Conceptual Framework for Utilizing Chatbots as Domain Experts in Organizations

Mihael Škarabot,[1] Robert Leskovar[2]

[1] Cloudvenia Ltd., Kranj, Slovenia
mihael.skarabot@cloudvenia.com
[2] University of Maribor, Faculty of Organizational Sciences, Kranj, Slovenia
robert.leskovar@um.si

This paper articulates conceptual framework for investigating the deployment of Large Language Models (LLMs) in the capacity of expert-level chatbot interfaces within organizational settings. Commencing with an exhaustive review of the pertinent literature, this study delineates the landscape of LLM application in corporate environments. The challenges encompass the heterogeneity of human-LLM interactions, the propensity for inadvertent errors, and the consequential effects on employee engagement and motivation. Foremost among these is the examination of the intricacies involved in the symbiosis of LLMs with extant business information systems, particularly evaluating the utility of LLMs as dynamic, bi-directional communicative interfaces. Moreover, the study anticipates the prospective impacts that LLMs may exert on prevailing human-machine interfaces within such information systems. Conclusively, this paper introduces high-level theoretical model for the integration of LLM-driven chatbots into business information systems, setting a platform for future investigations. This model is advancing the understanding of the transformative role of LLMs in augmenting and refining organizational information processing and decision-making paradigms.

Keywords: chatbots, AI domain experts, artificial intelligence, framework, organizations

University of Maribor Press

DOI https://doi.org/10.18690/um.fov.3.2024.67
ISBN 978-961-286-842-0

1 Introduction

In the waning days of November 2022, OpenAI unveiled ChatGPT, a groundbreaking development in the realm of conversational artificial intelligence. By January 2023, a mere two-month post-launch, ChatGPT had amassed an unprecedented user base exceeding 100 million individuals. This rapid and widespread adoption not only marked ChatGPT as the fastest growing consumer application in recorded history but also catalysed a significant surge in the AI sector, heralding what many have termed an "Artificial Intelligence Boom" (ChatGPT – Wikipedia, n.d.).

Prior to this juncture, consumer-facing AI chatbots were not an unfamiliar concept. Notable among them was Bank of America's AI-driven virtual assistant, Erica, which had been successfully integrated into the banking ecosystem, servicing millions of customers with banking-specific inquiries. Furthermore, in the finance sector, BlackRock's Robo-Advisor 4.0 stood as a testament to the amalgamation of AI and Machine Learning capabilities. This sophisticated system demonstrated a proficiency in advising clients on financial matters, showcasing a performance that could potentially surpass traditional human stock-pickers (Pal, A., Gopi, S., and Lee, K., 2023).

The primary distinguishing characteristic of chatbots powered by the Generative Pre-trained Transformer (GPT) model lies in their exceptional contextual understanding capabilities. This feature marks a significant advancement over previous generations of chatbots, which, despite incorporating some level of Natural Language Processing (NLP), often lacked the depth and nuance necessary for truly natural conversation flow. GPT-based chatbots transcend these limitations through their advanced and sophisticated NLP capabilities, ensuring interactions that are remarkably fluid and human-like.

A critical aspect of the GPT model is its generative nature. Unlike traditional chatbots that primarily rely on pre-defined responses, GPT-based systems have the unique ability to generate original content. This capability extends beyond simple response formulation. The model can synthesize or summarize inputs or combine its extensive trained knowledge (Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019) to create entirely original content. Such a feature is pivotal in

providing bespoke interactions, as it allows the chatbot to tailor its responses to individual users dynamically, considering the specific context of each interaction.

Moreover, the voluminous and diverse training material that forms the foundation of the GPT model endows it with a broad understanding of language, context, and subject matter. This extensive training enables the chatbot to engage in a wide array of topics, further enhancing the naturalness and relevance of its conversations (Roumeliotis, K., Tselikas, N., 2023).

In essence, chatbots built on the GPT framework offer a more human-like interaction experience than their predecessors. This is not merely in terms of the sophistication of the responses but also in the overall conversational quality, where the chatbot can adapt, respond, and even anticipate user needs in a manner reminiscent of human interaction. The GPT innovative approach to NLP represents a paradigm shift in the field of chatbot technology, setting a new benchmark for what is achievable in AI-driven conversational agents.

This paper aims to provide a comprehensive literature review of GPT-driven chatbots within organizational settings.

1.1 Understanding the Scope of Research

The emergence of chatbots has primarily been researched and analysed from the perspective of business-to-consumer (B2C) engagement. Recent advancements in generative chatbot technology, exemplified by systems such as ChatGPT, present a multitude of robust use-cases within various business domains (Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., Chen, L., 2023). Their usage extends to critical business functions including marketing, information technology/engineering, as well as legal and healthcare sectors. Within these domains, the capabilities of generative AI for content creation, data summarization, and language translation are of paramount importance. The efficacy of generative AI in these areas has demonstrated remarkable potential, offering significant advancements in operational efficiencies and strategic capabilities.

Chatbot utilization in direct integrations with business information systems for the purpose of supporting employees in executing business processes remains an under-explored and under-utilized area. This research proposes to fill this gap by examining the role of chatbots as domain experts in information systems such as Financial Management, Human Resources Management (HRM), Supply Chain Management (SCM), Manufacturing, Inventory Management, Purchasing/procurement, Project Management, Order Processing and Business Intelligence (BI). The aim is to outline how to implement integrations and to understand how employees can effectively utilize AI driven chatbots to optimize internal process workflows, enhance data analytics and decision making, and improve overall operational efficiency.

While acknowledging the complex and intricate nature of Large Language Models (LLM) on which ChatGPT is based on, this study will limit its examination to a comprehensive understanding of the capabilities and limitations inherent in these models. The research does not delve deeply into the intricate technical mechanics and algorithmic foundations of LLMs. Instead, it will appreciate their practical applicability and existing constraints.

This approach ensures a balanced perspective, concentrating on the pragmatic aspects of LLM implementation in business environments while maintaining an informed awareness of the underlying technological principles. The goal is to offer valuable insights into how LLMs can be strategically integrated into business systems to optimize decision-making efficacy, process efficiency and facilitating a more human-like interaction with information systems. The latter is also viable from the perspective of user preference as analysed by research studies (Sakirin, T., Ben Said, R., 2023).

1.2 Challenges of Implementing Chatbots in Organizational Environments

A suite of large language model (LLM)-based chatbot systems are currently accessible via the internet for public utilization. Prominent examples of these systems include OpenAI ChatGPT, Microsoft Bing Copilot, Google Gemini, etc. These chatbots are particularly suitable for scenarios that do not require extensive integration with business-specific contexts. In most cases, uploading files for building an indexed knowledge base, writing prompts and copy-pasting text back is

sufficient. This research does not delve into this context as we target more advanced automated integrations with business information systems.

A major challenge with Large Language Models (LLMs) is their tendency to produce "hallucinations" or outputs that, despite being presented confidently, may include partially fictional content. This issue is compounded by the risk of LLMs relying on outdated information. Such drawbacks pose significant obstacles in business contexts where LLMs are used for knowledge acquisition, data processing, and decision-making (Alkaissi, H., McFarlane, S., 2023). This article aims to address these challenges by presenting strategies to not only reduce instances of hallucinations but also to ensure the currency and accuracy of the information provided by chatbots, thereby significantly improving their reliability and credibility.

Further challenge lies in bridging the gap between structured data in information systems and the unstructured text generated by Large Language Models (LLMs), as well as the reverse process of translating unstructured LLM outputs into structured data for actionable insights within targeted information systems. This interoperability is critical, as it enables chatbots to not only access data from existing business information systems but also to initiate and execute business-related events within these systems. Developing an efficient model to facilitate this seamless integration and interaction is essential for the effective utilization of chatbots in business environments.

To facilitate the integration of business information systems and their respective data and interfaces with these chatbots, two primary LLM deployment methods are employed: utilization of cloud-based LLM APIs or on-premises LLM installations. To choose the viable approach consideration of data security, accessibility, performance, privacy, latency, technical expertise, and cost is needed.

Data protection and authorization are of paramount importance in the utilization of chatbots within organizations. Regulating access to specific segments of business information is essential for maintaining organizational coherence, ensuring continuous operations, and safeguarding security. However, this issue assumes comparatively lesser significance in the context of smaller businesses. These smaller entities often have more straightforward operational structures and reduced data complexities, which may result in a lower risk profile for data breaches or

unauthorized access. As such, the stringent measures required for larger organizations may not be as critical for smaller businesses, though they should still maintain a basic level of data protection and access control.

The following key research questions in the scope of this topic have been identified:

- How to integrate chatbot into business information systems and what are the challenges of this integration?
- What are the advantages of using chatbots as bi-directional interfaces to business information systems compared to existing methods?
- How do LLMs affect the accuracy and reliability of structured data in business information systems?
- What are the possible security, privacy and intellectual property concerns when using LLMs to access sensitive business data?
- What are the expected impacts of using chatbots as interfaces for structured data on productivity and competitiveness of businesses?

1.3 Literature Review

Comprehensive analysis of scholarly publications reveals a substantial dedication of research resources towards the deployment of chatbots in Business-to-Consumer (B2C) communication. These studies focus on chatbot applications in customer support, eCommerce, advisory services, education, healthcare, and interactions between citizens and government (Luo, B., Lau, R., Li, C., Si, Y., 2022). The increasing adoption of chatbots in these domains, alongside the corresponding escalation in research activities, can be attributed in part to the principle of economies of scale. Chatbots operating within the B2C sector benefit from a broader user base and heightened levels of utilization. This expansive reach typically results in a more favourable return on investment (ROI), underlining clear business incentives for their implementation. These types of chatbots are not targeted in this research.

In comparison, chatbots deployed within internal business systems primarily for employee usage face a notably smaller user base. This reduced scale of interaction significantly impacts the economic justification for investing in the development and

implementation of chatbots in these scenarios. The limited number of users often results in a lower return on investment, making the economic case for their development in such internal contexts less compelling and more challenging to validate.

With the introduction of Large Language Model (LLM) based chatbots, there is a paradigm shift in this perspective. This paper posits that LLM-based chatbots are set to revolutionize and justify their application within business systems. This evolution is expected to redefine the economic viability and functional relevance of chatbots in these internal business environments, overcoming the constraints posed by smaller user base.

Since the breakthrough of ChatGPT and similar LLMs occurred only two years ago, specific literature on their application in this context is scarce. Existing research related to LLMs tends to concentrate on assessing the extent of employees utilizing publicly available LLMs which are not directly integrated with information systems. Other studies (Eloundou, T., Manning, S., Mishkin, P., and Rock, D. 2023), explore the potential long-term effects of LLMs on the labour market. These studies generally agree on the expected impact of LLMs, yet they do not delve into chatbots enhanced with specific business domain knowledge tailored for use by employees within their organizations. This gap indicates an emerging area of research, focusing on the integration of domain-specific knowledge into LLM-based chatbots to optimize their utility in internal business settings.

2 Methodology

The literature review for this study was meticulously conducted following the established guidelines outlined by Snyder (Snyder, H., 2019). The search for pertinent articles was methodically carried out across several renowned academic databases, including Web of Science, Scopus (Elsevier), ProQuest, and Google Scholar. This process began with a strategic keyword search, followed by a refinement of the results using specific keyword combinations.

To evaluate the gathered articles, an initial review was conducted, focusing on abstracts and skim reads to ascertain the relevance and depth of the content. This approach facilitated the selection of the most pertinent articles, which were crucial

for acquiring a comprehensive understanding of previous research findings in the domain of Large Language Models (LLMs).

Given the rapid and recent evolving nature of LLM, a key criterion for the relevance of these articles was their publication year. This study prioritized recent publications, recognizing that the field of LLM (GPT) research has a relatively short history, spanning no more than two years. Additionally, the selection process placed significant emphasis on the number of references in each article. This measure served as an indicator of the article's impact and the extent to which it has contributed to the field, ensuring that the most influential and informative works were included in this review.

3 Intermediate Research Results

3.1 Deciding Between Cloud-Based and On-Premises Solutions

Utilizing a cloud-based API offers a cost-effective solution, primarily due to the distributed cost allocation among multiple API consumers and the adoption of pay-per-use pricing models. However, this approach raises concerns regarding on-premises IS integration, data privacy and security. In contrast, implementing an LLM on-premises is feasible but incurs substantial costs related to computing power, maintenance effort and environmental impact (Doo, F., Kulkarni, P., Siegel, E., Toland, M., Yi, P., Carlos, R., Parekh, V., 2023). Additionally, this approach necessitates ongoing maintenance and potential further training of the LLM models.

Notably, there are highly capable open-source LLMs available, such as Llama 2 from Meta and Falcon 180B from Technology Innovation Institute of the UAE. They can be leveraged for on-premises deployments, however, not without significant running costs (Wodecki, B., 2023).

In the context of research, the cloud-based Azure OpenAI API, has been selected due to its accessibility and robust capabilities. This choice allows for a balance between resource availability and the need for advanced language processing functionalities. Further thought will be given also to questions of privacy and security in such environments.

3.2 Model of LLM Integration with Business Information Systems

Incorporating actual business data into Large Language Models (LLMs) can be effectively achieved through Retrieval-Augmented Generation (RAG), as explained by Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. in 2021. The conceptual model of integration is presented on Fig 1.

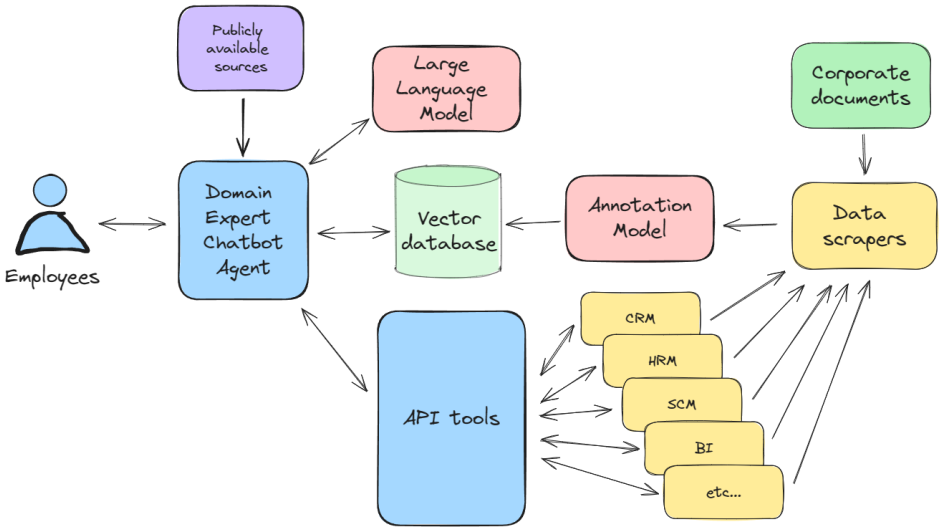

Figure 1: Conceptional model for LLM chatbot integration with information systems
Source: Own

This process necessitates the integration of data from business information systems into an indexed vector database. The data undergoes a transformation into vector embeddings, which are then stored in a searchable vector database. This database is subsequently queried on-demand through a chain-of-thought mechanism employed by agents. These agents execute transient chat threads that perform a series of operations. Such operations include interactions with LLMs itself, utilizing various tools such as web searches, business information system queries, vector database searches, and even human interactions. The determination of the steps necessary to address a user's request and the point at which to deliver the final response are governed by the internal LLM's reasoning capabilities. This framework allows chatbots to become a potent interface within specific business domains in a business environment.

Current research efforts have culminated in the development of a chatbot prototype on top of cloud based LLM API. It is capable of harnessing above mentioned tools through autonomous reasoning. Future developments are directed towards integrating this prototype into a real-world business setting. Implementation of user feedback mechanism will be implemented. Further research in data privacy, data access authorization will follow. After substantial active user base has been established a survey type of research will be executed to establish chatbot usage, efficacy, performance, and limitations.

4 Conclusions

Despite its potential, the integration of LLM based chatbots in internal business processes remains underutilized, presenting a novel area of exploration and research for enhancing business information systems and decision-making processes. A major challenge for LLMs in business contexts is ensuring the accuracy and currency of information, as well as effectively managing the translation between structured data in information systems and unstructured data in human like communication. The research identifies a gap in the literature regarding the application of LLMs in internal business contexts, suggesting an emerging area of research focused on integrating domain-specific knowledge into LLM-based chatbots.

References

Alkaissi, H., McFarlane, S. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus 2023, e35179, 15(2).

ChatGPT – Wikipedia (n.d.). Retrieved January 1, 2024, from https://en.wikipedia.org/wiki/ChatGPT.

Doo, F., Kulkarni, P., Siegel, E., Toland, M., Yi, P., Carlos, R., Parekh, V. (2023). Economic and environmental costs of cloud for medical imaging and radiology artificial intelligence. Journal of the American College of Radiology 2023, month 12.

Eloundou, T., Manning, S., Mishkin, P., Rock, D. (2023). GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. arXiv preprint arXiv:2303.10130 2023.

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. Journal of Information Technology Case and Application Research 2023, 277-304, 25(3).

Luo, B., Lau, R., Li, C., Si, Y. (2022). A critical review of state-of-the-art chatbot designs and applications. WIREs Data Mining and Knowledge Discovery 2022, 12(1)

Pal, A., Gopi, S., Lee, K. (2023). Fintech Agents: Technologies and Theories. Electronics 2023, 12, 3301.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Language Models Are Unsupervised Multitask Learners. OpenAI Blog. 2019. Retrieved from

https://life-extension.github.io/2020/05/27/GPT%E6%8A%80%E6%9C%AF%E5%88%9D%E6%8E%A2/language-models.pdf.

Roumeliotis, K., Tselikas, N. (2023). ChatGPT and Open-AI Models: A Preliminary Review. Future Internet 2023, 15, 192.

Sakirin, T., Ben Said, R. (2023). User preferences for ChatGPT-powered conversational interfaces versus traditional methods. Mesopotamian journal of Computer Science 2023, pp 24-31.

Shuster, K., Poff, S., Chen, M., Kiela, D., Weston, J. (2021). Retrieval Augmentation Reduces Hallucination in Conversation. Facebook AI Research, arXiv:2104.07567v1, from https://arxiv.org/pdf/2104.07567.pdf

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. Journal of Business Research, (2019), 333-339, 104

Wodecki, B. (2023). Open Source vs. Closed Models: The True Cost of Running AI, AI Business, 2023, from https://aibusiness.com/nlp/open-source-vs-closed-models-the-true-cost-of-running-ai.