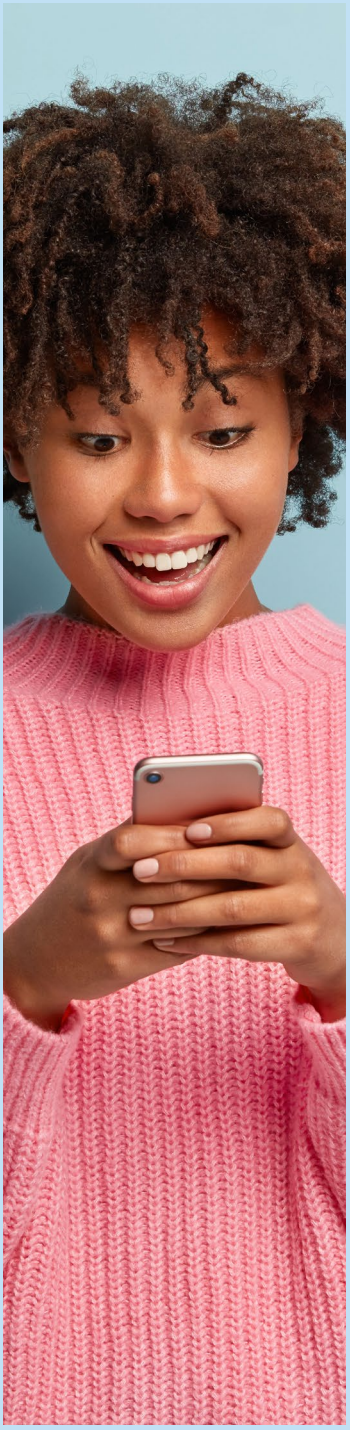
VIEW POINT



IMPACT OF GENERATIVE AI ON CHATBOTS

Abstract

The advent of Generative AI is having a deep impact on Conversational AI. Generative AI refers to the ability of machine learning algorithms to generate content based on the patterns and data they have been trained on.

In the context of chatbots, this means that generative AI can enable chatbots to provide more meaningful and relevant responses to users.

While it has the potential to enhance the functionality of chatbots, there are also risks and challenges that must be considered in its implementation.

The roadmap for Conversational AI and Generative AI's role in it is outlined in this POV.

Infosys®
Navigate your next

Context

Chatbots have become a crucial element in the automation strategies of many organizations. Since their rise in popularity in 2016-2017, there has been significant progress in this technology, with many leading market players such as AWS, Microsoft, IBM, NVIDIA and Google, investing heavily to help adoption in large organizations.

Most organizations initially focused on internal automations, but many have since used chatbots to automate a substantial part of their customer service operations.

The emergence of large language models (LLMs), like GPT-4, FLAN-UL2 20B, Cohere.AI, Bloom, LLaMa, GPT NeoX, Stanford Alpaca, Google PALM and Google BARD has, however, caused doubts amongst enterprise customers who fear their investment in chatbots might be futile. This document explores how the advent of large language models will impact traditional chatbots.

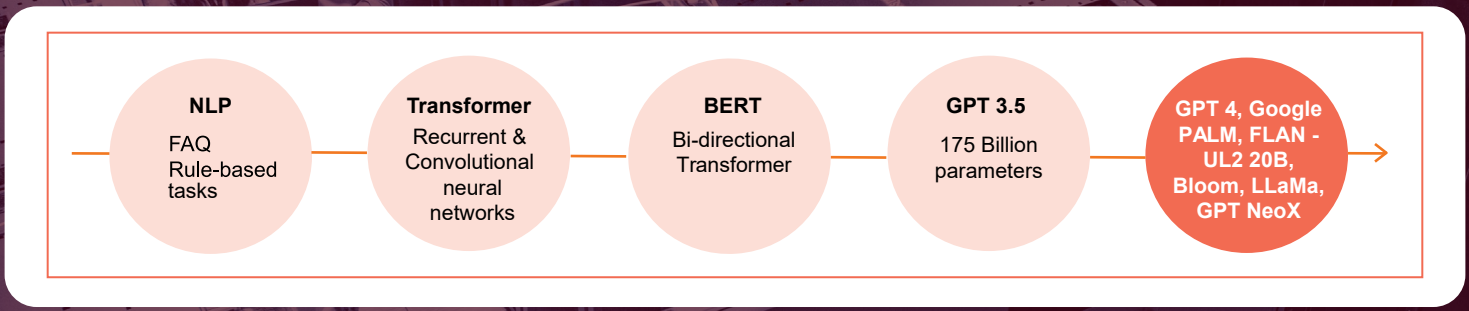


FIG : LLM EVOLUTION

AI-enhanced Chatbots

First it is essential to consider the various types of information that chatbots can process, such as transactional data stored in databases, knowledge articles with instructions on how to troubleshoot problems, legal contracts, frequently asked questions, policy documents and more. Many organizations began with simple FAQ bots, which relied on either intent-based or deep learning-based models to answer basic questions. Chatbots understood the intent, extracted entities and connected with backend systems to retrieve relevant information. Rather than replacing existing Chatbots, organizations can benefit from Generative AI by augmenting existing Chatbots with generative AI capabilities. Hyperscalers offer generative AI capabilities, and organizations can finetune open-source large language models to suit their needs. Lightweight version of models like LLaMa and recently released ALPACA-LORA may also be more manageable and sustainable than larger models like GPT. By adopting a strategy that combines these approaches, organizations can improve overall Chatbot experience

Significant changes expected in Chatbots with Generative AI

Advanced multi-lingual and multi-modal capabilities: Chatbots can now interact in multiple languages and through various modes, such as text, audio, and video, which enhances user experience. Enterprises have started working on solutions with voice cloning and 3D avatars using computer vision and Voice AI to provide multi-modal experience.

Personalized and human-like interactions: Chatbots can provide personalized and lifelike interactions, creating more natural and engaging conversations. For instance, users can inquire about the status of their tickets, jobs, or requests and run scripts through the assistant more naturally and effectively. We can also expect more empathetic and sentiment-based conversations, for e.g., if user is unsatisfied with the product and connecting with Chatbot in a bad mood, the chatbot should be able to detect user's sentiments and respond accordingly.

Enhanced problem-solving abilities: Chatbots can be trained to answer complex questions, beyond basic FAQs, leveraging a deeper understanding of organization's policies and operations. By mining knowledge base articles and generating responses, chatbots can act as an effective first line of defense, deflecting significant number of tickets.

Efficient issue resolution: Chatbots can now be utilized to effectively troubleshoot issues, thereby improving speed and efficiency of resolutions.

Language neutralization: Chatbots can be designed to be language-neutral, allowing users to communicate in their preferred language and receive responses in the same language, breaking down language barriers.

Content generation: In the past, chatbots were limited to sharing or responding based on pre-existing data. With generative AI, chatbots are now capable of creating content. This includes generating reports from data provided to the chatbot, producing training or product documentation, summarizing conversations and sharing results with a live agent and more.

Zero shot learning / Few shot learning: These learning methods allow language models to quickly adapt to new tasks even with less training data or no training data, saving operational efforts in conversation designing & training. This also enables chatbot to learn and improve quickly from conversations.

Conversation Design Prototypes: Generative AI could be used to create conversation flows and chatbot prototypes, reducing overall timeline of requirements gathering phase

Generative AI challenges to be addressed

Trust & Hallucinations: The generated outputs may be incorrect or biased, which raises concerns around trust. Additionally, generative AI may not be suitable for scenarios where deterministic answers are required. It's crucial to have humans in the loop to validate data and mitigate these risks. It would also be difficult to control the output or responses of a Generative AI Model, making it difficult to be used for specific tasks in enterprises and get fixed results. Hallucinations is another issue observed with LLMs. Recent occurrences of chatbots behaving as humans and providing incorrect or offensive responses is a case in point.

Cost and Infra: The cost of training a generative AI model can be exorbitant. This makes it difficult for smaller organizations to afford necessary resources. Enterprises looking to use Generative AI models should note that large amounts of data are required for models to be trained on and data preparation might not be an easy task, along with the huge infra and computational resources that is needed.

Copyright: There are concerns about authorship and copyright for AI-generated content, as it may be impossible to trace the source of the generated output given the size of training data. This creates ambiguity around ownership and usage rights. It also lacks creativity as output would always be based on training data and not a unique thought or idea.

Repetitiveness: Generative AI models can often produce repetitive and similar output due to the nature of the data they were trained on. To create novel and creative content, models need to be trained on diverse data sets.

Privacy and Regulation: Deepfakes and Voice clones are a real threat to one's privacy and generative AI has made it easier to create human-like digital clones. More stringent policies and guidelines are required to protect privacy and sensitive data present in conversations. Another major concern is the ability of AI to bypass authentication mechanisms like voice checks, captcha etc.

Does this mean Generative AI isn't relevant for enterprises?

Not necessarily, a hybrid approach may be more beneficial for different use cases within an organization. For example, an intent-based approach may be more effective for retrieving data from transactional data systems, while a transformer-based architecture may be better suited for retrieving and presenting policy related information as-is to end-users. Generative AI may be more beneficial where responses can be generated, for example, troubleshooting a problem, questions on installing or uninstalling a software etc.

Conclusion

While Generative AI is still evolving, its potential impact on the role of Chatbots within organizations is significant. The investment by Hyperscalers in developing Generative AI technology and services will make its adoption by enterprises easier.

However, challenges such as bias, data privacy, and copyright must continue to be addressed. By adopting Generative AI models such as GPT-4, FLAN-UL2 20B, Bloom, LLaMa, GPT NeoX, Cohere. AI, Chatbots will become the primary channel of interaction within organizations, rather than just an alternate channel.This will significantly enhance problem-solving abilities of Chatbots and enable more personalized and human-like interactions with users.

Overall, the adoption of Generative AI will strengthen the role of Chatbots in organizations and provide a more versatile and efficient user experience.

Authors

Guruprasad NV

AVP - Senior Principal Technology Architect, Infosys

Amit Kumar

Senior Technology Architect, Infosys

Arpit Bhardwaj

Senior Associate Consultant, Infosys

References:

Artificial Intelligence Is Booming—So Is Its Carbon Footprint

ChatGPT and generative AI are booming, but at a very expensive price (cnbc.com)

For more information, contact askus@infosys.com

Infosys®
Navigate your next