10-9-2023

Towards Hybrid Architectures: Integrating Large Language Models in Informative Chatbots

Arnold F. Arz von Straussenburg
University of Koblenz, Germany, arz@uni-koblenz.de

Anna Wolters
University of Koblenz, Germany, awolters@uni-koblenz.de

Recommended Citation

Arz von Straussenburg, Arnold F. and Wolters, Anna, "Towards Hybrid Architectures: Integrating Large Language Models in Informative Chatbots" (2023). *Wirtschaftsinformatik 2023 Proceedings*. 9.
https://aisel.aisnet.org/wi2023/9

Towards Hybrid Architectures: Integrating Large Language Models in Informative Chatbots

Research in Progress

Arnold F. Arz von Straussenburg[1] and Anna Wolters[1]

University of Koblenz, Koblenz, Germany
{arz, awolters}@uni-koblenz.de

Abstract. Informative chatbots embedded in an organization-specific context can provide a reliable, interactive, and engaging source of information for users. However, traditional chatbot techniques have limitations in processing and understanding user input and generating human-like responses. On the other hand, the latest implementations of large language models show promising results in these domains but have limitations in providing accurate and up-to-date facts from domain-specific knowledge bases. With the advent of popular chatbots like ChatGPT, they are increasingly becoming part of organizations' digital infrastructure. In this research-in-progress paper, we argue that the strengths and weaknesses of traditional techniques and large language models are complementary. Therefore, we propose a hybrid chatbot architecture that utilizes inter-agent communication to compensate for disadvantages while enhancing the chatbot's abilities, as perceived by the user. This approach will form the basis for development and evaluation using Design Science Research (DSR) as part of our research.

Keywords: Chatbots, Hybrid Chatbot Architecture, Large Language Models

1 Introduction & Motivation

Chatbots provide an interactive way to access information, perform tasks, and act as friendly interlocutors (Adamopoulou & Moussiades 2020*a*). They can be used in various sectors for different purposes to help organizations by providing a source of information with control over accuracy and topicality. However, challenges include understanding user intent, generating human-like responses, and maintaining data sources.

More powerful Large Language Models (LLMs) offer advanced chatbot capabilities beyond existing Natural Language Processing (NLP) and Machine Learning (ML) techniques, excelling in generating human-like text and understanding unstructured data sources (Brown et al. 2020). ChatGPT, an LLM-based chatbot trained with reinforcement learning from human feedback, has had a significant impact on the public since its release (OpenAI 2022). Despite their advantages, these new technologies still have drawbacks, including the potential to provide responses that are not supported by the underlying knowledge base. This may become a significant issue as chatbots become more integrated into organizations' digital infrastructure and operations. Furthermore, while commercial chatbots already use LLMs (Shuster et al. 2022), there is limited research on using LLMs for informative chatbots.

18th International Conference on Wirtschaftsinformatik
September 2023, Paderborn, Germany

This paper not only contrasts established chatbot methods with LLM-based chatbots to understand their respective strengths and weaknesses but also pioneers the integration of LLMs and traditional chatbot technologies, proposing a novel hybrid approach aimed at improving efficacy and user experience in information retrieval tasks. Furthermore, we argue that a combination of both techniques can enable the creation of powerful chatbot architectures that mitigate weaknesses and emphasize strengths. To achieve this, the research paper addresses the following research questions: *What are the advantages and shortcomings of both traditional and LLM chatbots?* (RQ1) and *How can these approaches be integrated/ combined to enable accurate informative chatbots that provide natural-sounding answers?* (RQ2).

2 Background

2.1 Classification of Chatbots

Chatbots, also referred to as conversational agents, digital assistants, interactive agents, or natural dialogue systems (Adamopoulou & Moussiades 2020*a*,*b*, Diederich et al. 2019) are computer systems that enable humans to interact with computers using natural language (Lokman & Ameedeen 2019). While not all chatbots can be fit neatly into categories, several dimensions can classify the types of chatbots considered in scientific discourse, each with different characteristics that form the basis for our research. The combined chatbot classifications or taxonomies presented by Adamopoulou & Moussiades (2020*b*), Diederich et al. (2019), Lokman & Ameedeen (2019), and Hussain et al. (2019) are depicted in Table 1.

Chatbots can have different objectives, namely providing information, supporting users in fulfilling tasks, or focusing on conversing with their users (Adamopoulou & Moussiades 2020*b*, Diederich et al. 2019). There are also different approaches for processing input texts and generating responses. Rule-based techniques apply pattern-matching to understand the context and the user's intent (Adamopoulou & Moussiades 2020*b*), while the most human-like response can be achieved by generating the answer using ML techniques, depending on the quality of the training data, the complexity of the users' input and other factors. Additionally, text processing might be performed using word embeddings or based on the text itself as represented by the Latin alphabet (Lokman & Ameedeen 2019).

Chatbots can be implemented by programming, modeling, supervised learning, or applying a hybrid approach, combining the previous implementation techniques (Diederich et al. 2019). Moreover, users might interact with the chatbot via text, voice, or both, while the chatbot could support a single or multiple languages (Hussain et al. 2019, Diederich et al. 2019).

2.2 Large Language Models

An Large Language Model (LLM) is an advanced AI-based text model that sets itself apart from other text models by utilizing *massive* amounts of data, surpassing what is commonly referred to as big data (O'Leary 2022). The resulting pre-trained, often autore-

Table 1. Chatbot Classification

<table>
<tr><th>Dimension</th><th colspan="12">Characteristics</th></tr>
<tr><td>Knowledge Domain</td><td colspan="6">Open-domain/ General-purpose</td><td colspan="6">Closed-domain/ Domain-specific</td></tr>
<tr><td>Service Provided</td><td colspan="4">Interpersonal</td><td colspan="4">Intra-personal</td><td colspan="4">Inter-agent</td></tr>
<tr><td>Goal</td><td colspan="4">Informative</td><td colspan="4">Task-based</td><td colspan="4">Conversational/ Non-task based</td></tr>
<tr><td>Input Processing/ Response Generation</td><td colspan="4">Rule-based</td><td colspan="4">Retrieval-based</td><td colspan="4">Generative</td></tr>
<tr><td>Human-aid</td><td colspan="6">Human mediation</td><td colspan="6">Autonomous</td></tr>
<tr><td>Permission</td><td colspan="6">Open-source</td><td colspan="6">Commercial</td></tr>
<tr><td>Text Processing</td><td colspan="6">Word Embeddings</td><td colspan="6">Text-level (Latin Alphabet)</td></tr>
<tr><td>Interaction/ Communication Mode</td><td colspan="4">Text-based</td><td colspan="4">Voice-based</td><td colspan="4">Both</td></tr>
<tr><td>Implementation</td><td colspan="3">Programming</td><td colspan="3">Modeling</td><td colspan="3">Supervised Learning</td><td colspan="3">Hybrid</td></tr>
<tr><td>Language</td><td colspan="6">Single Language</td><td colspan="6">Multi Language</td></tr>
</table>

gressive deep learning models, aim at producing natural language text (Floridi & Chiriatti 2020) and often feature a massive number of parameters, far exceeding those of previous models. As a result, these models offer a versatile and task-agnostic foundation for a wide range of text-based tasks, such as text summarization or generation (Brown et al. 2020).

Large organizations such as Google, OpenAI and Meta offer multiple commercial LLMs like Google's LaMDA (Thoppilan et al. 2022), T5 (Raffel et al. 2020), and BERT (Devlin et al. 2019) models, as well as OpenAI's GPT-4 and GPT-3.5, and Meta's BlenderBot 3 (Shuster et al. 2022) and LLaMA (Touvron et al. 2023). Out of these models, GPT-4 is currently the largest, consisting of 175 billion parameters, while the first and second generations used 110M (GPT-1) and 1.5B (GPT-2) parameters, respectively (Floridi & Chiriatti 2020, Thoppilan et al. 2022). Besides commercial LLMs, there are also freely and publicly available models that are larger than GPT-2. Examples of these GPT-Neo (2.7B parameters) (Black et al. 2021), GPT-J-6B (Wang & Komatsuzaki 2021), and GPT-NeoX-20B (Black et al. 2022), which were developed by EleutherAI based on the GPT-NeoX platform (Andonian et al. 2023). Other examples are PanGu-α (13B) (Zeng et al. 2021) and FairSeq (2.7B, 6.7B, 13B) (Artetxe et al. 2022).

To apply pre-trained models to context-specific tasks, fine-tuning is typically required. This involves a supervised learning process that updates the weights of the training parameters and generally requires large datasets with thousands of labeled data instances (Brown et al. 2020). An example of a context-specific model is ChatGPT (OpenAI 2022). However, to avoid extensive fine-tuning, pre-trained LLMs can be used with one-shot or few-shot learning. These techniques involve in-context learning using 10-100 examples as prompts. Unlike fine-tuning, one-shot or few-shot learning doesn't update model parameters, but few-shot learning has shown promising results. Another variant is zero-shot learning, which does not use any demonstrations but relies on a natural language description of the task (Brown et al. 2020).

3 Research Method

Our research-in-progress aims to build a hybrid chatbot that integrates LLMs with informative approaches, for which we present the conceptualization in the given paper. Therefore, we are working in the field of design-oriented information systems research (Hevner et al. 2004, Österle et al. 2011), with the goal of designing an artifact, which is the hybrid chatbot. To achieve this, our research follows the Design Science Research (DSR) methodology proposed by Peffers et al. (2007). In this paper, Section 1 and Section 2 contribute to problem identification and motivation. Section 4 compares LLMs with informative chatbots, which helps define the objectives for hybrid chatbots. Then, we develop a concept and present a prototypical implementation in Section 5, which contributes to the design and development phase, as well as the demonstration. Since this manuscript is still in progress, we have yet to conduct a comprehensive demonstration and evaluation. In the future, the resulting artifact will be validated on the basis of a set of experts. However, by communicating our intermediate results, we contribute to the sixth phase of Peffers et al. (2007). This allows us to integrate feedback from our peers with upcoming process iterations.

4 Comparison of Different Chatbot Approaches

Existing research in chatbot development has established a general architecture that can be used to design chatbots for different categories, as shown in Figure 1. This architecture, adapted from Adamopoulou & Moussiades (2020*b*), comprises three main components: NLP; dialogue management; and information retrieval, which involves gathering data from sources such as databases or the web (Adamopoulou & Moussiades 2020*b*). Different types of chatbots focus on different parts of the general architecture. Informative chatbots rely on comprehensive information retrieval to answer user queries, while conversational chatbots focus on dialogue management and user message analysis to sustain a conversation. The approach used for language understanding is also affected by the choice of input processing and response generation method.

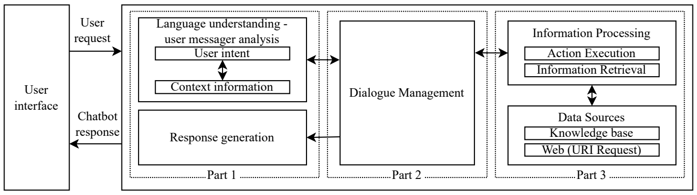

Figure 1. General chatbot architecture.

As mentioned in Section 2, there are two approaches for developing chatbots: rule-based and ML-based. Rule-based chatbots are specific to a task and follow a decision tree-like path with hand-coded or structured knowledge (Adamopoulou & Moussiades 2020*b*, Ramesh et al. 2017). ML-based chatbots consider the dialogue context and do not require predefined responses but need (at least incremental) training on labeled datasets for new tasks, making them less applicable when a chatbot gains new knowledge

(Adamopoulou & Moussiades 2020*a*). These chatbots are typically more focused on producing human-like responses.

Task-agnostic LLMs, like Generative Pre-trained Transformer 4 (GPT-4), have created new possibilities for building chatbots. They differ from traditional ML approaches in implementation and capabilities, with LLMs being able to generalize and adapt to specific tasks using a relatively small amount of task-specific data and transfer downstream tasks easily (Brown et al. 2020). Given suitable training data and computational resources, they can cover an extensive range of topics, generate more organic responses, and update more efficiently than their traditional counterparts. However, there are also drawbacks to using LLMs. They can generate inappropriate text when inappropriate language is fed into them, and their generative nature may result in plausible but unsupported answers (O'Leary 2022). Moreover, LLMs require more computational resources and can be more expensive to operate than traditional chatbots. Thus, both chatbot approaches have their own advantages and drawbacks, which answers the first research question.

5 Towards Hybrid Chatbots

5.1 Design of a Hybrid Architecture

The difference between rule-based and ML-based traditional chatbots results in divergent integration with LLM approaches. Rule-based chatbots offer developers tight control over included information and answers, making them well-suited for informative chatbots. Thus we focus on the integration with generative LLMs to enhance their ability to generate natural text in the following.

To balance the opposing strengths and weaknesses of both rule-based and LLM approaches, we propose a generic concept of a hybrid chatbot. Specifically, we plan to integrate LLMs into traditional informative chatbots, building on the notion of hybrid chatbot implementations presented by Diederich et al. (2019). The proposed hybrid chatbot will leverage a combination of chatbots with different approaches to provide the best possible architecture with their specific advantages. This can be achieved by chaining two interdependent chatbot components that are not functional on their own and combining them through inter-agent communication, as shown in Figure 2.

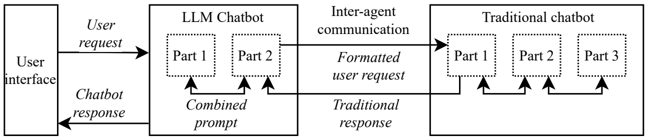

Figure 2. LLM chatbot that enhances the answering capabilities of a traditional chatbot.

The figure illustrates a traditional, rule-based chatbot on the right-hand side, which consists of all three parts of the architecture shown in Figure 1: user message analysis and response generation (Part 1), dialogue management (Part 2), and information retrieval (Part 3) and a LLM-based chatbot, without Part 3 of the architecture, on the left-hand side. To overcome the limitations of Part 1 in the traditional chatbot, an LLM is utilized to analyze the incoming *user request* and transforms it via inter-agent communication, resulting in a *formatted user request*. This formatted user request provides the user intent and necessary context, like intents derived from previous questions, to the traditional chatbot. Subsequently, the traditional chatbot can accurately provide relevant information in a condensed manner (*traditional response*) by accessing the knowledge base. Finally, the chatbot response includes the output of the combined prompt, which includes the original user request, the traditional chatbot's response, and instructions on the answer format.

5.2 Implications

The proposed architecture combines the accuracy guarantees of rule-based chatbots with the conversational capabilities of LLMs, which are often superior to those of rule-based chatbots. This hybrid architecture can be used to enhance existing traditional chatbots, allowing adoption in many application cases. These enhancements have the potential to significantly improve a variety of use cases in many organizations, like B2B customer service by providing prompt and precise responses to FAQs, reducing representatives' workload, and personalizing interactions to improve customer satisfaction and retention. Simultaneously, they enable a more personalized customer experience which can improve customer satisfaction and loyalty, acting as a key driver for retention.

6 Conclusion & Outlook

In this research-in-progress paper, we aim to enhance our understanding of how the combination of LLMs with established chatbot technologies can improve the reliability of informative chatbots. Drawing on existing chatbot classifications and architectures, we have concluded that addressing the possibility of including LLMs, such as GPT-4, in chatbot research is essential. By analyzing the strengths and weaknesses of different chatbot approaches, we propose a conceptualization of a hybrid chatbot implementation. We use inter-agent communication to combine a LLM and a traditional chatbot, creating an architecture that leverages the advantages of both. We argue that LLMs outperform traditional NLP methods in generating human-like responses while relying on the benefits of accessing structured information through the information retrieval part of a traditional chatbot. In this way, we ensure that the generated response is based on valid information. With this concept, we make a contribution to the research on developing hybrid chatbot architectures without restricting it to specific technologies. In the next step of our research, we will prototype the conceptualization and evaluate its ability to generate human-like responses and answer questions based on valid and reliable information. Future research needs to follow an iterative design process by taking multiple technologies into consideration, and by continuously adjusting the concept based on the latest evaluation results.

Acknowledgements

This research has been funded by The Federal Ministry of Education and Research (BMBF) under Research Grant No. 16DHBKI039.

References

Adamopoulou, E. & Moussiades, L. (2020*a*), 'Chatbots: History, technology, and applications', *Machine Learning with Applications* **2**, 1–18.

Adamopoulou, E. & Moussiades, L. (2020*b*), An overview of chatbot technology, *in* I. Maglogiannis, L. Iliadis & E. Pimenidis, eds, 'Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology', Vol. 584, Springer, pp. 373–383.

Andonian, A., Biderman, S., Black, S., Gali, P., Gao, L., Hallahan, E., Levy-Kramer, J., Leahy, C., Nestler, L., Parker, K., Pieler, M., Purohit, S., Songz, T., Phil, W. & Weinbach, S. (2023), 'GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch', Zenodo, doi: 10.5281/zenodo.7714278.

Artetxe, M., Bhosale, S., Goyal, N., Mihaylov, T., Ott, M., Shleifer, S., Lin, X. V., Du, J., Iyer, S., Pasunuru, R., Anantharaman, G., Li, X., Chen, S., Akin, H., Baines, M., Martin, L., Zhou, X., Koura, P. S., O'Horo, B., Wang, J., Zettlemoyer, L., Diab, M., Kozareva, Z. & Stoyanov, V. (2022), 'Efficient Large Scale Language Modeling with Mixtures of Experts', arXiv:2112.10684.

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U. S., Purohit, S., Reynolds, L., Tow, J., Wang, B. & Weinbach, S. (2022), GPT-NeoX-20B: An open-source autoregressive language model, *in* 'Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models'.

Black, S., Gao, L., Wang, P., Leahy, C. & Biderman, S. (2021), 'GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow', Zenodo, doi: 10.5281/zenodo.5297715.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. (2020), Language models are few-shot learners, *in* H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan & H. Lin, eds, 'Advances in Neural Information Processing Systems', Vol. 33, pp. 1877–1901.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019), 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', arXiv:1810.04805.

Diederich, S., Brendel, A. B. & Kolbe, L. M. (2019), Towards a taxonomy of platforms for conversational agent design, *in* 'Proceedings of 14th International Conference on Wirtschaftsinformatik', pp. 1100–1114.

Floridi, L. & Chiriatti, M. (2020), 'Gpt-3: Its nature, scope, limits, and consequences', *Minds and Machines* **30**(4), 681–694.

Hevner, A. R., March, S. T., Park, J. & Ram, S. (2004), 'Design Science in Information Systems Research', *MIS Quarterly* **28**(1), 75–105.

Hussain, S., Ameri Sianaki, O. & Ababneh, N. (2019), A survey on conversational agents/chatbots classification and design techniques, *in* L. Barolli, M. Takizawa, F. Xhafa & T. Enokido, eds, 'Proceedings of the Workshops of the 33rd International

Conference on Advanced Information Networking and Applications (WAINA-2019)', Vol. 927, pp. 946–956.

Lokman, A. S. & Ameedeen, M. A. (2019), Modern chatbot systems: A technical review, *in* 'Proceedings of the Future Technologies Conference (FTC) 2018', Vol. 881, Springer Verlag, pp. 1012–1023.

O'Leary, D. E. (2022), 'Massive data language models and conversational artificial intelligence: Emerging issues', *Intelligent Systems in Accounting, Finance and Management* **29**(3), 182–198.

OpenAI (2022), 'Introducing ChatGPT'. Accessed: 2023-06-23.
URL: *https://openai.com/blog/chatgpt*

Österle, H., Becker, J., Frank, U., Hess, T., Karagiannis, D., Krcmar, H., Loos, P., Mertens, P., Oberweis, A. & Sinz, E. J. (2011), 'Memorandum on design-oriented information systems research', *European Journal of Information Systems* **20**(1), 7–10.

Peffers, K., Tuunanen, T., Rothenberger, M. A. & Chatterjee, S. (2007), 'A Design Science Research Methodology for Information Systems Research', *Journal of Management Information Systems* **24**(3), 45–77.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. & Liu, P. J. (2020), 'Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer', arXiv:1910.10683.

Ramesh, K., Ravishankaran, S., Joshi, A. & Chandrasekaran, K. (2017), A survey of design techniques for conversational agents, *in* S. Kaushik, D. Gupta, L. Kharb & D. Chahal, eds, 'Information, Communication and Computing Technology', Vol. 750, Springer Verlag, pp. 336–350.

Shuster, K., Xu, J., Komeili, M., Ju, D., Smith, E. M., Roller, S., Ung, M., Chen, M., Arora, K., Lane, J., Behrooz, M., Ngan, W., Poff, S., Goyal, N., Szlam, A., Boureau, Y.-L., Kambadur, M. & Weston, J. (2022), 'Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage'.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhao, V., Zhou, Y., Chang, C.-C., Krivokon, I., Rusch, W., Pickett, M., Srinivasan, P., Man, L., Meier-Hellstern, K., Morris, M. R., Doshi, T., Santos, R. D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera-Arcas, B., Cui, C., Croak, M., Chi, E. & Le, Q. (2022), 'Lamda: Language models for dialog applications', arXiv:2201.08239.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E. & Lample, G. (2023), 'LLaMA: Open and efficient foundation language models', arXiv:2302.13971.

Wang, B. & Komatsuzaki, A. (2021), 'Gpt-j-6b: A 6 billion parameter autoregressive language model'. Accessed: 2023-06-23.
URL: *https://github.com/kingoflolz/mesh-transformer-jax*

Zeng, W., Ren, X., Su, T., Wang, H., Liao, Y., Wang, Z., Jiang, X., Yang, Z., Wang, K., Zhang, X., Li, C., Gong, Z., Yao, Y., Huang, X., Wang, J., Yu, J., Guo, Q., Yu, Y., Zhang, Y., Wang, J., Tao, H., Yan, D., Yi, Z., Peng, F., Jiang, F., Zhang, H., Deng, L., Zhang, Y., Lin, Z., Zhang, C., Zhang, S., Guo, M., Gu, S., Fan, G., Wang, Y., Jin, X., Liu, Q. & Tian, Y. (2021), 'PanGu-α: Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation', arXiv:2104.12369.