Project presentation accepted for CONVERSATIONS 2024 – the 8th International Workshop on Chatbots and Human-Centred AI, hosted by CERTH, Thessaloniki, Greece, December 4-5, 2024.

Using a large language model to support informal carers.

Hanna Allemann[1[0000-0003-1498-7021]], Daniel Holmer[2[0000-0003-1861-7706],
Mattias Arvola[2[0000-0003-2919-098X]], Tom Ziemke[2[0000-0001-6883-2450]],
Anna Strömberg[1 [0000-0002-4259-3671]]

[1] Nursing Sciences and Reproductive Health, Department of Health, Medicine and Caring Sciences, Linköping University, Linköping, Sweden [2] Human-Centered Systems, Department of Computer and Information Science, Linköping University, Linköping, Sweden

hanna.allemann@liu.se

Abstract. This paper describes a project exploring the feasibility of utilizing a chatbot to support informal carers of persons with heart failure, using the contents of a pre-existing co-designed online support programme. We describe the development and evaluation of a prototype using a large language model and discuss future research directions as well as expected impacts.

Keywords: Chatbot, Large language models, Retrieval-Augmented Generation, Health care, Support, Informal carers, Heart failure

1 Introduction

With increasing life expectancy and better treatment and care, more people are living longer with chronic conditions [1] such as heart failure (HF). HF is a serious and unpredictable syndrome that affects about 2% of adults and progressively worsens over time. The prevalence increases with age, and having HF implies both physical and psychological symptoms. Commonly, HF leads to symptoms like breathlessness and fatigue which affects the ability to be physically active. Furthermore, symptoms of depression and a lower perceived health related quality of life are common [2].

1.1 The scope of informal care

Informal care typically refers to unpaid support, help, and care provided by family members, friends, or neighbors. In Europe approximately 15% care for someone [3] and from a life perspective, everyone will, at some point, provide informal care. Informal carers support persons with HF both practically and emotionally [4] and, with self-care [5]. Diligent self-care is important for both health and survival [6]. Although caregiving can be rewarding [7], it can also be experienced as strenuous and demanding [8]. The unpredictability of HF influence caregiving [9], and carers may lack support, especially from health care providers [4].

1.2 A co-designed online support developed within a healthcare context

As part of approaching health care and welfare's challenges, digitalization is highlighted as an opportunity to address the discrepancy between resources and needs [10]. This is a motive for finding ways to (gently) push the utility of digital solutions as part of regular health care and welfare. Therefore, to support those carers who wish to remain in their caring roles, we employed a co-design methodology with the intent of identifying what could be considered relevant online support. The content in the online support was developed in cooperation between informal carers, practitioners with relevant competence (e.g., nurses, physicians, physiotherapists and a social worker) and researchers. The co-design process is described in detail elsewhere [11]. The online support is hosted on a national health portal called 1177. On this platform, Swedish citizens can access their medical records and book appointments with health care professionals. The platform also includes professionally developed and reviewed information about diseases, treatments, and prevention. Despite the platform's widespread use, we have identified challenges in offering online support, specifically to informal carers of people with HF. One of these challenges concerns that even though carers who have used the online support identified how it could be valuable for them, they may find it overwhelming to go through all content, which they perceive as comprehensive (unpublished work). A more tailored solution could, for example, be time-efficient by helping carers find material or answers they need or wish for more easily.

1.3 Opportunities and challenges of large language models for tailored support to carers

LLMs have gained widespread recognition, especially through the introduction of the generative Artificial Intelligence (AI) model ChatGPT introduced by OpenAI in November 2022. LLMs can be made useful in a wide variety of user scenarios, such as interactions with humans in health care settings [12,13] where it may even be able to provide higher quality and more empathetic answers than physicians in some cases [14]. Benefits of using LLMs could include more personalized medicine and care, e.g., information and advice adjusted to fit different (health) literacy levels [15,16] and life situations. Furthermore, it could be helpful for avoiding 'information overload' [17], i.e., provide comprehensive answers to personal questions. In relation to its use within a health care context, concerns about its ability to align with human values and ethics have been raised [18,19]. The use of LLMs with an interface such as ChatGPT has also been considered to involve risks of, for example, giving biased answers [19] and, contrary to what is mentioned above, being unable to be empathetic in communication with users [15]. Moreover, when put under pressure, a LLM used as a chatbot has been noted to provide false answers, i.e., hallucinations [12]. In relation to this risk, De Angelis et al. (2023) [18], pp.5 have pointed out that: "ChatGPT's ability to follow users' instructions is a double-edged sword: on one hand, this approach makes it great at interacting with humans, on the other hand being submissive 'ab origine' exposes it to misuse, for example by generating convincing human-like misinformation.".

2 Objective, Method and Planning

To address the need for more tailored and personalized support for informal carers of persons with HF, we wish to explore the potential of using an LLM in conversations to provide valid and reliable answers to carers' specific questions and concerns. In this paper we describe the development of a prototype for such a conversational agent and outline the approach to evaluating the prototype. We also present the preliminary findings from our initial testing sessions.

2.1 Creating textual data for a conversational agent using the co-designed content

The online content on 1177 comprises 15 modules, each addressing a specific topic and consisting of several separate web pages. Topics include providing information and input concerning HF, what it could mean to be a carer, what support is available for carers, emotional reactions related to living with someone with a serious illness, intimacy and sexuality when living with someone with a heart condition, and end-of-life considerations. In total, the online support contains 83 separate web pages with text, videos, links, and images. The material in the support programme was converted to text, from which relevant parts were selected for the inclusion in the conversational agent. For instance, we discarded image captions and sections only consisting of links to webpages for further reading.

2.2 Using Retrieval-Augmented Generation

In addition to the above-mentioned risks with current LLMs, another significant challenge with them is their reliance on the vast amounts of pre-existing training data. If one wants to apply an LLM in a more specific context, or in a domain with specialized data, these models will 'out of the box' probably not be able to deliver the appropriate output, because they have no 'knowledge' of the domain. To address this gap, Retrieval-Augmented Generation (RAG) has emerged as a promising solution [13]. RAG integrates the power of LLMs with the precision of retrieval-based systems, enabling the incorporation of specialized domain data into the generation process. This hybrid approach leverages the strengths of both paradigms: the generative prowess of LLMs and the context-specific retrieval of information in the given domain. Here we describe the development of such a system in the domain of an online support for informal carers of persons with HF.

The concept of RAG can be broken down into two parts: The retrieval component and the generation component [20]. The first component applies methods to index, search and fetch relevant documents given the user's query, while the generation component incorporates the information provided by the retrieval component together with the user query and prompting instructions to generate a response from (for instance) an LLM. However, the structure and inner workings of a RAG-pipeline, employing the two components, can be realized in many ways depending on the needs of the domain at hand. For instance, Zhao et al (2024) propose a taxonomy of four different types of

RAG foundations: Query-based, Latent Representation-based, Logit-based, and Speculative [21]. To our knowledge, the Query-based foundation is the most widely adopted due to its flexible nature. In this work we focus on this approach, and when referring to RAG, we therefore mean a Query-based RAG. For a description of the other pipeline types, see Zhao et al [21].

2.3 Technical description

From a software development point of view, a clear goal of the project has been to create a highly modular system with components that could easily be replaced or modified during the development process. While a basic system could be implemented with the help of pre-existing Python libraries such as LangChain or LlamaIndex, they (during the time of development) lacked the chunking system described in the following section. We therefore opted to not use any of these libraries. This section gives a brief overview of the parts that are implemented, and some suggestions for further additions. Both the embedding models and LLM can easily be replaced. Until now, we have been using the text-embedding-ada-002 and GPT-4, provided via Microsoft Azure.

Retriever
The retriever component serves as a fundamental mechanism for sourcing relevant information from a large corpus of data, which is subsequently used by the generator component to construct responses. The retriever's primary function is to quickly and efficiently query a vast dataset and return the most relevant documents or text passages in response to a user's input.

Vector database
The primary purpose of a vector database is to facilitate the efficient and effective retrieval of semantically relevant information. This is achieved by transforming textual data into high-dimensional vector embeddings, which capture the underlying semantic meanings of the text. The vector database stores these embeddings, allowing for rapid and accurate similarity searches based on these semantic representations. By storing text as vector embeddings, a vector database enables the system to perform high-dimensional similarity searches efficiently. However, how the text is stored in the database has to be considered. A naive approach is to select spans of N characters, words, or sentences in the text, and embed each span as a separate entry in the database. However, this will often give suboptimal search results. When text passages are arbitrarily divided, important contextual links between segments are lost. For example, a detailed explanation of a concept might be split into two parts, with the first part ending with a crucial introductory explanation and the second part beginning with the core details. If these segments are stored as separate entries, without any relational links, a query relevant to this topic might retrieve only one of the segments.

Instead, we use a "chunking" method with the purpose of keeping similar topics or themes found in the source text as individual units. This process involves generating sentence-level embeddings. These embeddings are subsequently processed using a

sliding-window algorithm to evaluate the relationships between consecutive sentences. The key metric employed is the cosine distance between the embeddings of adjacent groups of sentences. This distance measures the semantic divergence between sentences. When the cosine distance surpasses a predefined threshold, it indicates a significant shift in semantic content, prompting the algorithm to group these sentences into the same chunk. Each chunk thus represents a coherent concept or idea that extends over multiple sentences. The chunking process is visualized in Figure 1. For instance, if chunk #7 contains information about training exercises suitable for someone with HF, and the database search finds a good match with some group of sentences in the chunk, 'all' the sentences from the chunk will be delivered as the search answer. This provides complete context given the user query.

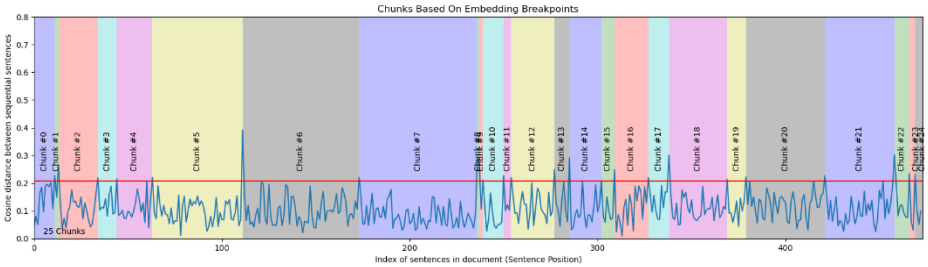

Fig. 1. Vizualisation of the chunking process on example data. When the cosine distance between adjacent groups of sentences exceeds the threshold, a new chunk is initiated.

Contextualiser

The main purpose of this component is to improve how the system understands and responds to user inputs. This component rewrites user input from the chat history, which is important for several reasons. First, it helps clarify unclear or vague inputs from users, making it easier for the system to identify what the user is asking and to fetch the most relevant information. This clarity is crucial for accurate information retrieval. Additionally, by incorporating relevant details from earlier in the conversation, the contextualizer improves the quality and appropriateness of the system's responses. In our implementation this is achieved by prompting an LLM with a combination of the user query and the conversation history. In the prompt we instruct the model to clarify the user input, both regarding any references to earlier exchanges, and vague wordings that are hard to find relevant passages for in the vector database.

Generator

The purpose of this component is to synthesize responses based on the information retrieved during the retrieval step. In our system, this consists of a conversational prompt that contains the retrieved information along with detailed instructions on how the LLM should respond, its role in the conversation, terms to avoid, etc. Since the focus of the project so far has been on the retrieval part and ensuring the system responds with accurate information, not much development has occurred in this area. Moving forward, additional components could be added to this step, such as

functionality to handle conversation states and a post-processing module to correct factual errors or other undesirable responses from the LLM.

2.4 Approach to evaluation of the prototype

To test the prototype, we are applying what could be referred to as a 'human-in-the-loop approach' [19] to qualitatively evaluate answers that the model produces. Currently evaluation includes a fixed number of questions (n17) relating to different topics/modules in the support programme. The intention is to have the same questions in each testing session. This is to balance the potential to identify patterns of desired and undesired features in the answers (e.g., hallucinations, tone, bias, etc.) with being able to conduct the necessary number of testing sessions without them becoming too time-consuming. Each question is 'asked' two times, and answers are evaluated by the model's choice of context for answers (precision and recall of context) and the correctness of the answers, and if any, the type of inaccuracies in answers (answer relevancy and faithfulness). This approach is inspired by Es, James, Espinosa-Anke and Schockaert [22]. The answers are compared, and determination is made as to which answer is considered qualitatively the best. Each testing session is summarized and documented and is discussed with the technicians who then can adjust parameters in the model. This approach allows both an iterative and structured process.

Initially evaluation has involved two researchers (HA, AS) who both are registered nurses and together have expertise in Swedish health care organization and more specifically also in HF and in informal caregiving. Both have been involved in the development of the support programme used for context and one of the researchers (HA) is very well familiar with the material since she also was involved in building the programme online.

3 Preliminary findings and expected contributions

In relation to our initial evaluations, we have so far implemented the above described contextualizer. We have also adjusted the model's temperature to make it less creative (hence more stable in each answer) and moreover increased the number of contexts that the model can use to support more complete answers. We have identified that our model is able to provide relevant answers with no obvious or harmful hallucinations. The model has the potential to be 'stable', in the sense that it provides similar answers when asked the same question twice. Providing predictable answers can enhance the perception of accuracy, which could also imply trustworthiness [13]. Current challenges lie in 'helping' the model to choose more accurate context from the online support as we can see that some types of questions seem to support relevant choice of context more accurately than others. Another challenge is to find a way to help the model adopt the right tone in its responses, providing carers with empathetic answers that consider the foundational values identified through the co-design process. Being able to provide answers that are correct and trustworthy as well as empathetic can be deemed especially

important in a health care context [23,12]. Including healthcare professionals in the evaluation process of these aspects, as we have done, is therefore considered pivotal.

When we have a more refined prototype, we are planning to continue with a co-design process including potential end-users (informal carers), designers, and researchers to further develop and study interactions with the model to investigate how to design a system (chatbot) that takes advantages of technology without risking important values in health care. This can be important to increase the likelihood that the effort put into creating relevant support will be implemented in clinical settings, and thereby contribute to addressing future challenges with resource shortages in health care and welfare.

Acknowledgments. We wish to thank Erik Allemann, who has been invaluable in both reasoning and hands-on work in the project. This project is supported by a research grant from the strategic research area eHealth of Linköping University and Linköping University Hospital.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. WHO (2022). Ageing and health. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health. Accessed 7th Nov 2024.
2. McDonagh, T. A., Metra, M., Adamo, M., Gardner, R. S., Baumbach, A., Böhm, M., et al. (2021). 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur. Heart J., 42*(36), 3599-3726, doi:10.1093/eurheartj/ehab368.
3. European Commission (2021). *Study on exploring the incidence and costs of informal long-term care in the EU*: Publications Office of the European Union.
4. Nicholas Dionne-Odom, J., Hooker, S. A., Bekelman, D., Ejem, D., McGhan, G., Kitko, L., et al. (2017). Family caregiving for persons with heart failure at the intersection of heart failure and palliative care: a state-of-the-science review. *Heart Fail. Rev.*, doi:10.1007/s10741-017-9597-4.
5. Kitko, L., McIlvennan, K.C., Bidwell T.J., Dionne-Odom J., N., Dunlay M. S., Lewis M.L., Meadows, G. Sattler L.P., E., Schulz, R., Strömberg, A. (2020). Family Caregiving for Individuals With Heart Failure: A Scientific Statement From the American Heart Association. *Circulation, 141*(22), e864-e878, doi:doi:10.1161/CIR.0000000000000768.
6. Yang, M., Kondo, T., Adamson, C., Butt, J. H., Abraham, W. T., Desai, A. S., et al. (2023). Knowledge about self-efficacy and outcomes in patients with heart failure and reduced ejection fraction. *Eur. J. Heart Fail.*, doi:10.1002/ejhf.2944.
7. Bangerter, L. R., Griffin, J. M., & Dunlay, S. M. (2019). Positive Experiences and Self-Gain among Family Caregivers of Persons with Heart Failure. [Article]. *Gerontologist, 59*(5), e433-e440, doi:10.1093/geront/gny162.
8. Grant, J. S., & Graven, L. J. (2018). Problems experienced by informal caregivers of individuals with heart failure: An integrative review. *Int. J. Nurs. Stud., 80*, 41-66, doi:10.1016/j.ijnurstu.2017.12.016.

9. Choi, S., Kitko, L., Hupcey, J., & Birriel, B. (2021). Longitudinal family caregiving experiences in heart failure: Secondary qualitative analysis of interviews. [Article]. *Heart Lung, 50*(5), 627-633, doi:10.1016/j.hrtlng.2021.05.002.
10. WHO (2021). Global strategy on digital health 2020-2025. Geneva: World Health Organization.
11. Allemann, H., Andréasson, F., Hanson, E., Magnusson, L., Jaarsma, T., Thylén, I., et al. (2023). The co-design of an online support programme with and for informal carers of people with heart failure: A methodological paper. *J Clin Nurs*, doi:10.1111/jocn.16856.
12. Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N. Engl. J. Med., 388*(13), 1233-1239, doi:10.1056/NEJMsr2214184.
13. Zhang, P., & Kamel Boulos, M. N. (2023). Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges. [Review]. *Future Internet, 15*(9), doi:10.3390/fi15090286.
14. Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., et al. (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med, 183*(6), 589-596, doi:10.1001/jamainternmed.2023.1838.
15. Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthc (Basel), 11*(6), doi:10.3390/healthcare11060887.
16. Dunn, A. G., Shih, I., Ayre, J., & Spallek, H. (2023). What generative AI means for trust in health communications. *J Commun Healthc, 16*(4), 385-388, doi:10.1080/17538068.2023.2277489.
17. Skjuve, M., Brandtzaeg, P. B., & Følstad, A. (2024). Why do people use ChatGPT? Exploring user motivations for generative conversational AI*., 29*(1), doi:10.5210/fm.v29i1.13541.
18. De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., et al. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front. Public Health, 11*, 1166120, doi:10.3389/fpubh.2023.1166120.
19. Ferrara, E. (2023). Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday, 28*(11).
20. Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2024). Evaluation of Retrieval-Augmented Generation: A Survey. arXiv:2405.07437 [cs.CL].
21. Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., et al. (2024). Retrieval-Augmented Generation for AI-Generated Content: A Survey. arXiv:2402.19473 [cs.CV].
22. Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. RAGAS: Automated Evaluation of Retrieval Augmented Generation. In O. D. C. Nikolaos Aletras (Ed.), *EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of System Demonstrations, St. Julians, Malta, 2024* (Vol. Proceedings of the 18th conference of the European chapter of the Association for Computational Linguistics: System Demonstrations, pp. 150-158): Association for Computational Linguistics.
23. Wang, C., Liu, S., Yang, H., Guo, J., Wu, Y., & Liu, J. (2023). Ethical Considerations of Using ChatGPT in Health Care. *J. Med. Internet Res., 25*, e48009, doi:10.2196/48009.