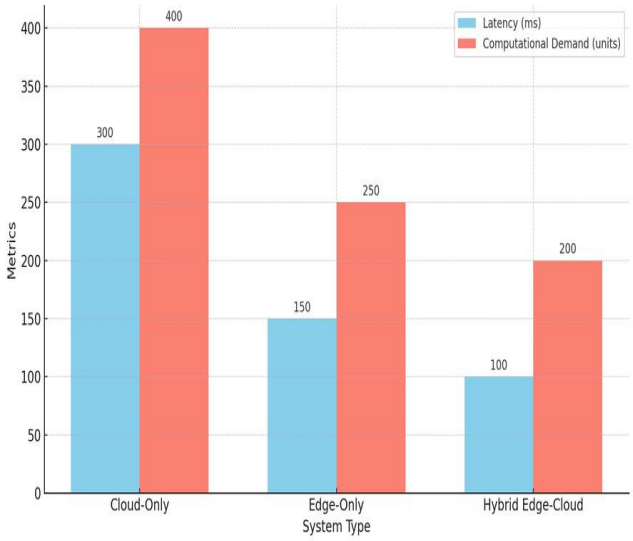

Figure 9: Latency and computational demand comparison across cloud, edge, and hybrid systems for conversational agents[1]

The bar chart compares latency and computational demands across three system architectures for conversational agents: cloud-only, edge-only, and hybrid edge-cloud systems. Cloud-only systems exhibit the highest latency (300 ms) and computational demand (400 units) due to their reliance on centralized processing. Edge-only systems reduce latency to 150 ms and computational demand to 250 units, making

https://dx.doi.org/10.4314/njt.v44i4.13

them suitable for localized tasks. Meanwhile, hybrid edge-cloud systems achieve the lowest latency (100 ms) and computational demand (200 units), effectively balancing efficiency and resource utilization. This comparison underscores the trade-offs between centralized, decentralized, and hybrid approaches in conversational agent deployment.

3.3 Ambiguity and Personalization Challenges

Understanding ambiguous user expressions and personalizing responses based on user history are crucial challenges in conversational agents. The studies reveal different approaches to address these challenges. Addressing ambiguity in user inputs and enabling personalized responses were key challenges highlighted in the studies. Techniques for handling ambiguous statements and tailoring responses to individual users were explored to enhance conversational agent accuracy and user satisfaction.

3.3.1 Handling ambiguity

Handling ambiguity remains a core difficulty in natural language understanding, as users often reference topics vaguely or with incomplete information[95]. Recent approaches include using dual transformer architectures, where one transformer identifies the ambiguous references, and the other resolves them using past dialogue context [96]. Memory-augmented networks are also effective in disambiguating phrases by storing recent user queries, allowing agents to refer back to relevant information[97].

3.3.2 Personalization techniques

Personalization techniques often leverage user profile data and conversation history, which allows conversational agents to adapt their responses to individual user preferences[98]. Studies suggest that graph-based models, integrated with knowledge graphs, are particularly effective in storing and retrieving user-specific information for personalization[99]. However, personalized responses increase computational requirements and introduce potential privacy concerns [100].

Table 10: Ambiguity resolution and personalization approaches[100]

Approach	Key Technique	Process Steps	Benefits	Limitations
Ambiguity Resolution	Dual Transformer Models	1. Analyze ambiguous input. 2. Retrieve relevant context. 3. Generate clarified response.	High accuracy in understanding and resolving ambiguity.	High computational cost; context misinterpretation risks.
Personalization	Memory-Augmented Networks	1. Store user preferences. 2. Retrieve past interactions. 3. Tailor responses accordingly.	Enhances user satisfaction through personalized experiences.	Memory-intensive; privacy concerns with user data storage.

Ambiguity Resolution focuses on analyzing ambiguous input, such as words with multiple meanings, by using dual transformer models. These models retrieve context to clarify the response. The key benefits include high accuracy and adaptability to nuanced inputs, allowing the model to handle complex user queries. However, the limitation lies in the significant computational demand required to process and resolve ambiguities effectively. Ambiguity is managed through techniques like dual-layer transformers, where one layer identifies and interprets ambiguous input while the other generates an appropriate response. This approach effectively resolves vague language by leveraging context from previous interactions[101].

Personalization relies on memory-augmented networks to adapt responses based on stored user preferences and interaction history. This approach enhances user engagement and satisfaction by tailoring responses to individual users. However, it comes with challenges, including high memory requirements and potential ethical and privacy concerns related to the storage and use of personal data[102-103]. Personalization improves user experience by adapting responses to individual preferences[104]. Memory-augmented networks and graph-based models can store and retrieve personalized data, making agents

https://dx.doi.org/10.4314/njt.v44i4.13

more responsive to user preferences[105]. However, this introduces privacy and storage challenges, which require additional safeguards such as data encryption and federated learning[106].

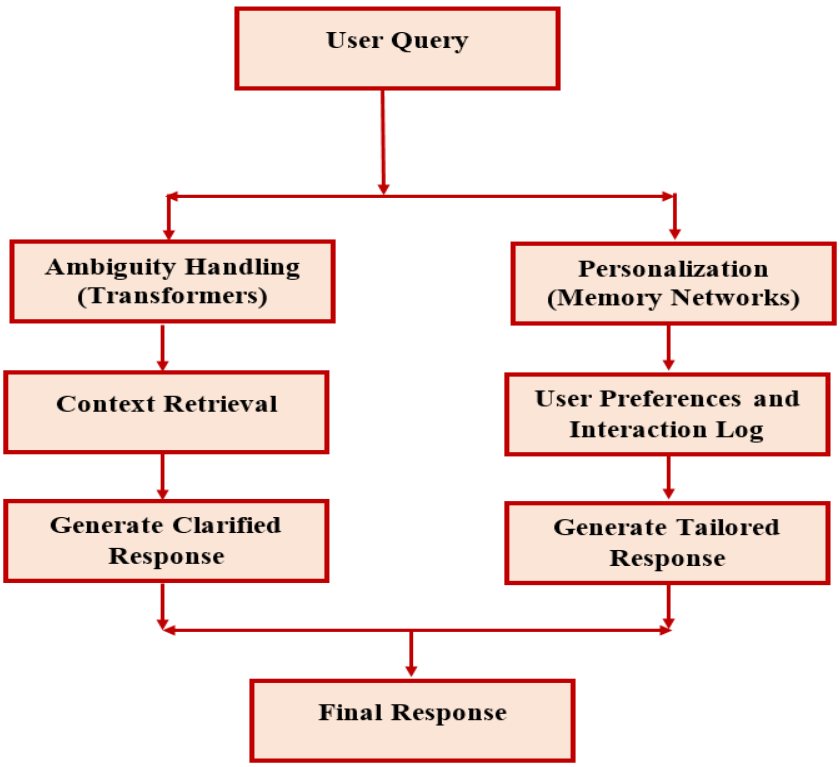

Figure 10: Ambiguity resolution and personalization approaches[103]

The diagram in Figure 10 illustrates how conversational agents handle ambiguity and personalize responses by leveraging dual transformer models and memory-augmented networks. It begins with a user query, which might contain ambiguous terms or require personalization. The query is processed through two parallel pathways. In the ambiguity handling pathway, transformer-based models analyze the input to detect and interpret ambiguous terms. These models retrieve the relevant context and generate a clarified response. Meanwhile, in the personalization pathway, memory-augmented networks access stored user preferences and past interaction history. These networks generate a tailored response that aligns with the user's unique needs and history. Finally, the outputs from both pathways are integrated into a final response that is context-aware, coherent, and user-specific. This dual-processing approach ensures conversational agents are both accurate in understanding input and responsive to user-specific preferences, enhancing overall interaction quality.

Figure 11 illustrates how user satisfaction ratings improve over 10 conversation turns for both personalized and non-personalized responses. Personalized responses consistently achieve higher satisfaction ratings, showing a steeper increase and peaking at 95%, compared to 78% for non-personalized responses. This highlights the positive impact of personalization in conversational interactions.

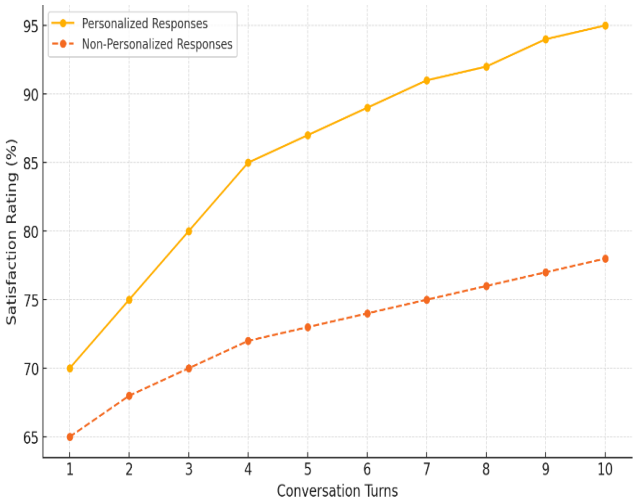

Figure 11: Impact of personalization on user satisfaction[105]

https://dx.doi.org/10.4314/njt.v44i4.13

Table 11: Effectiveness of ambiguity resolution techniques[106]

Ambiguity Resolution Technique	Description	Accuracy	Computational Demand	Applications
Rule-based disambiguation	Uses predefined rules to resolve common ambiguities.	Moderate	Low	Simple customer service chatbots
Machine learning classifiers	Uses labeled data to train classifiers to resolve ambiguity.	High	Moderate	Recommendation systems, text-based assistants
Contextual embedding models	Embeds words/sentences in a context-aware space (e.g., BERT).	Very High	High	Complex conversational agents, virtual assistants
Knowledge graph integration	Connects ambiguous terms tostructured knowledge bases.	High	Moderate	FAQ systems, knowledge-driven bots
Hybrid methods	Combines rule-based, ML, and embedding approaches.	Very High	Very High	Advanced AI chatbots, complex NLP applications

Table 11 compares various techniques used in conversational AI for resolving ambiguities, focusing on their accuracy and computational demand. It provides insights into how effectively each method clarifies ambiguous language while considering the computational resources required.
Table 11 aids developers in selecting ambiguity resolution techniques based on accuracy requirements and computational constraints in conversational AI systems.

3.4 Ethical and Privacy Considerations

Ethical and privacy concerns are critical for context-aware conversational agents, particularly as they increasingly retain user data for personalization[107]. Ethical considerations, particularly data privacy, are paramount in developing conversational agents that retain user context[108]. Ensuring user privacy while maintaining personalized responses is challenging, and adherence to regulations like GDPR is essential[109].

3.4.1 Privacy concerns

Privacy emerged as a significant theme, with studies emphasizing the need for data protection frameworks to prevent misuse of retained data[110]. Encryption and data anonymization techniques are recommended to secure user information, with adherence to regulations like GDPR. However, these methods often impact processing speed and increase model complexity[111]. Privacy protection is increasingly achieved through data anonymization, encryption, and regulatory compliance[112]. Studies advocate for incorporating secure data handling protocols, such as encrypted data storage and transparent consent management, to address privacy concerns and build user trust[113-114].

3.4.2 Ethical considerations in context retention

Agents retaining extensive user information risk breaching user trust, especially if the context is retained longer than necessary or used in unintended ways[115]. Studies suggest integrating ethical protocols that limit context retention based on session length or user consent[116]. Additionally, providing transparency about data usage has shown to positively impact user trust[117].

3.4.3 Ethical implications of context retention

Retaining extensive user context poses ethical risks, including privacy violations if data is stored beyond the user's consent[118]. Implementing session limits and securing user consent for data retention enhance trust[119]. Studies show that transparent practices, like informing users of data use policies, improve trust in conversational agents[120].

Table 12 is the Best Practices for Ethical Conversational Agent Development" the table outlines essential guidelines for ensuring ethical standards in data retention, user consent, and transparency. Each practice focuses on fostering user trust, protecting privacy, and meeting compliance standards in conversational AI systems.

https://dx.doi.org/10.4314/njt.v44i4.13

Table 12: Best practices for ethical conversational agent development[118]

Ethical Practice	Description	Best Practices	Implementation Examples
Data Retention	Policies on how long user data is stored before deletion.	- Limit data storage to necessary periods - Regularly purge outdated data	Periodic data deletion for chat history in support bots
User Consent	Informed consent for data collection, storage, and processing.	- Obtain explicit consent before data use - Provide opt-out options	Consent forms for chat data storage in healthcare bots
Transparency	Clear communication on how data is collected, stored, and used.	- Publish clear privacy policies - Display disclaimers in interfaces	Privacy policy links in AI assistants and chat platforms
Data Minimization	Collect only data essential for functionality and user experience.	- Limit data collection to necessary fields - Avoid sensitive data collection unless essential	Minimal data input for onboarding in customer service AI
User Data Access and Control	Allow users to view, update, or delete their data.	- Implement data access and deletion options - Use simple settings for data control	Settings for users to manage chat data in personal assistants
Regular Audits and Compliance	Conduct periodic reviews to ensure adherence to ethical standards and legal compliance.	- Schedule regular audits - Maintain compliance with regulations like GDPR, CCPA	Compliance reviews for conversational agent platforms

Data retention policies are essential for minimizing privacy risks by regularly deleting outdated or unnecessary data. Ensuring user consent is vital to keeping users informed and in control of their personal data, thereby fostering trust and meeting compliance requirements. Transparency practices clarify how user data is handled, enhancing trust and making systems more accessible and user-friendly.

Data minimization focuses on using only essential information, reducing the risk of unnecessary exposure while maintaining system functionality. Providing user data access and control empowers individuals with autonomy over their data, aligning with privacy rights and enhancing user experience. Regular audits and compliance checks are crucial for ensuring continuous adherence to legal and ethical standards, which helps organizations maintain credibility. These practices collectively guide the ethical development of conversational agents, emphasizing the importance of privacy, transparency, and user autonomy in building trustworthy and effective systems.

Table 13 helps guide the ethical development of conversational agents by highlighting core practices that enhance privacy, transparency, and user autonomy. Figure 12 adopts a layered structure to illustrate essential privacy protocols. At the **core layer**, data encryption ensures secure transmission and storage of sensitive information. The **middle layer** focuses on data anonymization, safeguarding user identities by removing or obfuscating personal details. Finally, the **outer layer** implements retention limits, defining time-based rules for data deletion to maintain compliance with privacy standards. Each layer represents an additional level of security, emphasizing a robust and comprehensive approach to privacy and ethical practices.

https://dx.doi.org/10.4314/njt.v44i4.13

Table 13: Ethical and privacy techniques

Technique	Description	Purpose	Examples of Use
Data Retention Policies	Defines how long user data is stored and ensures it is deleted after a set period.	Minimizes unnecessary data storage and reduces privacy risks	Common in chatbots that handle personal information
End-to-End Encryption	Encrypts data during transfer to prevent unauthorized access.	Secures data in transit from interception	Used in messaging apps and customer service bots
Differential Privacy	Adds noise to data to protect individual identities in aggregated datasets.	Ensures anonymity of user data in analytics	Applied in large-scale conversational analytics
Transparency Measures	Provides users with clear information on data usage, model limitations, and privacy policies.	Builds trust and informs users about data practices	Common in virtual assistants and AI-based support systems
Anonymization	Removes identifiable information from user data to protect privacy.	Prevents tracking or identification of users	Frequently used in health and support chatbots
Access Controls	Limits who can access user data within an organization.	Prevents unauthorized data access	Applied in enterprise chatbots with sensitive information

Table 13 presents key practices used to ensure ethical handling of data in conversational agents. This includes strategies for data retention, encryption, and transparency, which are essential for protecting user privacy and fostering trust.

Data Retention Policies ensure data is stored only as long as needed, mitigating privacy risks**.** End-to-End Encryption and Access Controls protect data from unauthorized access, crucial in applications handling sensitive information. Differential Privacy and Anonymization protect individual identities, especially in large datasets used for analysis. Transparency Measures enhance user trust by openly communicating data practices.

This table highlights essential practices for maintaining privacy and ethical standards, guiding developers in implementing responsible conversational AI systems.

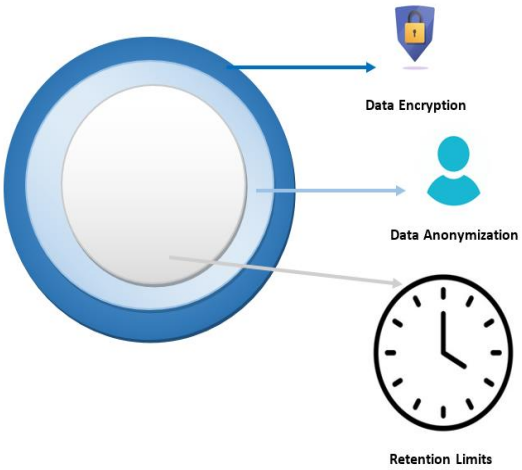

Figure 12: Privacy and ethical framework for conversational agents[116]

Table 13: Summary of key findings across research focus areas

Research Focus Area	Challenges	Recent Advancements	Possible Improvements
Contextual Understanding	Difficulty in maintaining context over long conversations; context drift.	Transformer-based models (e.g., BERT, GPT); memory networks.	Development of multi-turn conversation models; improved context tracking.
Dialogue Coherence	Ensuring responses are relevant and logically follow previous exchanges.	Sequence-to-sequence models; reinforcement learning for dialogue.	Hybrid models combining rule-based and deep learning approaches for consistency.
Scalability and Efficiency	High computational costs limit deployment on resource-constrained devices.	Techniques like pruning, quantization, and distillation.	Enhanced optimization methods for edge computing applications.

https://dx.doi.org/10.4314/njt.v44i4.13

Privacy and Data Security	Risks of data exposure and user identity compromise.	Differential privacy; anonymization techniques; secure data storage.	Development of more robust privacy-preserving AI techniques and user-controlled data settings.
Ethical Transparency	Lack of clarity on model decision-making; issues with bias.	Frameworks for ethical AI; improved transparency policies.	Implementation of more transparent, interpretable models with reduced bias.
User Satisfaction	Difficulty in measuring satisfaction accurately; varied user expectations.	Use of user feedback loops and satisfaction surveys.	Enhanced real-time sentiment analysis and adaptive dialogue systems.
Response Time Optimization	Delays in response generation, especially with complex queries.	Real-time response techniques using caching and lightweight models.	Improved real-time processing methods and model optimization for speed.

3.5 Summary of Key Findings

To consolidate insights, a summary table and radar chart are suggested to provide a holistic view of advancements, challenges, and promising areas for future research.

Table 13 offers a detailed view of each primary research focus area in conversational AI. This table includes associated challenges, recent advancements, and possible areas for improvement, providing a comprehensive snapshot of the state of the field.
Contextual Understanding and Dialogue Coherence focus on improving how agents understand and respond within multi-turn conversations. Scalability and Efficiency address the need for lightweight models, making conversational AI feasible on devices with limited resources. Privacy and Data Security and Ethical Transparency outline ongoing advancements in protecting user information and ensuring ethical AI practices. User Satisfaction and Response Time Optimization emphasize direct user experience, targeting faster, more satisfying interactions. This table provides a quick reference to understand current research challenges and suggests practical improvements to address these needs in future developments.
Figure 13 is a chart comparing contextual models' strengths and weaknesses across dimensions such as scalability, ambiguity resolution, user personalization, and privacy compliance.

The findings outlined in this section reflect the current landscape of contextual understanding in conversational agents, highlighting both the progress made and areas that require further research and innovation. In particular, advancements in transformer and graph-based models offer promising paths, although challenges in computational efficiency, ambiguity handling, and ethical considerations remain pressing. This comprehensive analysis provides a foundation for future research focused on creating scalable, contextually aware, and ethically responsible conversational agents.

- **Scalability:** This column reflects the model's ability to handle increasing workloads or more complex tasks without significant degradation in performance. Transformer-based and hybrid models score highest in scalability.

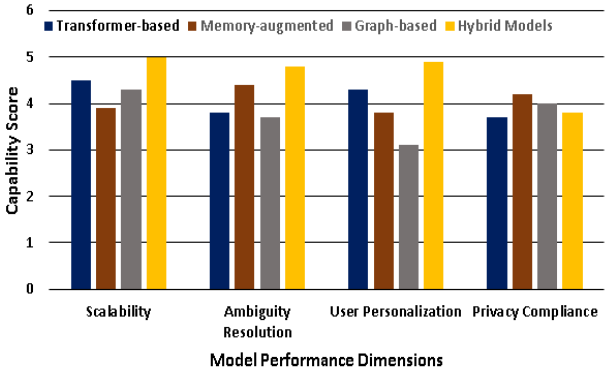

Figure 13: Chart of model capabilities vs. challenges [6]

https://dx.doi.org/10.4314/njt.v44i4.13

Table 14: Chart of model capabilities vs. challenges [6]

Model	Scalability	Ambiguity Resolution	User Personalization	Privacy Compliance
Transformer-based	High (4.5)	Moderate (3.8)	High (4.3)	Moderate (3.7)
Memory-augmented	Moderate (3.9)	High (4.4)	Moderate (3.8)	High (4.2)
Graph-based	High (4.3)	Moderate (3.7)	Low (3.1)	High (4.0)
Hybrid Models	Very High (5.0)	Very High (4.8)	Very High (4.9)	Moderate (3.8)

- **Ambiguity Resolution:** This column shows the model's ability to handle ambiguous user inputs. Memory-augmented and hybrid models excel here, providing clear, accurate responses to unclear queries.

- **User Personalization:** Reflects how well a model can adapt to individual user preferences and behaviors. Hybrid models stand out with high personalization capabilities, while graph-based models lag behind in this dimension.

- **Privacy Compliance:** This dimension evaluates the model's adherence to privacy standards and its ability to ensure user data protection. Memory-augmented and graph-based models perform well in privacy compliance, while transformer-based models show moderate performance.

Figure 14 allows for an easy comparison of how different models perform across multiple critical dimensions. It highlights the strengths and weaknesses of each model type and can guide decisions based on the specific needs of a conversational agent application.

Table 15 also represented in Figure 15, compares six key research areas based on their importance and feasibility, with scores ranging from 1 to 5 for each attribute.

Table 15: Comparison of the importance and feasibility of various research areas[108]

Research Area	Importance (1-5)	Feasibility (1-5)
Scalability enhancements	4.8	4.0
Ethical frameworks	4.5	3.8
Contextual understanding	5.0	4.5
Real-time processing	4.3	3.9
Personalization techniques	4.7	4.2
Data security	4.9	4.1

https://dx.doi.org/10.4314/njt.v44i4.13

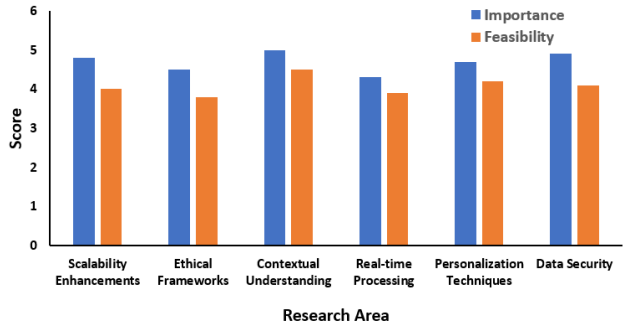

Figure 14: A radar chart illustrating the importance and feasibility of various future research areas identified in recent studies[108]

Scalability enhancements is highly important, with a score of 4.8, but its feasibility is somewhat lower at 4.0. This suggests that while scalability is a priority, it may require further technological development to achieve effectively. Ethical Frameworks hold significant importance, scoring 4.5, but their feasibility is a bit more limited, with a score of 3.8. This reflects the challenges in implementing robust ethical standards in real-world applications. Contextual Understanding is the most important research area, scoring 5.0 for importance, with high feasibility at 4.5. This shows that improving contextual understanding is both crucial and relatively achievable with current technology. Real-time Processing is moderately important, with a score of 4.3, but it faces some challenges in feasibility, with a score of 3.9. This highlights the difficulty of integrating real-time capabilities into conversational agents. Personalization Techniques are considered important, with a score of 4.7, and are moderately feasible, scoring 4.2. While personalizing responses is a key area of focus, it still presents some challenges in practical implementation. Data Security is very important, with a score of 4.9, and has a good level of feasibility at 4.1. This indicates that securing user data is crucial and achievable, though there is still room for improvement. Overall, the table provides insights into areas that require more attention, especially those that are highly important but face challenges in terms of feasibility.

4.0 SUMMARY OF MAJOR FINDINGS

Recent advancements in context modeling techniques highlight the dominance of transformer-based models, such as BERT, GPT-3, and T5, which excel in retaining context over long dialogues by leveraging self-attention to track dependencies across multi-turn conversations. Hybrid models that combine transformers with memory-efficient approaches are gaining traction for balancing accuracy and computational efficiency. Memory-augmented networks are effective in dynamically recalling past interactions, making them ideal for long-term context retention in applications like customer service and therapy chatbots. Graph-based models utilize knowledge graphs to capture structured relationships and entity dependencies, often integrated with transformers for domain-specific applications like healthcare and legal systems. The trend toward combining multiple modeling approaches, such as transformers with memory-augmented or graph-based models, has been significant in advancing contextual understanding.

In contextual understanding, scalability is a major challenge due to the high computational demands of transformers, particularly in real-time deployment scenarios with lengthy and complex dialogues. Techniques such as model pruning, quantization, and knowledge distillation show promise for optimizing scalability. Ambiguity handling also remains a key issue, as agents often struggle to resolve vague or incomplete user references without robust context. Advanced techniques like dual transformer layers and contextual cueing have improved ambiguity resolution. Computational efficiency is another concern, with transformer-based models requiring substantial memory and processing resources, which limits their application in low-resource settings. Maintaining context consistency over long conversations, known as context drift, is a persistent challenge as conversations grow in complexity.

Different context modeling approaches have a significant impact on dialogue coherence, personalization, and user satisfaction. Transformers greatly enhance dialogue coherence by capturing long-range dependencies, while hybrid approaches combining rule-based systems with machine learning are being explored for improving consistency. Personalization is achieved effectively with memory-augmented networks, which adapt to user preferences by retaining long-term data, and with graph-based models integrated with knowledge graphs, enabling personalized responses in domain-specific scenarios. However, personalization introduces privacy challenges, requiring careful handling of user data. Personalized systems consistently lead to higher user satisfaction, with ratings improving more significantly compared to non-personalized systems.

Ethical and computational considerations emphasize the importance of privacy protocols, such as encryp-

tion, anonymization, and adherence to regulations like GDPR, to safeguard user data. Techniques like differential privacy and federated learning are recommended to balance data utility and privacy. Ethical transparency is also crucial, with clear communication of data usage policies shown to improve user trust. Retaining user context raises ethical concerns, particularly regarding consent and the duration of data storage. Best practices include limiting context retention based on session length or explicit user consent and implementing transparency measures such as clear privacy policies and user data controls.

5.0 CONCLUSION

This paper provides an in-depth examination of the complexities and advancements in contextual understanding and modeling techniques. Despite the significant progress made in the field, several critical challenges persist, including the scalability of models, high computational costs, and ethical concerns such as bias and fairness in contextual systems. These challenges underscore the necessity of continued innovation and interdisciplinary collaboration to build robust systems capable of performing effectively in diverse, real-world environments.

The findings highlight the importance of balancing technical innovation with ethical considerations to ensure the development of reliable, fair, and contextually aware systems. This balance is critical for advancing AI technologies that can address pressing societal needs.

5.1 Suggestions for Future Research

Future research should prioritize the development of dynamic and flexible models capable of adapting to complex and evolving contexts across diverse applications. To enable broader adoption, efforts must focus on creating scalable algorithms and architectures that minimize computational demands without sacrificing performance. Addressing bias in data and models is another critical area, as developing equitable and inclusive AI systems requires robust techniques for detecting, mitigating, and preventing such biases.

Additionally, researchers should explore methods for integrating multimodal data, such as text, speech, and visuals, to enhance contextual understanding and decision-making capabilities. Emphasis should also be placed on designing systems that facilitate intuitive human-AI collaboration, empowering users in complex tasks while maintaining transparency and trust. Testing models in diverse, real-world scenarios is essential to validate their adaptability, robustness, and ability to generalize effectively.

Finally, collaboration between researchers, policymakers, and ethicists is necessary to establish governance frameworks that guide the responsible deployment of contextual systems. Such frameworks should address societal implications, ensure accountability, and provide a foundation for ethical AI development. By pursuing these directions, the research community can drive the creation of intelligent systems that are both technically innovative and ethically sound.

REFERENCES

[1] Nadir, Z., Al Lawati, H. M., Mohammed, R. A., Al Subhi, M. and Hossen, A. “SQUbot: Enhancing Student Support Through a Personalized Chatbot System,” *Technologies*, 13 (9), pp. 416, 2025.

[2] Schachner, T., Keller, R. and Wangenheim, F. “Artificial Intelligence-Based Conversational Agents for Chronic Conditions: Systematic Literature Review,” *Journal of Medical Internet Research*, 22 (9), pp. e20701, 2020.

[3] Wahde, M. and Virgolin, M. “Conversational Agents: Theory and Applications,” *Handbook on Computer Learning and Intelligence: Volume 2 – Deep Learning, Intelligent Control and Evolutionary Computation*, pp. 497–544, 2022.

[4] Casheekar, A., Lahiri, A., Rath, K., Prabhakar, K. S. and Srinivasan, K. “A Contemporary Review on Chatbots, AI-Powered Virtual Conversational Agents, ChatGPT: Applications, Open Challenges and Future Research Directions,” *Computer Science Review*, 52, pp. 100632, 2024.

[5] Tudor Car, L., Dhinagaran, D. A., Kyaw, B. M., Kowatsch, T., Joty, S., Theng, Y. L. and Atun, R. “Conversational Agents in Health Care: Scoping Review and Conceptual Analysis,” *Journal of Medical Internet Research*, 22 (8), pp. e17158, 2020.

[6] Diederich, S., Brendel, A. B., Morana, S. and Kolbe, L. “On the Design of and Interaction With Conversational Agents: An Organizing and Assessing Review of Human–Computer

Interaction Research," *Journal of the Association for Information Systems*, 23 (1), pp. 96–138, 2022.

[7] Fareedi, A. A., Ismail, M., Gagnon, S., Ghazanweh, A. and Arooj, Z. "Digital Health Transformation: Leveraging a Knowledge Graph Reasoning Framework and Conversational Agents for Enhanced Knowledge Management," *Systems*, 13 (2), pp. 72, 2025.

[8] Qureshi, F. K. "The Evolution of AI Algorithms: From Rule-Based Systems to Deep Learning," *Frontiers in Artificial Intelligence Research*, 1 (02), pp. 250–288, 2024.

[9] Sharma, D., Sundravadivelu, K., Khengar, J., Thaker, D. J., Patel, S. N. and Shah, P. "Advancements in Natural Language Processing: Enhancing Machine Understanding of Human Language in Conversational AI Systems," *Journal of Computational Analysis and Applications (JoCAAA)*, 33 (06), pp. 713–721, 2024.

[10] Brandtzaeg, P. B. and Følstad, A. "Chatbots: Changing User Needs and Motivations," *Interactions*, 25 (5), pp. 38–43, 2018.

[11] Chaves, A. P. and Gerosa, M. A. "How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design," *International Journal of Human–Computer Interaction*, 37 (8), pp. 729–758, 2021.

[12] He, L., Braggaar, A., Basar, E., Krahmer, E., Antheunis, M. and Wiers, R. "Exploring User Engagement Through an Interaction Lens: What Textual Cues Can Tell Us About Human–Chatbot Interactions," *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, pp. 1–14, 2024.

[13] Ma, Y., Abbas, T. and Gadiraju, U. "ContextBot: Improving Response Consistency in Crowd-Powered Conversational Systems for Affective Support Tasks," *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, pp. 1–14, 2023.

[14] Martins, A., Nunes, I., Lapão, L. and Londral, A. "Unlocking Human-Like Conversations: Scoping Review of Automation Techniques for Personalized Healthcare Interventions Using Conversational Agents," *International Journal of Medical Informatics*, pp. 105385, 2024.

[15] Mariani, M. M., Hashemi, N. and Wirtz, J. "Artificial Intelligence Empowered Conversational Agents: A Systematic Literature Review and Research Agenda," *Journal of Business Research*, 161, pp. 113838, 2023.

[16] Huang, W., Hew, K. F. and Fryer, L. K. "Chatbots for Language Learning—Are They Really Useful? A Systematic Review of Chatbot-Supported Language Learning," *Journal of Computer Assisted Learning*, 38 (1), pp. 237–257, 2022.

[17] Bhattacharjee, A., Williams, J. J., Beltzer, M., Meyerhoff, J., Kumar, H., Song, H. and Kornfield, R. "Investigating the Role of Situational Disruptors in Engagement With Digital Mental Health Tools," *Proceedings of the ACM on Human-Computer Interaction*, 9 (7), pp. 1–35, 2025.

[18] Nuruzzaman, M. and Hussain, O. K. "IntelliBot: A Dialogue-Based Chatbot for the Insurance Industry," *Knowledge-Based Systems*, 196, pp. 105810, 2020.

[19] Asar, S. H., Jalalpour, S. H., Ayoubi, F., Rahmani, M. R. and Rezaeian, M. "PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses," *Journal of Rafsanjan University of Medical Sciences*, 15 (1), pp. 68–80, 2016.

[20] Kusal, S., Patil, S., Choudrie, J., Kotecha, K., Mishra, S. and Abraham, A. "AI-Based Conversational Agents: A Scoping Review From Technologies to Future Directions," *IEEE Access*, 10, pp. 92337–92356, 2022.

[21] Jiao, L., Shao, Y., Sun, L., Liu, F., Yang, S., Ma, W., Li, L., … et al. "Advanced Deep Learning Models for 6G: Overview, Opportunities, and Challenges," *IEEE Access*, 12: pp. 133245–133314, 2024.

[22] Sajun, A. R., Zualkernan, I. and Sankalpa, D. "A Historical Survey of Advances in Transformer Architectures," *Applied Sciences*, 14 (10), pp. 4316, 2024.

[23] Hadi, M. U., Al Tashi, Q., Shah, A., Qureshi, R., Muneer, A., Irfan, M. and Shah, M. "Large Language Models: A Comprehensive Survey of Its Applications, Challenges, Limitations, and Future Prospects," *Authorea Preprints*, 2024.

[24] Yang, B., Wang, L., Wong, D. F., Shi, S. and Tu, Z. "Context-Aware Self-Attention Networks for Natural Language Processing," *Neurocomputing*, 458, pp. 157–169, 2021.

https://dx.doi.org/10.4314/njt.v44i4.13

[25] Sun, Y., Yuan, N. J., Xie, X., McDonald, K. and Zhang, R. "Collaborative Intent Prediction With Real-Time Contextual Data," *ACM Transactions on Information Systems (TOIS)*, 35 (4), pp. 1–33, 2017.

[26] Abu Tami, M., Ashqar, H. I., Elhenawy, M., Glaser, S., and Rakotonirainy, A. "Using Multimodal Large Language Models (MLLMs) for Automated Detection of Traffic Safety-Critical Events," *Vehicles*, 6(3): pp. 1571–1590, 2024.

[27] Liu, Q., Yogatama, D. and Blunsom, P. "Relational Memory-Augmented Language Models," *Transactions of the Association for Computational Linguistics*, 10, pp. 555–572, 2022.

[28] Ma, G., Vo, V. A., Willke, T. L. and Ahmed, N. K. "Memory-Augmented Graph Neural Networks: A Brain-Inspired Review," *IEEE Transactions on Artificial Intelligence*, pp. 2011, 2023.

[29] Ren, X., Chen, T., Nguyen, Q. V. H., Cui, L., Huang, Z. and Yin, H. "Explicit Knowledge Graph Reasoning for Conversational Recommendation," *ACM Transactions on Intelligent Systems and Technology*, 15 (4), pp. 1–21, 2024.

[30] Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G. and Chen, E. "When Large Language Models Meet Personalization: Perspectives of Challenges and Opportunities," *World Wide Web*, 27 (4), pp. 42, 2024.

[31] Gillioz, A., Casas, J., Mugellini, E. and Abou Khaled, O. "Overview of the Transformer-Based Models for NLP Tasks," *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pp. 179–183, 2020.

[32] McTear, M. and Ashurkina, M. "Transforming Conversational AI: Exploring the Power of Large Language Models in Interactive Conversational Agents," *Springer Nature*, 2024.

[33] Noh, J. and Kavuluru, R. "Improved Biomedical Word Embeddings in the Transformer Era," *Journal of Biomedical Informatics*, 120, pp. 103867, 2021.

[34] Metheniti, E. "What Do You Know, BERT? Exploring the Linguistic Competencies of Transformer-Based Contextual Word Embeddings," *Doctoral Dissertation, Université Toulouse le Mirail–Toulouse II*, 2023.

[35] Ghaith, S. "Deep Context Transformer: Bridging Efficiency and Contextual Understanding of Transformer Models," *Applied Intelligence*, 54 (19), pp. 8902–8923, 2024.

[36] Manu, B. K. "A Bridge Between Graph Neural Networks and Transformers: Positional Encodings as Node Embeddings," 2023.

[37] Sharaf Al-deen, H. S., Zeng, Z., Al-Sabri, R. and Hekmat, A. "An Improved Model for Analyzing Textual Sentiment Based on a Deep Neural Network Using Multi-Head Attention Mechanism," *Applied System Innovation*, 4 (4), pp. 85, 2021.

[38] Naseem, U., Razzak, I., Musial, K. and Imran, M. "Transformer-Based Deep Intelligent Contextual Embedding for Twitter Sentiment Analysis," *Future Generation Computer Systems*, 113, pp. 58–69, 2020.

[39] Singh, S., Kumar, M., Kumar, A., Verma, B. K., Abhishek, K. and Selvarajan, S. "Efficient Pneumonia Detection Using Vision Transformers on Chest X-Rays," *Scientific Reports*, 14 (1), pp. 2487, 2024.

[40] Hernández, A. and Amigó, J. M. "Attention Mechanisms and Their Applications to Complex Systems," Entropy, 23 (3), pp. 283, 2021.

[41] He, T., Tan, X., Xia, Y., He, D., Qin, T., Chen, Z. and Liu, T. Y. "Layer-Wise Coordination Between Encoder and Decoder for Neural Machine Translation," *Advances in Neural Information Processing Systems*, 31, pp. 1–??, 2018.

[42] Li, S., Zhang, L., Wang, Z., Wu, D., Wu, L., Liu, Z. and Li, S. Z. "Masked Modeling for Self-Supervised Representation Learning on Vision and Beyond," *arXiv Preprint*, arXiv:2401.00897, 2023.

[43] Asadi, A. and Safabakhsh, R. "The Encoder–Decoder Framework and Its Applications," *Deep Learning: Concepts and Architectures*, pp. 133–167, 2020.

[44] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M. and Gao, J. "Deep Learning-Based Text Classification: A Comprehensive Review," *ACM Computing Surveys (CSUR)*, 54 (3), pp. 1–40, 2021.

[45] Bansal, G., Chamola, V., Hussain, A., Guizani, M. and Niyato, D. "Transforming Conversations With AI—A Comprehensive

https://dx.doi.org/10.4314/njt.v44i4.13

Study of ChatGPT," *Cognitive Computation*, pp. 1–24, 2024.

[46] Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J. and Azam, S. "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," *IEEE Access*, pp. 26839-26874, 2024.

[47] Zhao, T., Wang, S., Ouyang, C., Chen, M., Liu, C., Zhang, J., Yu, L., *et al.* "Artificial Intelligence for Geoscience: Progress, Challenges, and Perspectives," *The Innovation*, 5 (5):, 2024.

[48] Nazir, A. and Wang, Z. "A Comprehensive Survey of ChatGPT: Advancements, Applications, Prospects and Challenges," *Meta-Radiology*, pp. 100022, 2023.

[49] Nassiri, K. and Akhloufi, M. "Transformer Models Used for Text-Based Question Answering Systems," *Applied Intelligence*, 53 (9), pp. 10602–10635, 2023.

[50] Gruetzemacher, R. and Paradice, D. "Deep Transfer Learning and Beyond: Transformer Language Models in Information Systems Research," *ACM Computing Surveys (CSUR)*, 54 (10s), pp. 1–35, 2022.

[51] Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y. and Zhu, J. "Pre-Trained Models: Past, Present and Future," *AI Open*, 2, pp. 225–250, 2021.

[52] Sur, C. "RBN: Enhancement in Language Attribute Prediction Using Global Representation of Natural Language Transfer Learning Technology Like Google BERT," *SN Applied Sciences*, 2 (1), pp. 22, 2020.

[53] Žužemič, J. "Memory-Augmented Neural Networks: Analyzing Memory-Augmented Neural Network Architectures for Incorporating External Memory to Enhance Learning and Reasoning," *Australian Journal of Machine Learning Research & Applications*, 3 (2), pp. 269–278, 2023.

[54] Esteva, A., Kale, A., Paulus, R., Hashimoto, K., Yin, W., Radev, D. and Socher, R. "COVID-19 Information Retrieval With Deep-Learning Based Semantic Search, Question Answering and Abstractive Summarization," *NPJ Digital Medicine*, 4 (1), pp. 68, 2021.

[55] Ali, M., Agrawal, A. and Roy, K. "RAMANN: In-SRAM Differentiable Memory Computations for Memory-Augmented Neural Networks," *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, pp. 61–66, 2020.

[56] Wu, C. S. "Learning to Memorize in Neural Task-Oriented Dialogue Systems," *arXiv Preprint*, arXiv:1905.07687, 2019.

[57] Belser, C. A. "Comparison of Natural Language Processing Models for Depression Detection in Chatbot Dialogues," *Doctoral Dissertation, Massachusetts Institute of Technology*, 2023.

[58] Guo, J., Li, N., Qi, J., Yang, H., Li, R., Feng, Y. and Xu, M. "Empowering Working Memory for Large Language Model Agents," *arXiv Preprint*, arXiv:2312.17259, 2023.

[59] Sakhinana, S. S., Aripirala, K. S. S., Gupta, S. and Runkana, V. "Joint Hypergraph Rewiring and Memory-Augmented Forecasting Techniques in Digital Twin Technology," *arXiv Preprint*, arXiv:2408.12634, 2024.

[60] Lv, Z., Qiao, L., Yang, S., Li, J., Lv, H. and Piccialli, F. "Memory-Augmented Neural Networks Based Dynamic Complex Image Segmentation in Digital Twins for Self-Driving Vehicle," *Pattern Recognition*, 132, pp. 108956, 2022.

[61] Adel, A., Ahsan, A., and Davison, C. "ChatGPT Promises and Challenges in Education: Computational and Ethical Perspectives," *Education Sciences*, 14 (8): pp. 814, 2024.

[62] Belser, C. A. "Comparison of Natural Language Processing Models for Depression Detection in Chatbot Dialogues," Doctoral Dissertation, Massachusetts Institute of Technology, 2023.

[63] Guo, J., Li, N., Qi, J., Yang, H., Li, R., Feng, Y. and Xu, M. "Empowering Working Memory for Large Language Model Agents," *arXiv Preprint*, arXiv:2312.17259, 2023.

[64] Sakhinana, S. S., Aripirala, K. S. S., Gupta, S. and Runkana, V. "Joint Hypergraph Rewiring and Memory-Augmented Forecasting Techniques in Digital Twin Technology," *arXiv Preprint*, arXiv:2408.12634, 2024.

[65] Lv, Z., Qiao, L., Yang, S., Li, J., Lv, H. and Piccialli, F. "Memory-Augmented Neural

Networks Based Dynamic Complex Image Segmentation in Digital Twins for Self-Driving Vehicle," *Pattern Recognition*, 132, pp. 108956, 2022.

[66] Liang, F., Qian, C., Yu, W., Griffith, D. and Golmie, N. "Survey of Graph Neural Networks and Applications," *Wireless Communications and Mobile Computing*, 2022 (1), pp. 9261537, 2022.

[67] Jia, Y., Wang, J., Shou, W., Hosseini, M. R. and Bai, Y. "Graph Neural Networks for Construction Applications," *Automation in Construction*, 154, pp. 104984, 2023.

[68] Gao, C., Zheng, Y., Li, N., Li, Y., Qin, Y., Piao, J. and Li, Y. "A Survey of Graph Neural Networks for Recommender Systems: Challenges, Methods and Directions," *ACM Transactions on Recommender Systems*, 1 (1), pp. 1–51, 2023.

[69] Javdani Rikhtehgar, D., Tiddi, I., Wang, S., Schlobach, S. and Heylen, D. "Assessing the HI-Ness of Virtual Heritage Applications With Knowledge Engineering," *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pp. 173–187, 2024.

[70] Nazi, Z. A. and Peng, W. "Large Language Models in Healthcare and Medical Domain: A Review," *Informatics*, 11 (3), pp. 57, 2024.

[71] Theodorakopoulos, L., Theodoropoulou, A., and Stamatiou, Y. "A State-of-the-Art Review in Big Data Management Engineering: Real-Life Case Studies, Challenges, and Future Research Directions," *Eng*, 5(3): pp. 1266–1297, 2024.

[72] Rane, N. L., Mallick, S. K., Kaya, O. and Rane, J. "Machine Learning and Deep Learning Architectures and Trends: A Review," *Applied Machine Learning and Deep Learning: Architectures and Techniques*, pp. 1–38, 2024.

[73] Singh, S. and Mahmood, A. "The NLP Cookbook: Modern Recipes for Transformer-Based Deep Learning Architectures," *IEEE Access*, 9, pp. 68675–68702, 2021.

[74] Ray, P. P. "ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope," *Internet of Things and Cyber-Physical Systems*, 3: pp. 121–154, 2023.

[75] Keyvan, K. and Huang, J. X. "How to Approach Ambiguous Queries in Conversational Search: A Survey of Techniques, Approaches, Tools, and Challenges," *ACM Computing Surveys*, 55(6): pp. 1–40, 2022.

[76] Axelsson, A., Buschmeier, H. and Skantze, G. "Modeling Feedback in Interaction with Conversational Agents—A Review," *Frontiers in Computer Science*, 4: p. 744574, 2022.

[77] Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E. and Cieliebak, M. "Survey on Evaluation Methods for Dialogue Systems," *Artificial Intelligence Review*, 54: pp. 755–810, 2021.

[78] Kusal, S., Patil, S., Choudrie, J., Kotecha, K., Mishra, S. and Abraham, A. "AI-Based Conversational Agents: A Scoping Review from Technologies to Future Directions," *IEEE Access*, 10: pp. 92337–92356, 2022.

[79] Chow, J. C. and Li, K. "Ethical Considerations in Human-Centered AI: Advancing Oncology Chatbots Through Large Language Models," *JMIR Bioinformatics and Biotechnology*, 5: p. e64406, 2024.

[80] Edu, J., Mulligan, C., Pierazzi, F., Polakis, J., Suarez-Tangil, G. and Such, J. "Exploring the Security and Privacy Risks of Chatbots in Messaging Services," *Proceedings of the 22nd ACM Internet Measurement Conference*, pp. 581–588, 2022.

[81] Azad, M. A., Arshad, J., Mahmoud, S., Salah, K. and Imran, M. "A Privacy-Preserving Framework for Smart Context-Aware Healthcare Applications," *Transactions on Emerging Telecommunications Technologies*, 33(8): p. e3634, 2022.

[82] Gumusel, E. "A Literature Review of User Privacy Concerns in Conversational Chatbots: A Social Informatics Approach," *Journal of the Association for Information Science and Technology*, (online first), 2024.

[83] Abdollahi, M., Yeganli, S. F., Baharloo, M. and Baniasadi, A. "Hardware Design and Verification with Large Language Models: A Scoping Review, Challenges, and Open Issues," *Electronics*, 14(1): p. 120, 2024.

[84] Allouch, M., Azaria, A. and Azoulay, R. "Conversational Agents: Goals, Technologies, Vision and Challenges," *Sensors*,

https://dx.doi.org/10.4314/njt.v44i4.13

21(24): p. 8448, 2021.

[85] Ferrag, M. A., Alwahedi, F., Battah, A., Cherif, B., Mechri, A., and Tihanyi, N. "Generative AI and Large Language Models for Cyber Security: All Insights You Need," *SSRN Electronic Journal*, 2024

[86] Roller, S., Boureau, Y. L., Weston, J., Bordes, A., Dinan, E., Fan, A., … and Williamson, M. "Open-Domain Conversational Agents: Current Progress, Open Problems, and Future Directions," arXiv preprint arXiv:2006.12442, 2020.

[87] Petroşanu, D. M. and Pîrjan, A. "Exploring Compression Strategies for Large Language Models Towards Efficient Artificial Intelligence Implementations," *Journal of Information Systems & Operations Management*, 18(1): pp. 225–260, 2024.

[88] Bibi, U., Mazhar, M., Sabir, D., Butt, M. F. U., Hassan, A., Ghazanfar, M. A., … and Abdul, W. "Advances in Pruning and Quantization for Natural Language Processing," *IEEE Access*, (online first), 2024.

[89] Li, X., Yang, H., Yang, C. and Zhang, W. "Efficient Medical Knowledge Graph Embedding: Leveraging Adaptive Hierarchical Transformers and Model Compression," *Electronics*, 12(10): p. 2315, 2023.

[90] Yang, C., Zhu, Y., Lu, W., Wang, Y., Chen, Q., Gao, C., … and Chen, Y. "Survey on Knowledge Distillation for Large Language Models: Methods, Evaluation, and Application," *ACM Transactions on Intelligent Systems and Technology*, (online first), 2024.

[91] Chung, K., & Park, R. C. (2019). "Chatbot-based heathcare service with a knowledge base for cloud computing". *Cluster Computing*, *22*, 1925-1937.

[92] Chung, K. and Park, R. C. "Chatbot-Based Healthcare Service with a Knowledge Base for Cloud Computing," *Cluster Computing*, 22: pp. 1925–1937, 2019.

[93] Liang, Q., Shenoy, P., & Irwin, D. (2020, Liang, Q., Shenoy, P. and Irwin, D. "AI on the Edge: Characterizing AI-Based IoT Applications Using Specialized Edge Architectures," *Proceedings of the IEEE International Symposium on Workload Characterization (IISWC)*, pp. 145–156, 2020.

[94] Gunnam, G. R., Inupakutika, D., Mundlamuri, R., Kaghyan, S., and Akopian, D. "Assessing Performance of Cloud-Based Heterogeneous Chatbot Systems and A Case Study," *IEEE Access*, 2024.

[95] Yadav, A., Patel, A., and Shah, M. "A Comprehensive Review on Resolving Ambiguities in Natural Language Processing," *AI Open*, 2: pp. 85–92, 2021.

[96] Nassiri, K., and Akhloufi, M. "Transformer Models Used for Text-Based Question Answering Systems," *Applied Intelligence*, 53(9): pp. 10602–10635, 2023.

[97] Baek, J., Chandrasekaran, N., Cucerzan, S., Herring, A., and Jauhar, S. K. "Knowledge-Augmented Large Language Models for Personalized Contextual Query Suggestion," *Proceedings of the ACM on Web Conference 2024*, pp. 3355–3366, 2024.

[98] Kocaballi, A. B., Berkovsky, S., Quiroz, J. C., Laranjo, L., Tong, H. L., Rezazadegan, D., … and Coiera, E. "The Personalization of Conversational Agents in Health Care: Systematic Review," *Journal of Medical Internet Research*, 21(11): p. e15360, 2019.

[99] Dobbala, M. K., and Lingolu, M. S. S. "Conversational AI and Chatbots: Enhancing User Experience on Websites," *American Journal of Computer Science and Technology*, 10(3): pp. 62–70, 2024.

[100] Skjæveland, M. G., Balog, K., Bernard, N., Łajewska, W., and Linjordet, T. "An Ecosystem for Personal Knowledge Graphs: A Survey and Research Roadmap," *AI Open*, 5: pp. 55–69, 2024.

[101] Izadi, S., and Forouzanfar, M. "Error Correction and Adaptation in Conversational AI: A Review of Techniques and Applications in Chatbots," *AI*, 5(2): pp. 803–841, 2024.

[102] Ahmed, M., Khan, H., Iqbal, T., Alarfaj, F. K., Alomair, A., and Almusallam, N. "On Solving Textual Ambiguities and Semantic Vagueness in MRC Based Question Answering Using Generative Pre-Trained Transformers," *PeerJ Computer Science*, 9: p. e1422, 2023.

[103] Li, L., Li, J., Wang, H., and Nie, J. "Application of the Transformer Model Algorithm in Chinese Word Sense Disambiguation: A Case Study in Chinese Language," *Scientific Reports*, 14(1): p. 6320, 2024.

[104] Salammagari, A. R. R., and Srivastava, G.

"Adaptive Chatbots: Enhancing User Experience Through Interactive Learning and Dynamic Response Refinement," 2024.

[105] Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G., … and Chen, E. "When Large Language Models Meet Personalization: Perspectives of Challenges and Opportunities," *World Wide Web*, 27(4): p. 42, 2024.

[106] Shumanov, M., and Johnson, L. "Making Conversations with Chatbots More Personalized," *Computers in Human Behavior*, 117: p. 106627, 2021.

[107] Nama, P. "AI-Powered Mobile Applications: Revolutionizing User Interaction Through Intelligent Features and Context-Aware Services," 2023.

[108] Tufail, S., Riggs, H., Tariq, M., and Sarwat, A. I. "Advancements and Challenges in Machine Learning: A Comprehensive Review of Models, Libraries, Applications, and Algorithms," *Electronics*, 12(8): p. 1789, 2023.

[109] Henderson, P., Sinha, K., Angelard-Gontier, N., Ke, N. R., Fried, G., Lowe, R., and Pineau, J. "Ethical Challenges in Data-Driven Dialogue Systems," *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 123–129, 2018.

[110] Alom, N. B. "A Comprehensive Analysis of Customer Behavior Analytics, Privacy Concerns, and Data Protection Regulations in the Era of Big Data and Machine Learning," *International Journal of Applied Machine Learning and Computational Intelligence*, 14(5): pp. 21–40, 2024.

[111] Mai, N. D., Lee, B. G., and Chung, W. Y. "Affective Computing on Machine Learning-Based Emotion Recognition Using a Self-Made EEG Device," Sensors, 21(15): p. 5135, 2021.

[112] Giordani, J. "Mitigating Chatbots AI Data Privacy Violations in the Banking Sector: A Qualitative Grounded Theory Study," *European Journal of Applied Science, Engineering and Technology*, 2(4): pp. 14–65, 2024.

[113] Khurana, R. "Implementing Encryption and Cybersecurity Strategies Across Client, Communication, Response Generation, and Database Modules in E-Commerce Conversational AI Systems," *International Journal of Information and Cybersecurity*, 5(5): pp. 1–22, 2021.

[114] Bhardwaj, V., Khan, S. S., Singh, G., Patil, S., Kuril, D., and Nahar, S. "Risks for Conversational AI Security," *Conversational Artificial Intelligence*, pp. 557–587, 2024.

[115] Acquisti, A., Adjerid, I., Balebako, R., Brandimarte, L., Cranor, L. F., Komanduri, S., … and Wilson, S. "Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online," *ACM Computing Surveys (CSUR)*, 50(3): pp. 1–41, 2017.

[116] Mozafari, N., Weiger, W. H., and Hammerschmidt, M. "Trust Me, I'm a Bot–Repercussions of Chatbot Disclosure in Different Service Frontline Settings," *Journal of Service Management*, 33(2): pp. 221–245, 2022.

[117] Choudhury, A., and Shamszare, H. "Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis," *Journal of Medical Internet Research*, 25: p. e47184, 2023.

[118] Garcia Valencia, O. A., Suppadungsuk, S., Thongprayoon, C., Miao, J., Tangpanithandee, S., Craici, I. M., and Cheungpasitporn, W. "Ethical Implications of Chatbot Utilization in Nephrology," *Journal of Personalized Medicine*, 13(9): p. 1363, 2023.

[119] Gemiharto, I. "Analysis of Personal Data Preservation Policy in Utilizing AI-Based Chatbot Applications in Indonesia," *Medium*, 12(1): pp. 63–78, 2024.

[120] Dewitte, P. "Better Alone Than in Bad Company: Addressing the Risks of Companion Chatbots Through Data Protection by Design," *Computer Law & Security Review*, 54: p. 106019, 2024.

https://dx.doi.org/10.4314/njt.v44i4.13