International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 11 | Nov – 2025 DOI: 10.55041/ISJEM05214
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Survey on LLM-Powered Chatbots: Architectures, Applications, Challenges, and Future Directions

Varun Bantia[1], Prathyusha K[1], Venkatashiva Reddy[1], and Dr. Vishwanath Y[2]

[1]Department of Computer Science, [Presidency University ,Banglore], India
[2]Professor, Department of Computer Science, [Presidency University ,Banglore], India

Abstract

This survey analyzes nine recent research works in the domain of large language model (LLM)-powered chatbots, covering their architectures, applications, advantages, limitations, and open challenges. Traditional chatbots relied on rule-based or retrieval-based sys- tems, which limited flexibility, adaptability, and scala- bility. With the rise of LLMs such as GPT-3, GPT-4, and LLaMA, conversational agents have become more interactive, context-aware, and capable of performing complex tasks.

The nine studies reviewed in this paper cover do- mains including healthcare, education, nutrition, cy- bersecurity, interdisciplinary research, and sustain- ability. Each paper introduces unique architec- tural innovations—from retrieval-augmented genera- tion (RAG) in clinical decision-making, to multimodal orchestration for interdisciplinary research assistants. Collectively, they demonstrate the versatility of LLM- powered chatbots but also highlight persisting limita- tions such as hallucination, explainability gaps, bias, privacy risks, and unsustainable compute usage.

This survey contributes by providing:

- A taxonomy and classification of the nine works by domain, architecture, and technique.
- A comparative analysis summarizing strengths, limitations, and experimental contexts.
- A synthesis of challenges and open research ques- tions.
- Recommendations for future research directions in chatbot development.

1 Introduction

1.1 Motivation

The last three years (2023–2025) have seen a dra- matic surge in the adoption of LLM-powered chatbots. Within two months of its release, ChatGPT reached 100 million active users, making it the fastest-growing consumer application in history. Institutions in health- care, education, and industry quickly recognized the potential of conversational AI to scale expertise, de- mocratize access to knowledge, and improve human– computer interaction.

Despite their success, LLM-powered chatbots raise serious questions:

- Can they be trusted in high-stakes domains such as medicine or law?
- How do we ensure ethical, transparent, and unbi- ased responses?
- Can they be deployed sustainably given their car- bon footprints?

1.2 Why These Nine Papers?

The nine papers surveyed here were chosen because they represent recent (2024–2025), high-impact stud- ies addressing chatbot design in different domains. To- gether, they illustrate a wide spectrum of approaches: clinical assistants (Rau et al., Podoreanu et al.), dig- ital tutors (Alsafari et al., Neumann et al., Ilagan & Ilagan), nutrition advisors (Yang et al.), cybersecu- rity monitors (Shafee et al.), interdisciplinary research tools (Forootani et al.), and sustainability assessments (Jiang et al.).

1.3 Research Questions

This survey addresses the following guiding questions:

1. **Architectural Trends:** What design approaches (e.g., RAG, multimodal orchestration) dominate recent chatbot research?
2. **Applications & Impacts:** How are chatbots be- ing adapted for healthcare, education, cybersecu- rity, and other sectors?

3. **Limitations & Gaps:** What challenges persist across domains (e.g., hallucination, bias, privacy)?

4. **Future Directions:** What research frontiers (e.g., multimodal reasoning, green AI, personal- ization) are most promising?

2 Background

2.1 Transformers and LLMs

The backbone of modern chatbots is the transformer architecture (Vaswani et al., 2017), which uses self- attention to process text sequences in parallel. Models like GPT-4 are autoregressive LLMs trained on trillions of tokens, enabling few-shot and zero-shot learning.

2.2 Evolution of Chatbots

•Rule-based chatbots (1960s–2000s): ELIZA and AIML systems; simple pattern matching.

•Retrieval-based chatbots (2010–2018): Matching queries with FAQs or predefined answers.

•Neural seq2seq chatbots (2016–2020): Encoder-decoder models with limited coherence.

•LLM-powered chatbots (2020–present): Contex- tual, generative, and domain-adaptable, with in- creasing use of retrieval augmentation for factual grounding.

2.3 Key Concepts

•**Retrieval-Augmented Generation (RAG):** Combines knowledge retrieval with LLM reason- ing.

•**Multimodality:** Processing beyond text (voice, images, structured data).

•**Explainability:** Ability to justify or cite outputs.

•**Sustainability:** Measuring compute and carbon footprints of training/inference.

3 Related Surveys

Earlier surveys largely focused on traditional chatbots:

•Shawar & Atwell (2007): Reviewed AIML chat- bots.

•Jain et al. (2018): Examined ML-based dialogue systems.

•Xu et al. (2020): Covered pre-LLM conversational AI.

What differentiates this work:

• We focus exclusively on LLM-driven systems (2024–2025).

• We include six application domains, whereas most past surveys were single-domain (e.g., customer service, healthcare only).

• We integrate sustainability as a dimension of analysis—absent from earlier reviews.

4 Taxonomy of the Nine Papers

We classify the works along three dimensions:

By Domain:

• Healthcare: Rau et al. (accGPT-4), Podoreanu et al. (Elderly Care).

• Education: Alsafari et al., Neumann et al. (MoodleBot), Ilagan & Ilagan (Virtual Policy Agent).

• Nutrition: Yang et al. (ChatDiet).

• Cybersecurity: Shafee et al. (OSINT Chatbot).

• Research: Forootani et al. (Bio-Eng-LLM Assist).

• Sustainability: Jiang et al. (Carbon Footprints).

By Technique:

• LLM-only: Elderly chatbot, cybersecurity chat- bot.

• Hybrid LLM + RAG: accGPT-4, MoodleBot.

• Multimodal orchestration: ChatDiet, Bio-Eng-LLM Assist.

By Architecture:

• Encoder-decoder: GPT-4, Vicuna.

• Retrieval-Augmented: accGPT-4, MoodleBot.

• Multimodal pipelines: Bio-Eng-LLM Assist.

5 Detailed Review of the Nine Papers

5.1 Healthcare-Oriented Papers

Rau et al. (2024): accGPT-4 — This work addresses the challenge of evidence-based medical decision-making in radiology. Traditional chatbots of- ten hallucinate medical information, which can have dangerous consequences in clinical settings. Rau et al. mitigate this by embedding the American Col- lege of Radiology (ACR) guidelines into a retrieval- augmented pipeline that feeds into GPT-4. In practice, this means the chatbot can cite authoritative recom- mendations while still producing natural, flexible dia- logue. A major strength of this work lies in its abil- ity to provide clinically grounded justifications, mov- ing beyond generic text generation. However, a critical limitation is the need for constant updates to medi- cal guidelines, which requires careful pipeline mainte- nance and version control. Furthermore, the reliance on GPT-4 makes the system expensive and potentially inaccessible to smaller clinics.

Podoreanu et al. (2025): Chatbot for Mild Cognitive Impairment — This study targets a vul- nerable demographic: elderly patients with mild cogni- tive impairment (MCI). The chatbot analyzes linguistic biomarkers—patterns in speech and text that may in- dicate cognitive decline. Its dual purpose is therapeu- tic (providing conversational support) and diagnostic (flagging potential deterioration for clinicians). The in- novation lies in applying LLM conversational flow to a healthcare monitoring context. Yet, several limitations arise: (1) limited clinical trials mean we cannot gener- alize its effectiveness, (2) ethical issues exist around monitoring vulnerable populations, and (3) data pri- vacy concerns are particularly sensitive in healthcare. Nevertheless, it highlights the potential of chatbots as digital companions in elderly care.

5.2 Education-Oriented Papers

Alsafari et al. (2024): Teaching Assistants — This study compares intent-based chatbots with LLM- powered assistants. The authors demonstrate that while intent-based systems are predictable and con- trollable, they lack flexibility. LLM-based systems, by contrast, can handle a much wider range of stu- dent queries, adapting to novel contexts. A notable strength of this paper is its empirical comparison, which provides a clear benchmark for institutions de- ciding whether to adopt LLMs. However, the study also acknowledges risks: LLMs may provide overcon- fident but incorrect answers, raising issues of trust in educational environments. **Neumann et al. (2025): MoodleBot** — MoodleBot represents an integration of chatbots into Learning Management Systems (LMS). By connecting an LLM to structured course data, MoodleBot provides 24/7 support for students, an- swering questions about lecture content, assignments, and databases. Its architecture exemplifies a retrieval- augmented system, where the LLM is grounded in structured educational material. The strength of this work lies in its practical deployment and its impact on student engagement. The limitation, however, is that its effectiveness depends heavily on the quality of course materials; poor or outdated resources lead to poor chatbot performance. **Ilagan & Ilagan (2024): Policy Support Agent** — This study focuses on a specialized educational support task: guiding students through university policies. The authors use few-shot prompting and chain-of-thought (CoT) reasoning to improve accuracy. A unique strength is its attention to administrative queries, often overlooked in chatbot design. However, CoT reasoning still fails in complex or ambiguous cases, limiting reliability. This highlights the difficulty of using LLMs for tasks requiring formal legal or policy interpretation.

Case Study: Education-Oriented and Healthcare-Oriented Chatbots

The domains of healthcare and education provide two of the most socially impactful applications of LLM- powered chatbots. This case study synthesizes the works of Rau et al. [1], Podoreanu et al. [2], Alsa- fari et al. [3], Neumann et al. [4], and Ilagan & Ilagan [5], focusing on their contributions, similarities, and domain-specific challenges.

Healthcare-Oriented Case Study. In healthcare, chatbots serve either as clinical decision-support tools or as patient-facing assistants. Rau et al. [1] intro- duced accGPT-4, which integrates American College of Radiology (ACR) guidelines into a retrieval-augmented GPT-4 pipeline to provide evidence-based recommen- dations. This ensures higher reliability in radiology de- cisions, though it requires continuous updates of med- ical guidelines. Podoreanu et al. [2], in contrast, fo- cused on elderly patients with mild cognitive impair- ment, leveraging linguistic biomarkers from conversa- tions to detect early cognitive decline. While promis- ing for preventive healthcare, this raises privacy and ethical concerns due to the sensitive nature of patient monitoring. Together, these works demonstrate that healthcare chatbots can enhance both *decision-making accuracy* and *patient monitoring*, but clinical valida- tion and data protection remain critical bottlenecks.

Education-Oriented Case Study. Educational chatbots aim to expand access to learning and admin- istrative support. Alsafari et al. [3] compared intent- based chatbots with LLM-powered assistants, showing that LLMs outperform traditional systems in flexibility but pose risks of overconfidence in incorrect answers. Neumann et al. [4] developed MoodleBot, a retrieval- augmented assistant embedded in an LMS, which pro- vided round-the-clock course support but struggled when source materials were incomplete or outdated. Ilagan & Ilagan [5] proposed a chatbot for navigating university policies using few-shot prompting and chain- of-thought reasoning. While effective in handling ad-

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 11 | Nov – 2025 DOI: 10.55041/ISJEM05214
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

ministrative queries, it was limited in cases requiring nuanced interpretation of policy documents. Collec- tively, these studies highlight the strengths of LLM- based chatbots in *scalability, adaptability, and person- alization*, while underscoring the challenges of content reliability and error management.

Cross-Domain Insights. Both domains face over- lapping challenges in *trust, explainability, and privacy*, but the risks manifest differently: in healthcare, er- rors may affect patient safety, while in education, they may mislead or frustrate learners. The case study re- veals that success in both domains requires domain- specific grounding (e.g., guidelines in healthcare, cu- rated course materials in education), coupled with mechanisms to mitigate hallucinations and bias.

Future Opportunities. Healthcare and education chatbots can benefit from hybrid designs that com- bine LLM reasoning with symbolic models, as well as from privacy-preserving methods like federated learn- ing. These case studies confirm that while LLM- powered chatbots are already reshaping practice, their real-world adoption depends on overcoming reliability, ethics, and sustainability challenges.

5.3 Technical/Architectural Papers

Yang et al. (2024): ChatDiet — ChatDiet is a personalized nutrition assistant that integrates two knowledge sources: (1) an individual's personal health data (e.g., age, weight, health conditions), and (2) population-level nutritional models. These are or- chestrated by an LLM that generates tailored diet plans. This hybrid framework represents an important step towards personalized digital health assistants. A strength is its ability to bridge individual and general knowledge. A major limitation, however, is privacy: user health data is highly sensitive, and centralizing it in an LLM system raises regulatory challenges.

Shafee et al. (2025): Cybersecurity Chatbot — This paper explores the use of GPT-4 for OSINT- based cyber threat awareness. The chatbot processes public data streams and classifies threats. The au- thors show GPT-4 excels at text classification tasks but struggles with named entity recognition (e.g., ex- tracting specific threat actor names). The strength of this work lies in its security application, an underex- plored domain for chatbots. Its limitation highlights the broader weakness of LLMs in fine-grained extrac- tion tasks, where symbolic or traditional ML methods might outperform generative models.

5.4 Other Domains

Forootani et al. (2025): Bio-Eng-LLM Assist — This study introduces a modular, multimodal chatbot platform for interdisciplinary research. Unlike single-domain assistants, it supports text, voice, and im- ages, and orchestrates multiple models (LLMs, diffu- sion models, knowledge retrieval). The strength here is its versatility, allowing researchers in bioengineering to collaborate with AI across modalities. However, com- plexity is a major barrier—integrating and maintaining multiple pipelines is resource-intensive.

Jiang et al. (2024): Sustainability of LLM Chatbots — Unlike the other works, this paper takes a meta perspective: studying the energy and carbon footprints of deploying LLM chatbots. The authors ar- gue that without interventions, the exponential scaling of LLMs will make them environmentally unsustain- able. The strength of this study is its novelty—few works connect sustainability with chatbots. A limi- tation is the lack of concrete mitigation strategies; it highlights the problem but does not propose detailed solutions.

6 Challenges and Open Issues

Despite progress, several challenges persist:

6.1 Hallucination and Reliability

LLMs still generate plausible but false information. In healthcare and policy contexts, hallucinations could have severe consequences. Even with Retrieval- Augmented Generation (RAG) pipelines, hallucina- tions are not eliminated but only reduced.

6.2 Explainability and Transparency

Most chatbots are still black-box systems. Only accGPT-4 attempts explainability by citing guidelines, but even that remains partial. Users often cannot tell why a chatbot gave a certain answer.

6.3 Bias and Fairness

LLMs inherit biases from their training data, which can lead to discriminatory outputs. For example, ChatDiet could unintentionally recommend diets unsuitable for certain populations if underlying data is biased.

6.4 Privacy and Security

Sensitive domains like healthcare and education involve data that must comply with GDPR, HIPAA, and other privacy regulations. ChatDiet and Elderly Care chat- bots face significant data security challenges, as do any systems that centralize personal health records.

ISSN: 2583-6129
DOI: 10.55041/ISJEM05214

6.5 Domain Adaptation and Maintenance

Many chatbots rely on constantly changing knowledge bases. Medical guidelines, cybersecurity threats, and academic policies all evolve, requiring expensive up- dates and retraining.

6.6 Sustainability

Jiang et al. show that the compute requirements of LLMs are unsustainable in the long run. The envi- ronmental impact of scaling models remains an open issue.

6.7 User Trust and Acceptance

Beyond technical performance, users may hesitate to trust chatbots, particularly in sensitive domains (e.g., elderly patients or students seeking academic guid- ance). Cultural factors and perceptions of AI strongly influence adoption.

7 Future Directions

The surveyed works suggest several directions for ad- vancing chatbot research:

7.1 Hybrid Architectures

Future systems should combine LLMs with symbolic reasoning, causal models, and domain-specific knowl- edge graphs. This could reduce hallucinations and in- crease explainability.

7.2 Green AI and Sustainability

Efficiency improvements are urgent. Promising strate- gies include model pruning, knowledge distillation, federated learning, and edge deployment. Exploring smaller but more specialized models may also help.

7.3 Privacy-Preserving Personalization

Future systems must integrate federated learning, on-device inference, and differential privacy to protect sen- sitive user data. ChatDiet and Elderly Care assistants especially need such safeguards.

7.4 Multimodal Expansion

Bio-Eng-LLM Assist demonstrates the value of multi- modal input (text, voice, image). Extending this to ed- ucation (e.g., diagrams + explanations) and healthcare (e.g., X-rays + reports) could unlock powerful applica- tions.

7.5 Ethical and Regulatory Standards

Clear governance frameworks are needed. This includes AI audits, transparent documentation, and compliance certifications for chatbot deployments in healthcare, education, and beyond.

8 Conclusion

The nine recent works reviewed in this survey col- lectively highlight the transformative potential of LLM-powered chatbots. In healthcare, they promise evidence-based decision support; in education, they of- fer scalable and adaptive tutoring; in cybersecurity, they provide new tools for threat monitoring; in nu- trition and research, they enable personalized and in- terdisciplinary collaboration.

At the same time, they share common limitations: hallucinations, lack of explainability, biases, privacy risks, sustainability concerns, and cultural barriers to adoption. Addressing these will require hybrid archi- tectures, green AI strategies, privacy-preserving meth- ods, multimodal expansion, and ethical governance.

Ultimately, LLM chatbots are not just tools for au- tomating conversation—they are becoming partners in knowledge work. If responsibly designed, they can move from “assistants” to reliable collaborators, shap- ing the future of healthcare, education, research, and beyond.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 11 | Nov – 2025 DOI: 10.55041/ISJEM05214
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Table 1: Comparative Analysis of the Nine Papers

Paper	Domain	Architecture	Dataset/Setting	Strengths	Limitations	Performance
Rau et al.	Healthcare	RAG + GPT-4	Radiology guidelines	Evidence-based outputs	Costly KB updates	High accuracy but expensive
Podoreanu et al.	Healthcare	LLM biomarkers	Elderly speech	Early detection	Needs clinical trials	Promising diagnos- tic support
Alsafari et al.	Education	Intent vs LLM	Student tasks	Clear comparison	Scalability issues	LLMs outperform intent-based sys- tems
Neumann et al.	Education	LMS + RAG	Moodle courses	24/7 support	Content limited	Good engagement
Ilagan & Ilagan	Education	Few-shot + CoT	Policy docs	Policy focus	Prone to errors	Moderate accuracy
Yang et al.	Nutrition	Orchestration	Diet data	Personalized advice	Privacy risk	Effective personal- ization
Shafee et al.	Cybersecurity	GPT-4 monitoring	OSINT feeds	Threat classifica- tion	Weak entity recog- nition	High recall; weak precision
Forootani et al.	Research	Multimodal	Bio/Eng datasets	Interdisciplinary collaboration	Complex design	Strong multimodal reasoning
Jiang et al.	Sustainability	Lifecycle model	Compute data	Novel dimension	No mitigation strategies	Highlights energy costs

References

[1] A. Rau et al., “Enhancing chatbot performance for imaging recommendations: Leveraging GPT-4 and context-awareness,” *European Journal of Ra- diology*, 2024.

[2] B. Podoreanu et al., “Chatbot for patients suf- fering from mild cognitive impairment,” *Procedia Computer Science*, 2025.

[3] B. Alsafari et al., “Towards effective teaching as- sistants: From intent-based chatbots to LLM- powered assistants,” *NLP Journal*, 2024.

[4] A. Neumann et al., “An LLM-Driven Chatbot in Higher Education for Databases and Information Systems,” *IEEE Trans. on Education*, 2025.

[5] J. Ilagan and J. Ilagan, “A prototype of a con- versational university support agent powered by LLMs,” *Procedia Computer Science*, 2024.

[6] Z. Yang et al., “ChatDiet: Empowering person- alized nutrition-oriented food recommender chat- bots,” *Smart Health*, 2024.

[7] S. Shafee et al., “Evaluation of LLM-based chat- bots for OSINT-based Cyber Threat Awareness,” *Expert Systems With Applications*, 2025.

[8] A. Forootani et al., “Bio-Eng-LLM AI Assist: A modular chatbot platform for interdisciplinary re- search,” *SoftwareX*, 2025.

[9] P. Jiang et al., “Preventing immense lifecycle en- ergy and carbon footprints of LLM-powered chat- bots,” *Engineering*, 2024.

[10] H. Touvron et al., “LLaMA: Open and Efficient Foundation Language Models,” *arXiv preprint arXiv:2302.13971*, 2023.

[11] OpenAI, “GPT-4 Technical Report,” *arXiv preprint arXiv:2303.08774*, 2023.

[12] T. B. Brown et al., “Language Models are Few- Shot Learners,” in *NeurIPS*, 2020.

[13] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *NeurIPS*, 2020.

[14] T. Gao, X. Yao, and D. Chen, “Retrieval-augmented generation for conversational AI,” *TACL*, 2023.

[15] R. Bommasani et al., “On the Opportunities and Risks of Foundation Models,” *arXiv preprint arXiv:2108.07258*, 2021.

[16] L. Weidinger et al., “Ethical and social risks of large language models,” *arXiv preprint arXiv:2112.04359*, 2021.

[17] Z. Wang, H. Yu, and H. Zhang, “Survey on bias and fairness in large language models,” *ACM Computing Surveys*, 2023.

[18] J.-B. Alayrac et al., “Flamingo: a Visual Language Model for Few-Shot Learning,” in *NeurIPS*, 2022.

[19] A. Zeng et al., “OpenFlamingo: An open-source framework for training multimodal LLMs,” *arXiv preprint arXiv:2308.01390*, 2023.

[20] E. Strubell, A. Ganesh, and A. McCallum, “En- ergy and Policy Considerations for Deep Learning in NLP,” in *ACL*, 2019.

[21] R. Schwartz, J. Dodge, N. A. Smith, and O. Et- zioni, “Green AI,” *Communications of the ACM*, 2020.

[22] D. Patterson, J. Gonzalez, Q. Le, C. Liang, and J. Dean, “The carbon footprint of machine learning training,” *Communications of the ACM*, 2022.