Building Certified Medical Chatbots: Overcoming Unstructured Data Limitations with Modular RAG

Leonardo Sanna*, **Patrizio Bellan***, **Simone Magnolini***,
Marina Segala*, **Saba Ghanbari Haez*†**, **Monica Consolandi***,
Mauro Dragoni*
*Fondazione Bruno Kessler, Trento (ITALY)
[lsanna, pbellan, magnolini, msegala, sghanbarihaez, mconsolandi, dragoni]@fbk.eu
†Free University of Bozen, Bozen (ITALY)

Abstract

Creating a certified conversational agent poses several issues. The need to manage fine-grained information delivery and the necessity to provide reliable medical information requires a notable effort, especially in dataset preparation. In this paper, we investigate the challenges of building a certified medical chatbot in Italian that provides information about pregnancy and early childhood. We show some negative initial results regarding the possibility of creating a certified conversational agent within the RASA framework starting from unstructured data. Finally, we propose a modular RAG model to implement a Large Language Model in a certified context, overcoming data limitations and enabling data collection on actual conversations.

Keywords: Conversational Agent, Digital Health, Retrieval-Augmented Generation

1. Introduction

In recent research, the demonstrated effectiveness of conversational agents and Large Language Models (LLMs) has expanded to include tasks that were once thought unlikely, marking a notable advancement in their capabilities. For instance, within the digital health area, it has been shown that conversational agents can provide emotional support to patients, possibly more efficiently than a standard interaction between a physician and a patient (Suppadungsuk et al., 2023; Ayers et al., 2023; Fadhil and Gabrielli, 2017).

In this paper, we present the work-in-progress of a project to create a conversational agent capable of providing certified medical information regarding pregnancy and the first thousand days of a child's life. With the expression "*certified information*" we mean textual content generated or validated by healthcare professionals, ensuring its verifiability and alignment with the current scientific knowledge in the respective domain.In addition, an essential attribute of "*certified information*" is its predictability, indicating that, given a specific question the response would always be the same. The agent will be implemented initially in Italian only.

To the best of our knowledge, there are no examples in the literature where conversational agents have been employed to aid patients in this particular field. Likewise, there are no examples of an Italian medical conversational solution capable of delivering certified medical advice. Current applications of conversational agents within the healthcare industry suffer problems of data certification and accuracy (Srivastava and Singh, 2020; Jungmann et al., 2019; Swick, 2021); consequently, there is a lack of evidence of their efficacy in clinical contexts (Bibault et al., 2019). Therefore, medical conversational agents are often limited to assisting medical staff rather than patients (Minutolo et al., 2022), or used as a tool to help diagnostics (Ni et al., 2017; Verma et al., 2022) and integrate the search for medical assistance (Soprano et al., 2023; Polignano et al., 2020). Also, the trust towards deploying this kind of technology is an aspect that needs to be addressed, as it directly impacts the potential efficacy (Seitz et al., 2022; Martens et al., 2024; Laumer et al., 2019). Creating a certified medical conversational agent would address some of these significant issues, especially when deploying these agents in the public sector.

In the following sections, we outline the main issues we have found in our workflow so far, summarize some text insights, and explore the possible solutions for the upcoming steps.

2. Dataset and Conversational Design

Our current corpus contains approximately 1300 texts sourced from verified medical channels [1], focusing predominantly on *informational cards*. These cards offer brief yet detailed medical information on various topics, providing verified advice on conditions, treatments, and procedures. They are commonly used in FAQ sections, offering patients reliable information without direct interaction with healthcare professionals.

However, working with certified information

[1]The content is sourced from texts curated by the Obstetrician Department of the Hospital of Trento and from UPPA, a reputable child care website https://www.uppa.it/

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 124–130
20 May, 2024.

poses challenges, particularly when adapting it for conversational use. Indeed, our dataset is not designed for integration into a conversational framework. One of the main challenges is that that editing options are severely limited when dealing with certified medical information. The optimal approach would be to use the texts in their original form to preserve their certification. Yet, they often tend to be excessively lengthy and informationally dense for effective conversation use.

Moreover, we must consider that extracting information from these texts is complicated due to their highly discursive nature. Automatic segmentation often results in imprecise responses, occasionally leading to grammatical inaccuracies since segments are extracted from an existing discursive context. There is also a notable risk of encountering information gaps, despite the fact they are densely packed with information. In fact, in a certified context, all the deliverable information must be present explicitly in the text; even the simplest inferences are impossible since they would require certification, ensuring that they correspond to correct medical knowledge.

Lastly, our informational cards come from specialized sites and are meant to be instructive, so they often use medical vocabulary. This characteristic complicates the process of generating additional data, especially when generating questions for training a conversational agent. Medical jargon is indeed quite influential in affecting question generation, often leading to the creation of improbable examples.

While using an LLM could compensate for the lack of conversational data, our requirement to provide reliable information without any changes prevents us from directly using an LLM for user interaction. LLMs' erratic nature doesn't align with the need for stable and predictable output in certified information contexts.

3. Workflow: Creating a RASA Chatbot

We began with an existing COVID-19 FAQ chatbot (Lucianer et al., 2022) named *Covibot*. Since this agent was realized within the RASA framework[2], we used RASA to create our first test conversational agent, focusing our efforts on the Natural Language Understanding (NLU) module, as its performance significantly impacts the overall conversation flow. This first experiment was therefore only focused on a simple classification pipeline, with the goal of associating each intent with a specific reply.

Using our data, we automated the generation of example questions with GPT-4 via the OpenAI ChatGPT API. We segmented the texts into shorter paragraphs using GPT-4 to generate the briefest meaningful paragraphs while considering the textual excerpt's topic. We then prompted the model to generate three simple questions for each text. These questions were then associated with specific intents linked to their corresponding answers.

Since RASA intent classifier[3] also supports custom word embeddings, we created a model (Le and Mikolov, 2014) from our data. While RASA supports various embedding techniques, support for highly specific domains, like ours, is limited [4].

Our custom embedding model showed promising results in improving the conversational agent's performance in an initial sample of around 50 intents and 1500 total examples. Performance assessment was conducted by partitioning the dataset into 80% for training and 20% for testing, progressively increasing the number of examples during the training phase. In the graph shown in Figure 1, the *UPPA* configuration uses the embeddings of our dataset; the *Spacy* configuration uses pre-trained Spacy embeddings [5] for Italian, whereas the *Base* configuration uses no pre-trained embeddings.

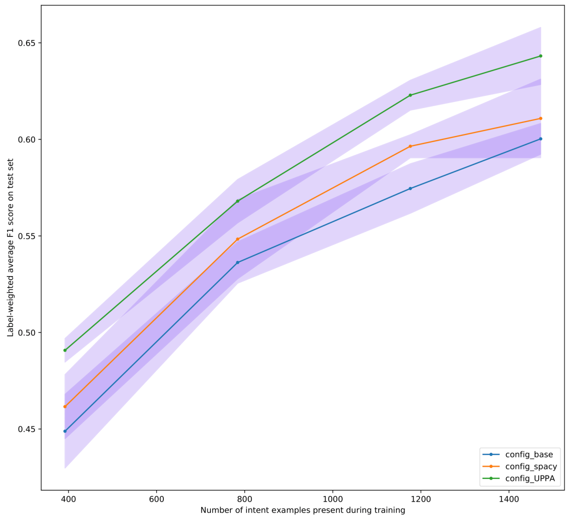

Figure 1: Comparison of custom word embedding impact on our first trained model.

Subsequently, we expanded our dataset to include 4500 intents and their corresponding answers. However, this dataset extension resulted in a noticeable decline in the RASA model's perfor-

[2]https://github.com/RasaHQ/rasa

[3]https://rasa.com/blog/introducing-dual-intent-and-entity-transformer-diet-state-of-the-art-performance-on-a-lightweight-architecture/

[4]Support is limited to Gensim embeddings: https://rasa.com/blog/custom-gensim-embeddings-in-rasa/

[5]https://spacy.io/usage/models

mance. This second evaluation assessed RASA's capacity for predicting the right intent class and, consequently, giving the right answer for each of the main topics in our dataset. Figure 2 illustrates the model's performance, which has been proven to be below acceptable standards.

Our RASA chatbot could classify correctly only an average of 28% of intents. Moreover, the model is quite sparse, with an average confidence on correct predictions of 0.27. Also, our custom embeddings lost their relevance in enhancing the traning; the model proved indeed highly sensitive to minor rephrasing operations, where even a small alteration in a training sentence could easily cause the model to fail.

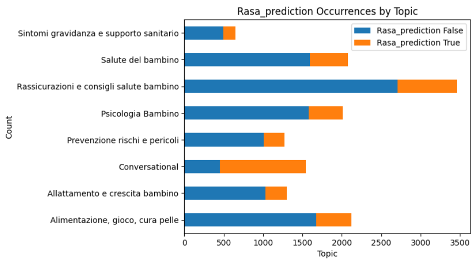

Figure 2: RASA performance across the main topics with 4500 intents. In orange, the correct replies.

4. Data Limitations

Considering the outcome of the first test, some additional considerations on data quality are necessary. The data that we have is all unstructured text. These texts have great stylistic heterogeneity, even within the same source, combined with great semantic homogeneity, all being part of a specific medical domain. This dual characteristic makes topic modeling problematic; we have currently tried different types of approaches, ranging from the more classic Latent Dirichlet Allocation (LDA) (Blei et al., 2003), keywords (Bondi and Scott, 2010; Gabrielatos and Marchi, 2011), and BERTopic [6], which has recently been shown as one of the most effective topic modeling techniques (Gan et al., 2023; Egger and Yu, 2022). Regardless of the method we used, we found that semantic areas in our data are always rather fragmented because of the great ramifications of sub-topics, even within the same thematic areas. For instance, in Figure 3 we show the topics found using BERTopic. The two main semantic macro-areas consist of one encompassing documents related to the newborn and another containing documents regarding pregnancy. Nevertheless, the extensive thematic fragmentation within these areas poses a significant challenge in training conversational agents to effectively associate intents with their respective topics.

Figure 3: Visualization of the topics found using BERTopic.

We would need fine-grained annotation on topics and other relevant linguistic aspects to effectively deliver certified information. Yet, since our semantic areas frequently overlap, automatic topic extraction does not produce qualitatively acceptable document groups. This means an in-depth qualitative analysis of the automatic topic extraction is required before annotation, also to highlight other elements like named entities and hardly quantifiable textual features (Hunston, 2004) such as relevant pragmatic aspects for medical conversations.

Moreover, having only unstructured texts is a substantial problem for RASA, since its intent classifier is designed to work with Named Entity Recognition. The existing state-of-the-art approaches such as MedBert (Egger and Yu, 2022) are also not focused on question answering nor entity recognition on unstructured texts like ours. Also, we have to consider that most of the approaches regarding medical conversational agents, especially for question answering (Kacupaj, 2022) have a knowledge-based approach (Dayal et al., 2023; Minutolo et al., 2017), which also requires annotated data.

5. Future Work: Annotation and RAG

In our case, the data quality is a major issue that might have different solutions. Looking at previous approaches, it becomes evident that using certified sources in a conversational context, even a basic one, necessitates a considerable amount of contextual information (Kadariya et al., 2019; Fenza et al., 2023; Alloatti et al., 2021). Hence, developing an annotation methodology is essential to improve the

[6] https://doi.org/10.48550/arXiv.2203.05794

performance of the conversational agent, irrespective of the chosen framework. Certain information required for building our knowledge base can only be obtained through fine-grained annotation. However, this process proves to be time-consuming, and its success remains uncertain.

Alternatively, an immediately implementable strategy could involve using an LLM to address the discursive aspects, while incorporating certified sources from our database. LLMs, especially ChatGPT, have proven to be reasonably reliable, at least on basic questions about medical care (Mihalache et al., 2024; Cheong et al., 2023; Cascella et al., 2023). In addition to this, techniques such as Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Karpukhin et al., 2020) can be used to increase the LLM's ability to correctly answer a question, minimizing hallucinations (Martino et al., 2023). Essentially, the user's request and the additional knowledge work together to guide the Language Model's response. This prevents the model from giving inaccurate information when it does not have it readily available. However, as we said before, in a certified context we cannot rely on an LLM to provide the information to a patient, since it is impossible to certify the model output because of its stochastic nature.

Furthermore, a key issue in the standard RAG approach is the possible mismatch between the user's query and the correct documents. Typically, RAG involves the transformation of a user query into a vector embedding representation, which is then used to assess semantic similarity among the repository of documents. However, the vector of the query and documents' vectors might be significantly different within the semantic space; this discrepancy introduces a consequential constraint, as it may lead to the exclusion of relevant documents during the retrieval process.

Modular RAG with HyDE We are working within the Hypothetical Document Embeddings (HyDE) framework (Gao et al., 2023) to address these two limitations. HyDE is a novel approach recently introduced that operates unsupervised. In a nutshell, HyDE uses an LLM to produce a hypothetical document (HyDoc) based on input queries and then it uses the HyDoc to retrieve the information from the certified repository. Despite the hallucinations that might be present in the HyDoc, the generated text should lie in the semantic space in a neighborhood of similar real documents that contain the correct and certified answer to provide to the user.

In the pipeline that we are implementing, given a specific question, we generate a hypothetical document that is used to query the certified document repository. Then, the *paraphrase-multilingual-mpnet-base-v2* Bi-Encoder model (Reimers and Gurevych, 2019) is used to retrieve the documents. However, the Bi-Encoder performs optimally when estimating similarity between documents of similar sizes. Given that our HyDoc and the certified documents may differ significantly in length, we use a cross-encoder, i.e. *ms-marco-MiniLM-L-6-v2* [7], to re-rank the retrieved documents and refine the list. Finally, the selected documents are used to augment the initial prompt, and a *Guard-Rail* module [8] ensures that the LLM reply is short enough. As shown in Figure 4, the conversational agent's final answer contains the documents' textual summary (80-120 words) and the pointers to the original certified sources. Although our RAG model represents a compromise, it facilitates testing in a production environment, enabling data collection from authentic conversations and facilitating data augmentation.

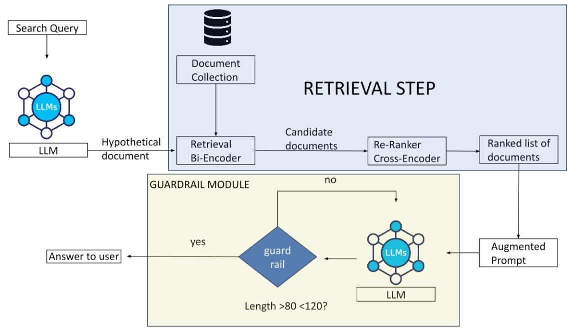

Figure 4: An overview of the RAG model we are implementing.

Preliminary testing with GPT-4-turbo on 100 user-generated questions yielded promising results, retrieving relevant documents in over 85% of cases. On the same test set, the RASA model achieved only 13% correct answers, with approximately on-topic responses in 25% of cases and off-topic replies in over 60% of cases. In terms of HyDoc generation, GPT-4-turbo demonstrated the ability to produce pertinent responses in over 95% of examples. Given that the initial module impacts the entire model, additional investigation is required to assess open-source LLMs [9] performance, both in generating HyDocs and in the quality of document summarization.

6. Acknowledgments

This paper is part of the project TrustAlert which has received funding from the Fondazione Compagnia

[7] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2

[8] https://doi.org/10.48550/arXiv.2402.15911

[9] For instance: https://huggingface.co/swap-uniba/LLaMAntino-2-70b-hf-UltraChat-ITA

San Paolo and Fondazione CDP under the “Artificial Intelligence” call. We acknowledge the support of the PNRR project INEST - Interconnected North-East Innovation Ecosystem (ECS00000043), under the NRRP MUR program funded by the NextGenerationEU. We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

7. Bibliographical References

Francesca Alloatti, Alessio Bosca, Luigi Di Caro, and Fabrizio Pieraccini. 2021. Diabetes and conversational agents: the aida project case study. *Discover Artificial Intelligence*, 1:1–21.

JW Ayers, A Poliak, M Dredze, 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.

Jean-Emmanuel Bibault, Benjamin Chaix, Pierre Nectoux, Arthur Pienkowski, Arthur Guillemasé, and Benoît Brouard. 2019. Healthcare ex machina: Are conversational agents ready for prime time in oncology? *Clinical and translational radiation oncology*, 16:55–59.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Marina Bondi and Mike Scott, editors. 2010. *Keyness in Texts*, volume 41. John Benjamins Publishing.

Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. *Journal of Medical Systems*, 47(1):33.

Ryan Chin Taw Cheong, Kenny Peter Pang, Samit Unadkat, Venkata Mcneillis, Andrew Williamson, Jonathan Joseph, Premjit Randhawa, Peter Andrews, and Vinidh Paleri. 2023. Performance of artificial intelligence chatbots in sleep medicine certification board exams: Chatgpt versus google bard. *European Archives of Oto-Rhino-Laryngology*, pages 1–7.

Raghav Dayal, Parv Nangia, Surbhi Vijh, Sumit Kumar, Saurabh Agarwal, and Shivank Saxena. 2023. Development of chatbot retrieving fact-based information using knowledge graph. In *Proceedings of International Conference on Recent Innovations in Computing: ICRIC 2022, Volume 1*, pages 153–164. Springer.

Roman Egger and Joanne Yu. 2022. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7:886498.

Ahmed Fadhil and Silvia Gabrielli. 2017. Addressing challenges in promoting healthy lifestyles: the al-chatbot approach. In *Proceedings of the 11th EAI international conference on pervasive computing technologies for healthcare*, pages 261–265.

Giuseppe Fenza, Francesco Orciuoli, Angela Peduto, and Alberto Postiglione. 2023. Healthcare conversational agents: Chatbot for improving patient-reported outcomes. In *International Conference on Advanced Information Networking and Applications*, pages 137–148. Springer.

Costas Gabrielatos and Anna Marchi. 2011. Keyness: Matching metrics to definitions. In *Theoretical-methodological Challenges in Corpus Approaches to Discourse Studies and Some Ways of Addressing Them*.

Lin Gan, Tao Yang, Yifan Huang, Boxiong Yang, Yami Yanwen Luo, Lui Wing Cheung Richard, and Dabo Guo. 2023. Experimental comparison of three topic modeling methods with lda, top2vec and bertopic. In *International Symposium on Artificial Intelligence and Robotics*, pages 376–391. Springer.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1762–1777. Association for Computational Linguistics.

Susan Hunston. 2004. Counting the uncountable: Problems of identifying evaluation in a text and in a corpus. *Corpora and discourse*, 9:157–188.

Stefanie Maria Jungmann, Timo Klan, Sebastian Kuhn, and Florian Jungmann. 2019. Accuracy of a chatbot (ada) in the diagnosis of mental disorders: comparative case study with lay and expert users. *JMIR formative research*, 3(4):e13863.

Endri Kacupaj. 2022. *Conversational Question Answering over Knowledge Graphs with Answer Verbalization*. Ph.D. thesis, Universitäts-und Landesbibliothek Bonn.

Dipesh Kadariya, Revathy Venkataramanan, Hong Yung Yip, Maninder Kalra, Krishnaprasad Thirunarayanan, and Amit Sheth. 2019. kbot: knowledge-enabled personalized chatbot for asthma self-management. In *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 138–143. IEEE.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Sven Laumer, Christian Maier, and Fabian Tobias Gubler. 2019. Chatbot acceptance in healthcare: Explaining user adoption of conversational agents for disease diagnosis. In *Proceedings of the 27th European Conference on Information Systems (ECIS)*, Stockholm & Uppsala, Sweden.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Marzia Lucianer, Francesca Perini, Giulia Malfatti, Lorenzo Gios, Alessandro Bacchiega, Claudio Giuliano, Andrea Nicolini, Stefano Forti, Roberta Corazza, Veronica Tretter, et al. 2022. A technology-enabled, public-driven, and multichannel communication strategy during covid19 pandemic in the province of trento, italy. *JOURNAL OF MEDICAL INTERNET RESEARCH*.

Marijn Martens, Ralf De Wolf, and Lieven De Marez. 2024. Trust in algorithmic decision-making systems in health: A comparison between ada health and ibm watson oncology. *Cyberpsychology*, 18(1).

Ariana Martino, Michael Iannelli, and Coleen Truong. 2023. Knowledge injection to counter large language model (llm) hallucination. In *The Semantic Web: ESWC 2023 Satellite Events*, pages 182–185, Cham. Springer Nature Switzerland.

Andrew Mihalache, Ryan S Huang, Marko M Popovic, and Rajeev H Muni. 2024. Chatgpt-4: an assessment of an upgraded artificial intelligence chatbot in the united states medical licensing examination. *Medical Teacher*, 46(3):366–372.

Aniello Minutolo, Emanuele Damiano, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2022. A conversational agent for querying italian patient information leaflets and improving health literacy. *Computers in Biology and Medicine*, 141:105004.

Aniello Minutolo, Massimo Esposito, and Giuseppe De Pietro. 2017. A conversational chatbot based on kowledge-graphs for factoid medical questions. In *SoMeT*, pages 139–152.

Lin Ni, Chenhao Lu, Niu Liu, and Jiamou Liu. 2017. Mandy: Towards a smart primary care chatbot application. In *International symposium on knowledge and systems sciences*, pages 38–52. Springer.

Marco Polignano, Fedelucio Narducci, Andrea Iovine, Cataldo Musto, Marco De Gemmis, and Giovanni Semeraro. 2020. Healthassistantbot: A personal health assistant for the italian language. *IEEE Access*, 8:107479–107497.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Lennart Seitz, Sigrid Bekmeier-Feuerhahn, and Krutika Gohil. 2022. Can we trust a chatbot like a physician? a qualitative study on understanding the emergence of trust toward diagnostic chatbots. *International Journal of Human-Computer Studies*, 165:102848.

Michael Soprano, Kevin Roitero, Vincenzo Della Mea, Stefano Mizzaro, et al. 2023. Towards a conversational-based agent for health services. In *Proceedings of the Italia Intelligenza Artificiale-Thematic Workshops co-located with the 3rd CINI National Lab AIIS Conference on Artificial Intelligence (Ital IA 2023)*, pages 278–283.

Prakhar Srivastava and Nishant Singh. 2020. Automatized medical chatbot (medibot). In *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, pages 351–354. IEEE.

S. Suppadungsuk, C. Thongprayoon, J. Miao, et al. 2023. Exploring the potential of chatbots in critical care nephrology. *Medicines*, 10:58.

Robert K Swick. 2021. The accuracy of artificial intelligence (ai) chatbots in telemedicine. *Journal of the South Carolina Academy of Science*, 19(2):17.

Shreya Verma, Mansi Singh, Ishita Tiwari, and BK Tripathy. 2022. An approach to medical diagnosis using smart chatbot. In *International Conference on Computational Intelligence in Pattern Recognition*, pages 43–56. Springer.