Holistic Multi-layered System Design for Human-Centered Dialog Systems

Roland Oruche
Department of EECS
University of Missouri
Columbia, USA
rro2q2@umsystem.edu

Rithika Akula
Department of EECS
University of Missouri
Columbia, USA
rahvk@umsystem.edu

Sai Keerthana Goruganthu
Department of EECS
University of Missouri
Columbia, USA
sgmhz@umsystem.edu

Prasad Calyam
Department of EECS
University of Missouri
Columbia, USA
calyamp@missouri.edu

***Abstract*—Human-centered techniques have become integral towards the rapid adoption and interaction of dialog systems. Despite this, most dialog system designs entail sparse human-centered approaches in the development lifecycle. Furthermore, no existing work has provided a holistic view on the recent advancements in dialog systems with human-centered techniques. In this paper, we propose a holistic multi-layered system design for human-centered dialog systems (HCDS) that considers human-centered methods in every facet of the development lifecycle. We first investigate the recent advancements on HCDS by proposing a taxonomy that classifies relevant methods and concepts in HCDS design. Next, we discuss open issues within our taxonomy and propose design requirements for enabling human-centered approaches to be distributed across end-to-end dialog system architectures. We then use the design requirements to detail our multi-layered system design in the aspects of data management, model development, and end-user application on a publication analytics platform for knowledge discovery. We exemplify our design approach through an existing HCDS and conduct a usability experiment that demonstrates the utility of the end-user application. Lastly, we list future research directions on emerging human-centered designs for increasing end-users' trust and enabling them to be more efficient in their decision-making processes.**

***Index Terms*—Dialog Systems, Human-Centered Design, Human-Computer Interaction, Natural Language Processing**

I. Introduction

Human-centered AI – which leverages human and machine intelligence for synergistic interaction and alignment – has been vital in the recent success of text-based dialog system technologies (e.g., chatbots, LLMs) within various applications such as healthcare [1] and education [2]. Despite these advancements, the majority of dialog system architectures entail sparse human-centered approaches, which in turn can compromise the safety of deployed chatbots when managing users private information and the utility in managing end-user goals in real-world applications. In addition, there is a lack of existing work that holistically presents the recent advancements on dialog systems with human-centered approaches.

This work is supported by the National Science Foundation under awards: OAC-2006816 and OAC-2007100. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

In this paper, we propose a holistic multi-layered system design for human-centered dialog systems (HCDS). Herein, we consider human involvement to be integral in the dialog system design life cycle in the layers of data management, model development, and end-user application. First, we survey the recent advancements in HCDS design and propose a taxonomy that categorizes relevant methods/concepts under three main categories: *User-Centric Design*, *User Evaluation & Experience*, and *Governance*. Next, we discuss open issues from our proposed taxonomy and use these issues to create a set of design requirements for enabling human-centered approaches to be distributed across end-to-end dialog system architectures. By considering the open issues and requirements, we then detail our holistic multi-layered system design as a prototype that utilizes relevant human-centered methods/concepts from our proposed taxonomy. We detail how each component plays a vital role in ensuring dialog safety and utility when applied on a publication analytics application for knowledge discovery over the COVID-19 pandemic viz., *KnowCOVID-19* [3].

We exemplify our multi-system design through an existing HCDS viz., Vidura Advisor [4] for question-answering tasks and perform a KnowCOVID-19 usability experiment with 20 subjects on a set of clinical research tasks to test user performance and user experience based on a set of quantitative metrics. Through our design approach, we perform a comparative study that demonstrates how KnowCOVID-19 assisted *with* Vidura improves the overall utility of the application versus KnowCOVID-19 *without* Vidura. Lastly, we discuss future research directions on emerging HCDS designs for increasing end-users' trust and enabling them to be more efficient in their decision-making processes.

Our novel contributions are as follows:

- We propose a taxonomy that classifies relevant methods and concepts in human-centered dialog systems to help guide effective dialog system design approaches.
- We address open issues in our taxonomy and detail the design requirements for building dialog systems with distributed human-centered approaches.
- We detail our holistic multi-layered system design for the Vidura chatbot on KnowCOVID-19 and perform a usability experiment on the utility of KnowCOVID-19 via our design approach.

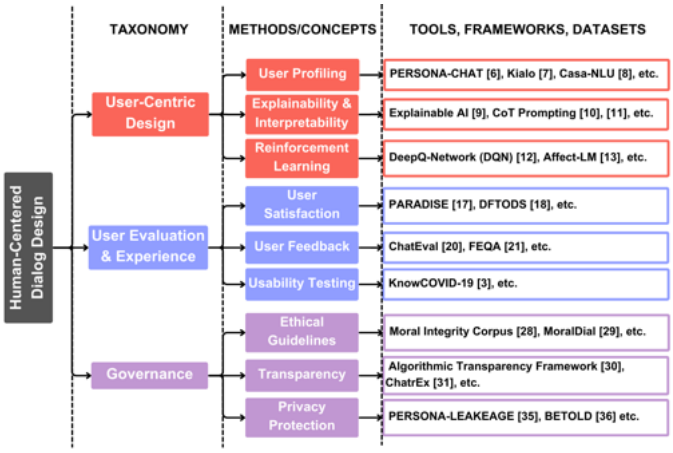

Fig. 1: Taxonomy of HCDS design: methods/concepts and tools to incorporate User-Centric Design, User Evaluation & Experience, and Governance.

Paper remainder is organized as follows: In Section II, we propose a taxonomy in HCDS design and detail the state-of-the-art methods/concepts within each taxonomy category. Section III details open issues within each taxonomy category along with each method/concept and proposes design requirements. In Section IV, we detail our holistic multi-layered system design approach for human-centered dialog systems applied to KnowCOVID-19 for knowledge discovery Finally, Section VI concludes the paper and lists future design considerations.

II. Recent Advancements in Human-Centered Dialog System Design

In this section, we survey the recent advancements in HCDS design. In Fig. 1, we propose a taxonomy in the form of a tree-diagram that encompasses *User-Centric Design*, *User Evaluation & Experience*, and *Governance* and associated methods/concepts. We show how this holistic perspective can influence the necessary design requirements to involve humans in every stage of dialog system design.

A. User-Centric Design

For user-centric chatbot design, it is essential to enable the chatbot to adapt and provide personalized, transparent, and user-driven interactions to enhance the overall user experience. We explore methods in user-centric design including user profiling, explainability and interpretability, and reinforcement learning (RL).

1) User Profiling: User-centric Design is implemented as the personalization of model responses by conditioning a generative dialog model on two important factors: (i) the respective user's profile, and (ii) the conversation's context history [5]. Authors in [6] break down user profile information into two categories of factual knowledge [7] and stylistic modifiers [8] to better incorporate personalization. The second factor in personalization is maintaining the context histories of the user to understand user input in relation to prior interactions [9].

2) Explainability & Interpretability: Building trustworthy conversational agents is crucial for user-centric design, and one of the key aspects of trustworthy AI is explainability and interpretability. Incorporating explainability in chatbot responses ensures transparency and guides to detect key drivers responsible for the decisions made by the AI system, and also to assess potential levels of bias. Authors in [10] propose frameworks for building human-centered algorithm-driven explainable AI based on extensive reviews, study probes, and saliency techniques. In recent times, Chain-of-Thought prompting in Large Language Models (LLMs) [11] has become a breakthrough in not just increasing interpretability, but also improving the chatbot performance as a whole [12].

3) Reinforcement Learning: Reinforcement learning for user-centric dialog system design involves training the system to optimize responses based on feedback from domain experts or user interactions, improving overall performance and adaptability [13]. Authors in [14] highlight the integration of reinforcement learning in dialog systems, using human feedback as a reward mechanism to enhance model performance. The work in [15] shows that it is possible to train an RL model even from small amounts of fairly reliable human feedback in tasks like German-to-English language translation. Using RL for human feedback has shown success in aligning to human preferences. For instance, authors in [16] improve LLMs by incorporating reinforcement learning from human feedback in the healthcare domain.

B. User Evaluation & Experience

In creating human-centered dialog systems that excel in practical utility and align with user expectations, it is imperative to integrate methods such as user satisfaction measurement, which assess system performance against user expectations, user feedback collection for direct insights, and usability testing to ensure real-world practicality.

1) User Satisfaction: User satisfaction in dialog systems, a crucial metric of system usability and effectiveness, is thoroughly examined in various studies. The work in [17] emphasizes measuring and modeling user satisfaction to gauge overall system performance. The approach detailed by [18] involves human evaluators using these systems and rating their experience, with data like task success rate and completion time being pivotal for predictive modeling and system improvements. The PARADISE framework [19] offers a systematic methodology for evaluating dialog strategies through user satisfaction metrics. Moreover, authors in [20] discuss a dialog framework for task-oriented dialog systems that focuses on efficient information structuring and system reasoning enhancements through module routers and Hierarchical Argument Structures.

2) User Feedback: User feedback, essential for evaluating system performance and user satisfaction, is gathered through surveys, interviews, and direct comments. The work in [21] emphasizes the importance of direct user feedback in refining

dialog systems' learning processes. They focus on human-centered evaluation, prioritizing user feedback on aspects like naturalness and task completion in dialog systems. Authors in [22] use the ChatEval toolkit for both automatic and human evaluations of conversational agents, enabling systematic comparisons. The work in [23] introduces FEQA, a QA-based metric for assessing faithfulness in summaries, comparing QA model answers derived from summaries to the original source.

3) Usability Testing: Usability testing, involving real users performing tasks under observation, is key for assessing system utility. The works in [24], [25] conduct usability tests using questionnaires to facilitate task completion, and using metrics such as time, quality, and flexibility to correlate system performance with usability. This method proved effective in enhancing initial application usability. Authors in [3] evaluate their chatbot framework through user performance and perception, providing insights on the effectiveness and user experience of chatbot-enabled search systems versus traditional search engines.

C. Governance

Recent human-centered methods have been incorporated for the assurance of designing safe and responsible dialog systems such as establishing ethical guidelines, transparency, and privacy protection.

1) Ethical Guidelines: Recent studies consider establishing or adopting ethical guidelines toward safe and robust dialog system design. The works in [26], [27] discuss the control of ethical design and must focus on eliminating data management biases and fairness and accountability in dialogue conversations with users. Other works employ ethical design principles in usability studies to study the effect that dialogue conversation [28] and language style [25] have on user perception. Ethical design considerations have also sparked the initiative in building datasets and frameworks for designing safe and responsible dialog systems that address ethical issues such as social bias [29], and morality [30], [31].

2) Transparency: In the context of dialog system design and governance, algorithmic techniques such as explainability and interpretability have been key drivers toward dialog system transparency. Authors in [32] propose the Algorithmic Transparency Framework that enables users to receive context from "black box" chatbot models that induce verifiable human reasoning in decision-making applications. The work in [33] proposes ChatrEx, to demonstrate how using explainable techniques in chatbot interfaces can improve transparency and trust. The work in [34] considers increasing transparency and human-chatbot collaboration by enabling humans to have more control in inspecting, chaining, and modifying LLM prompts for improving the quality of various task outcomes.

3) Privacy Protection: Privacy in data management and dialogue conversations is vital for ensuring governance in dialog system design. These privacy concerns have led to the investigation of potential risks such as data leakage in dialogue conversations [35] and data storage [36]. The work in [37] resolves such concerns through a detection scheme that prevents humans from being victims of social engineering. Authors in [38] create a privacy-preserving dataset for conversational breakdown detection and propose a scheme for the detection of potential breakdowns. Other works such as in [39] propose a dialog system architecture that preserves the privacy of users via an argumentation framework and adopts data protection regulations to limit the data processing of users' information.

III. Open Issues & Design Requirements

In this section, we identify several open issues within our taxonomy and associated methods and concepts. Using Table I, we summarize the issues associated with each method and concept linked to our taxonomy categories, the risks associated with them, and a potential solution for each issue in the requirements column. Following this, we elaborate on the necessary design requirements in a dialog system that unify the potential solutions from Table I for building an end-to-end dialog system with distributed human-centered approaches for all stages of the development lifecycle.

A. User-Centric Design

Building personalized dialog systems comes with challenges in the data collection and in the model-building phases. The datasets used might have used biased and/or faulty data collection techniques along with misrepresentation of user profile data which will negatively impact the dialog system by producing harmful and biased responses [40]. In the model development phase – which encompasses "human-in-the-loop" internal feedback training of the model for improving its performance, efficiency, and alignment (see Section IV-A) – developers are tasked to strike the right balance between response generalization and personalization. They might fail to completely understand the end user's preferences as they are highly singular and volatile.

To solve the issue of data bias and misrepresentation, proper data cleaning techniques must be used. To understand the level of personalization preferred by end-users, feedback should be incorporated internally within the model training by adding human-in-the-loop feedback.

B. User Experience & Evaluation

In the development of dialog systems, privacy emerges as a critical challenge, notably in crafting policies for the development team on handling sensitive data. There is a potential risk of misusing participant information, such as demographics, which can lead to privacy breaches. Additionally, user experience is at stake, as participants may inadvertently become victims of social engineering [37] by sharing personal information, compromising their privacy, and detracting from the intended experience of the dialog system.

Developers must navigate the delicate balance of collecting enough data to ensure a personalized experience while protecting user privacy, preventing the system from potentially leveraging the full richness of user data for customization and improvement. Dialogue systems must incorporate robust privacy-preserving mechanisms that protect user data, even

TABLE I: Open Issues and potential risks of each method/concept that relate to User-Centric Design, User Evaluation & Experience, and Governance for the development of human-centered dialog system applications.

	METHOD	**ISSUE**	**RISK**	**REQUIREMENT**
UCD	*User Profiling*	Data imbalance of common user demographics.	Model can learn to discriminate based on demographics and preferences.	Proper data cleaning techniques must be used.
	Explainability & Interpretability	"Black box" models make it difficult for chatbots to be transparent.	Lack of transparency can result in the chatbot misinterpreting user intent.	Model should be developed with explainability and traceability methods.
	Reinforcement Learning	Complex user feedback hinders model from fully capturing their preferences.	Develops a generalized model that overlooks user preferences.	Human-in-the-loop techniques should be incorporated in model training.
UE/UX	*User Satisfaction*	Personalized responses leaking personal user data during conversations.	Poor user experience and privacy risks deter dialog system use among users.	Data collection methods must balance personalization and user privacy.
	User Feedback	Recruitment bias may not represent the full target user population.	Non-diverse feedback can cause chatbot to favor one group over others.	A system must incorporate large-scale and diverse user feedback.
	Usability Testing	Diverse user expressions create input ambiguity; dialog systems must adapt for a smooth experience.	Over-scripted usability tests may not mirror real-world user interactions, masking spontaneous usage patterns.	Develop realistic scenarios for usability testing, including feedback mechanisms for ongoing user insights.
Governance	*Ethical Guidelines*	The lack of clear AI governance ethics hinders policy adoption.	Unethical chatbots can produce harmful and offensive responses to users.	Must form multi-disciplinary collaboration on the ethics for HCDS design.
	Data Transparency	Companies do not always comply and follow mandates truthfully.	Can cause poor personalization, data misuse, and reduced trust in HCDS.	External audits must be performed to provide an unbiased perspective.
	Privacy Protection	Leaking sensitive information on users in the training dataset.	Violation of data protection laws and data breaches.	Ensure HCDS compliance with data protection laws and secure user data.

if users choose not to opt in for data sharing. This entails creating a system that delivers a personalized experience without relying on sensitive information, ensuring that the system remains functional and user-friendly while adhering to the highest standards of data protection and user trust.

C. Governance

While studies have exemplified governance techniques that ensure safe and ethical practices of dialog system design, there is a lack of uniform guidelines and standard practices that can be adopted by practitioners and organizations. The adoption of such standards towards governance in organizations is scarce [41], and the adoption/establishment of safe and ethical design practices among entities are siloed and lack uniformity. Furthermore, bias in governance can arise when algorithmic and data audits are performed internally, which can lead to a loss of transparency and truthfulness in dialog system design if organizations are not enacting ethical standards.

To avoid these open issues, entities and researchers from different scientific fields must collaborate in establishing principles that can be adopted by practitioners to ensure the ethical design and development of dialog systems for real-world use case scenarios. In addition, external audits must be performed to provide an unbiased perspective on the development and design practices done by organizations. As a result, ethical guidelines can be put in place to help eliminate biases in data management (e.g., collection, processing) and model training.

IV. Design Prototype and its Applicability to Real-World Applications

This section employs the identified open issues and design requirements to develop our holistic multi-layered system design. This architecture incorporates the methods and concepts of HCDS from our proposed taxonomy, as depicted in Fig. 2. We then demonstrate the practicality and effectiveness of this design applied on a HCDS integrated for a real-world application.

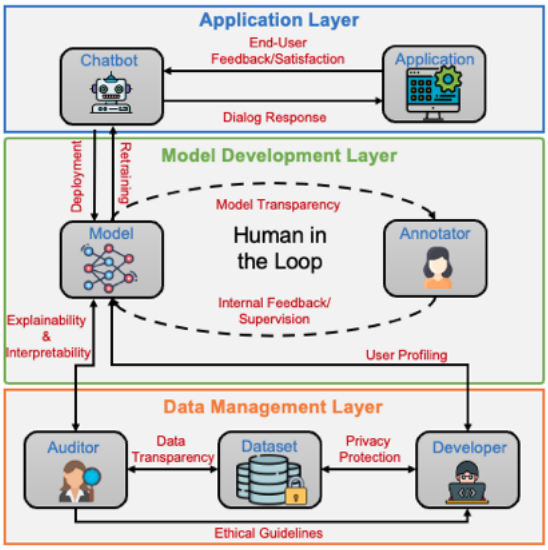

Fig. 2: Holistic multi-layered human-centered dialog system architecture for human-centered dialog system architectures.

A. Multi-layered Design for Human-Centered Dialog Systems

Data Management Layer: The initial phase lays the groundwork for the entire system. Here, data is meticulously collected, adhering to strict ethical guidelines. An auditor's role is pivotal, ensuring compliance with these guidelines and the safeguarding of data privacy. Simultaneously, developers are tasked with enforcing stringent privacy protection measures within the dataset to avoid data mismanagement and leakage. This phase also encompasses meticulous cleaning of the data to prepare it for the next stages, thus setting a high-quality baseline for the model development.

Model Development Layer: Building on clean and ethically sourced data, the second phase involves the development and rigorous testing of the model. A distinctive feature of this stage is the integration of a human-in-the-loop technique,

where internal users, annotators, or domain experts actively engage with the model to conduct interactions, testing responses, and furnishing invaluable feedback. This iterative feedback loop proves indispensable in optimizing the model's performance, efficiency and alignment with humans. Moreover, this phase also places an emphasis on model transparency, ensuring that the AI model's decision-making processes are accessible and comprehensible to stakeholders, thereby fostering trust and reliability in the system for deployment.

Application Layer: In the final phase of deployment, the focus shifts to user experience and evaluation. It involves the end-users who interact with the system, providing feedback based on their experience and satisfaction. This user feedback is crucial for identifying areas of improvement and for further refinement of the model. The deployment stage is not only a culmination but also a continuous process where user experience and feedback become integral for iterative development, ensuring model efficiency, user-friendly, and aligned with the evolving needs and expectations of its users.

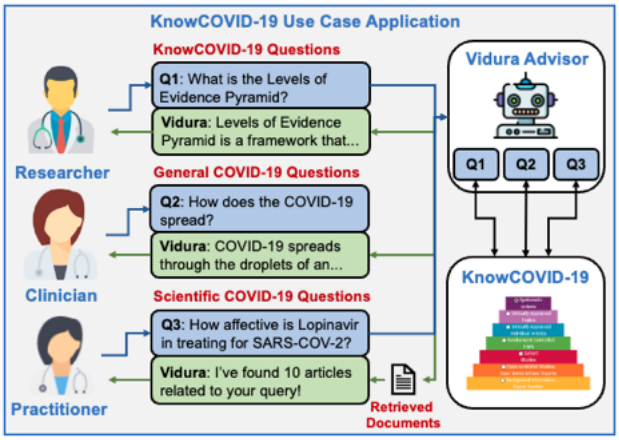

Fig. 3: Integration of the holistic muli-layered system design on the KnowCOVID-19 application for knowedge discovery.

B. Prototype Implementation in a Real-World Application

We detail our proposed holistic multi-layered design framework applied on a HCDS viz., Vidura Advisor [4], a question-answering chatbot that guides domain-specific users for knowledge discovery. In Fig. 3, we demonstrate the integration of the Vidura chatbot on an application viz., KnowCOVID-19 [3] that provides contextual guidance for medical users (e.g., researchers, clinicians, practitioners) on finding high quality literature evidence from clinical queries through a hierarchical framework called the Level of Evidence Pyramid [42]. In the following, we exemplify the main components of our systems design approach via Vidura integrated on KnowCOVID-19.

Data Management and Transparency: Prior to the development and integration of Vidura Advisor on KnowCOVID-19, we gathered a set of requirements from a group medical researchers for improving medical literature search. These include the following capabilities a HCDS must have: workflow automation, discerning user intentions for knowledge discovery, handling complex research intents, and displaying human-like responses. In the Data Management Layer, we ensured that their information throughout this study was kept confidential and used solely for the development of Vidura.

In the scope of medical literature, ensuring proper data management techniques is crucial for increasing the performance and transparency of your HCDS. In this context, we collect, and clean roughly 10,000 papers from the COVID-19 Open Research Dataset (CORD-19) [43]. CORD-19 is a publicly available dataset with over 1,000,000 scholarly articles surrounding the COVID-19 pandemic from notable medical journals as well as pre-print servers. In addition, we also gathered a set of drug and gene terms relevant to the common COVID-19 questions on treatment, diagnosis, and prevention. Our system design framework ensures that data usage complies with legal and ethical standards, particularly given the sensitive nature of medical information.

User Profiling and Model Development: The chatbot model is developed with a focus on ensuring that it understands and responds to the specific needs of different users, such as clinicians, medical researchers, or even novice users from different fields. Motivated by the works of [4], our multi-layer system design profiles users by gathering their background information (e.g. position title, research focus, and areas of interest) from the KnowCOVID-19 website and enable the Vidura chatbot to tailor its responses based on a given user's proficiency and interests.

Furthermore, our model development layer trains Vidura using natural language understanding techniques with a focus of discerning between different types of medical queries from users, namely general information on the KnowCOVID-19 application, Levels of Evidence, the COVID-19 virus, or scientific queries that returns medical literature archives. In addition, we train the chatbot to comprehend and respond to small-talk conversations to improve its likeness when conversing with users. A human-in-the-loop annotator is incorporated in the model development to provide internal feedback on dialog conversations on knowledge discovery tasks and small talk. This helps the model to learn human decision-making processes and understanding, thereby increasing its usability and effectiveness in real-world scenarios.

Usability Testing and Protocols: In the Application Layer, our holistic multi-layered system design focuses on user experience and evaluation, specifically incorporating feedback from end-users to continually refine the model. Hence, we leverage quantitative measurements based on user performance and user perception to test its utility on KnowCOVID-19. We adopt the Single Usability Metric (SUM) [44] score on user performance. SUM aims to convert the raw performance metrics of user input and standardize them in a single percentage score that measures their performance on each task as well as the overall application. Specifically, our metrics for each task based on previous work [3] include: *user success rate*, *completion time*, *difficulty score*, and the *number of user actions* performed on the application. In terms of user perception (e.g., user experience and satisfaction), we adopt

the USE Questionnaire [45] that measures their perception after the usability study. We utilize a 5-point Likert scale along the following metrics: *usefulness*, *ease of use*, *ease of learning*, and *satisfaction*. This stage also emphasizes usability protocols privacy protection for both users and the Vidura chatbot, ensuring secure and confidential information handling.

V. Human-Centered Design Performance Evaluation

We present a usability experiment that tests our proposed holistic multi-layered system design for human-centered dialogs. Specifically, we integrate the Vidura chatbot on the KnowCOVID-19 application and perform a usability study with 20 participants over clinical research tasks to capture their performance and experience. In the following, we detail our experimental setup and provide insights on the effectiveness of our design approach through the results and discussion.

A. Experimental Setup

We set up our experiment by recruiting 20 human subjects to participate in the KnowCOVID-19 usability study. The participants were recruited from various fields such as the medical domain, computer science, as well as media and communications. To demonstrate the comparison between the KnowCOVID-19 with and without our dialog-based system design, we assign 10 participants to use the application *without* the assistance of Vidura and the remaining 10 participants on KnowCOVID-19 *with* the assistance of the chatbot.

We used Zoom to conduct and record all experiments, as well as to record all experimental data for a post-hoc analysis. We ensured the protection of user privacy by asking participants for their consent to record over Zoom and keeping any of their personal information confidential. Prior to the experiment, we briefly introduced each participant to the KnowCOVID-19 application and clinical methodologies such as the Levels of Evidence that are helpful for performing evidence-based literature search. Each participant was given three literature research tasks based on COVID-19 clinical questions. Each task varies by difficulty, namely easy, medium, and hard, and participants were given a time limit of 3 minutes, 5 minutes, and 7 minutes, respectively. The three tasks are the following:

- **Easy Level Task**: *Find one article that is labeled as a case-control study.*
- **Medium Level Task**: *Find the most recent information regarding COVID-19 vaccine booster where lower level of evidence was accepted.*
- **Hard Level Task**: *Find high level of evidence about medicinal treatment of COVID-19 excluding vaccines.*

We use the aforementioned usability metrics, SUM and USE, to calculate user performance and user perception, respectively, and compare our proposed multi-layered system design for the Vidura chatbot on KnowCOVID-19 versus the standalone KnowCOVID-19 application.

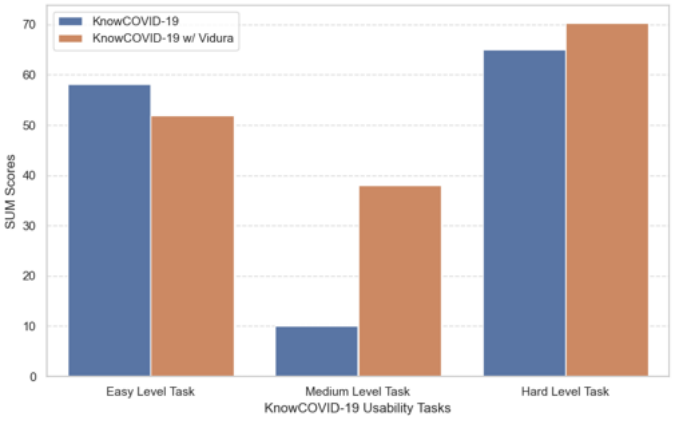

Fig. 4: The average SUM scores on user performance across all participants for each task on the KnowCOVID-19 application.

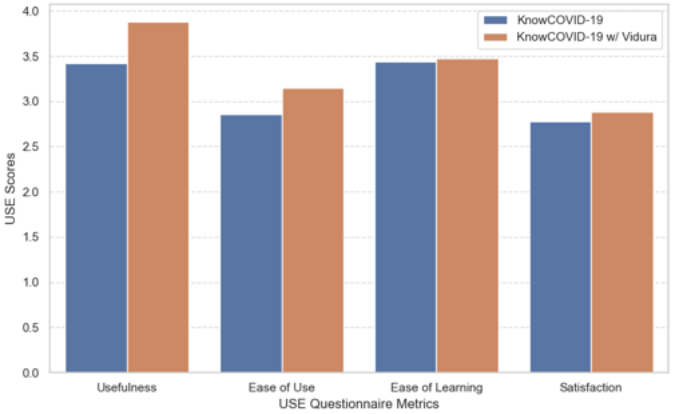

Fig. 5: The average USE scores across all participants for each metric in the questionnaire based on a 5-point Likert scale.

B. Results

The results on user performance are shown in Fig. 4. The KnowCOVID-19 application assisted with Vidura showed a higher SUM score on the whole application (53.34%) than KnowCOVID-19 only (44.42%). Participants who were assigned on KnowCOVID-19 integrated with our multi-system design for Vidura outperformed the participants using the standalone application on the medium level and hard level tasks by $\approx$ 28 percentage points and $\approx$ 5 percentage points, respectively. Participants using only KnowCOVID-19 showed a higher SUM score on the easy level tasks (58.13%) compared to KnowCOVID-19 assisted with Vidura (51.96%). Based on the recordings of the precipitants, we conclude that using the Vidura chatbot on the easy level task may have slowed down participants' time completion, thus bringing their overall SUM score by $\approx$ 6 percentage points. On the other hand, as the tasks had progressively gotten difficult, participants only using KnowCOVID-19 performed significantly worse compared to the subjects using the Vidura chatbot.

Fig. 5 shows the results on user perception using both platforms. Across all USE metrics, participants using Vidura

on KnowCOVID-19 consistently showed higher average scores compared to the subjects only using KnowCOVID-19. A significant number of participants on the medium level and hard level tasks did not finish in time, which brought down the average score of the USE metrics, particularly ease of use and satisfaction. While not all participants came from a medical background, we observe that our human-centered dialog, Vidura, enhanced their experience in performing effective literature tasks on the KnowCOVID-19 application. Ultimately, we show how our multi-layered design for the Vidura chatbot improves the overall utility of knowledge-intensive applications such as KnowCOVID-19.

C. Discussion

Herein, we aim to provide a discussion that describes how our KnowCOVID-19 usability study demonstrates the effectiveness of holistic multi-system design. As previously mentioned, the Vidura chatbot aims to guide researchers/practitioners in various knowledge intensive tasks. A reason for why Vidura is effective on guiding users on KnowCOVID-19 is largely due its ability to provide generalized knowledge while also tailor services to users given their domain proficiency. This enables our approach, specifically in the Model Development Layer, to balance between providing necessary and common knowledge among users as well as personalized responses. In terms of the Application Layer, we have deployed our NLU component within the Model Development Layer on the application to recognize friendly and emotional user utterances and provide responses that use human-like tone and empathize with the user.

Extensions of our work can involve studying crucial examples to improve the utility on various components within our multi-layer system design through evaluation techniques. Within the Data Management Layer, it is essential to test the robustness of the chatbot by detecting data biases to prevent biases in generated responses, as mentioned in previous works such as [46]. From the perspective of the Model Development Layer, establishing techniques or frameworks such as in [32], [33] are crucial in enabling the dialog to be more transparent in its predictive responses to mitigate users from blindly following the model's outcomes. In the Application Layer, there presents a need to provide a human-centric evaluation on generated responses that are also scalable in practice. For example, authors in [47] show that using automated response selection evaluation techniques on dialog generation better correlates with human evaluation. Investigating these aspects within our multi-system design can ultimately help foster a robust and dependable HCDS in real-world applications.

VI. Conclusion

To conclude, we propose a holistic multi-layered system design that considers human-centered approaches in every facet of the development lifecycle. We identified a taxonomy within the HCDS design and surveyed common methods/concepts among each category. We then identified open issues and potential risks within taxonomy and detailed design requirements. Based on these requirements, we proposed our end-to-end HCDS design in a real-world application for knowledge discovery and performed a usability study that demonstrated the utility of the application through improved user performance and experience.

Future directions in the scope of the design of dialog systems with a human-centered approach include safety evaluation of datasets and model prompts [48], algorithmic auditing tools to eliminate the misuse of unauthorized user data [49], and standard guidelines for ethical and safe development practices [41].

References

[1] N. Bhirud, S. Tataale, S. Randive, and S. Nahar, "A literature review on chatbots in healthcare domain," *International journal of scientific & technology research*, vol. 8, no. 7, pp. 225–231, 2019.

[2] G.-J. Hwang and C.-Y. Chang, "A review of opportunities and challenges of chatbots in education," *Interactive Learning Environments*, pp. 1–14, 2021.

[3] R. Oruche, E. D. Milman, X. Cheng, M. Joish, C. Kulkarni, A. Sharma, K. Kee, H. Regunath, and P. Calyam, "Measurement of utility in user access of covid-19 literature via ai-powered chatbot," in *2021 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. IEEE, 2021, pp. 1–13.

[4] A. A. Chandrashekara, R. K. M. Talluri, S. S. Sivarathri, R. Mitra, P. Calyam, K. Kee, and S. Nair, "Fuzzy-based conversational recommender for data-intensive science gateway applications," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 4870–4875.

[5] S. Kottur, X. Wang, and V. Carvalho, "Exploring personalized neural conversational models." in *IJCAI*, 2017, pp. 3728–3734.

[6] D. Yang and L. Flek, "Towards user-centric text-to-text generation: A survey," in *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings 24*. Springer, 2021, pp. 3–22.

[7] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?" in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2204–2213.

[8] T. Scialom, S. S. Tekiroğlu, J. Staiano, and M. Guerini, "Toward stance-based personas for opinionated dialogues," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 2625–2635.

[9] A. Gupta, P. Zhang, G. Lalwani, and M. Diab, "CASA-NLU: Context-aware self-attentive natural language understanding for task-oriented chatbots," 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)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1285–1290.

[10] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the ai: informing design practices for explainable ai user experiences," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–15.

[11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[12] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *The Eleventh International Conference on Learning Representations*, 2023.

[13] E. Ricciardelli and D. Biswas, "Self-improving chatbots based on reinforcement learning," in *4th Multidisciplinary Conference on Reinforcement Learning and Decision Making*, 2019.

[14] B. Liu, G. Tur, D. Hakkani-Tur, P. Shah, and L. Heck, "Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems," *arXiv preprint arXiv:1804.06512*, 2018.

[15] J. Kreutzer, S. Khadivi, E. Matusov, and S. Riezler, "Can neural machine translation be improved with user feedback?" in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. New Orleans - Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 92–105.

[16] P. Yu, H. Xu, X. Hu, and C. Deng, "Leveraging generative ai and large language models: A comprehensive roadmap for healthcare integration," in *Healthcare*, vol. 11, no. 20. MDPI, 2023, p. 2776.

[17] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak, "Survey on evaluation methods for dialogue systems," *Artificial Intelligence Review*, vol. 54, pp. 755–810, 2021.

[18] C. Siro, M. Aliannejadi, and M. de Rijke, "Understanding user satisfaction with task-oriented dialogue systems," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 2018–2023.

[19] A. Følstad and C. Taylor, "Investigating the user experience of customer service chatbot interaction: a framework for qualitative analysis of chatbot dialogues," *Quality and User Experience*, vol. 6, no. 1, p. 6, 2021.

[20] Y. Park, S. Kang, and J. Seo, "An efficient framework for development of task-oriented dialog systems in a smart home environment," *Sensors*, vol. 18, no. 5, p. 1581, 2018.

[21] X. Li, W. Wu, L. Qin, and Q. Yin, "How to evaluate your dialogue models: A review of approaches," *unpublished*, 2023.

[22] J. Sedoc, D. Ippolito, A. Kirubarajan, J. Thirani, L. Ungar, and C. Callison-Burch, "Chateval: A tool for chatbot evaluation," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, 2019, pp. 60–65.

[23] E. Durmus, H. He, and M. Diab, "FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 5055–5070.

[24] A. Malchanau, V. Petukhova, and H. Bunt, "Multimodal dialogue system evaluation: a case study applying usability standards," in *9th International Workshop on Spoken Dialogue System Technology*. Springer, 2019, pp. 145–159.

[25] L. Vanderlyn, G. Weber, M. Neumann, D. Väth, S. Meyer, and N. T. Vu, ""it seemed like an annoying woman": On the perception and ethical considerations of affective language in text-based conversational agents," in *Proceedings of the 25th Conference on Computational Natural Language Learning*, A. Bisazza and O. Abend, Eds. Online: Association for Computational Linguistics, Nov. 2021, pp. 44–57.

[26] J. Bang, S. Kim, J. W. Nam, and D.-G. Yang, "Ethical chatbot design for reducing negative effects of biased data and unethical conversations," in *2021 International Conference on Platform Technology and Service (PlatCon)*. IEEE, 2021, pp. 1–5.

[27] P. Henderson, K. Sinha, N. Angelard-Gontier, N. R. Ke, G. Fried, R. Lowe, and J. Pineau, "Ethical challenges in data-driven dialogue systems," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 123–129.

[28] T. Wambsganss, A. Höch, N. Zierau, and M. Söllner, "Ethical design of conversational agents: towards principles for a value-sensitive design," in *Innovation Through Information Systems: Volume I: A Collection of Latest Research on Domain Issues*. Springer, 2021, pp. 539–557.

[29] J. Zhou, J. Deng, F. Mi, Y. Li, Y. Wang, M. Huang, X. Jiang, Q. Liu, and H. Meng, "Towards identifying social bias in dialog systems: Framework, dataset, and benchmark," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3576–3591.

[30] C. Ziems, J. Yu, Y.-C. Wang, A. Halevy, and D. Yang, "The moral integrity corpus: A benchmark for ethical dialogue systems," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3755–3773.

[31] H. Sun, Z. Zhang, F. Mi, Y. Wang, W. Liu, J. Cui, B. Wang, Q. Liu, and M. Huang, "Moraldial: A framework to train and evaluate moral dialogue systems via moral discussions," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 2213–2230.

[32] S. Hepenstal, N. Kodagoda, L. Zhang, P. Paudyal, and B. Wong, "Algorithmic transparency of conversational agents," in *IUI 2019 Workshop on Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies*. 85y0v, 2019.

[33] A. Khurana, P. Alamzadeh, and P. K. Chilana, "Chatrex: Designing explainable chatbot interfaces for enhancing usefulness, transparency, and trust," in *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 2021, pp. 1–11.

[34] T. Wu, M. Terry, and C. J. Cai, "Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts," in *Proceedings of the 2022 CHI conference on human factors in computing systems*, 2022, pp. 1–22.

[35] M. Larson, N. Oostdijk, and F. Zuiderveen Borgesius, "Not directly stated, not explicitly stored: Conversational agents and the privacy threat of implicit information," in *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 2021, pp. 388–391.

[36] M. Hasal, J. Nowaková, K. Ahmed Saghair, H. Abdulla, V. Snášel, and L. Ogiela, "Chatbots: Security, privacy, data protection, and social aspects," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 19, p. e6426, 2021.

[37] Q. Xu, L. Qu, Z. Gao, and G. Haffari, "Personal information leakage detection in conversations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 6567–6580.

[38] S. Terragni, B. Guedes, A. Manso, M. Filipavicius, N. Khau, and R. Mathis, "BETOLD: A task-oriented dialog dataset for breakdown detection," in *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, X. Wu, P. Ruan, S. Li, and Y. Dong, Eds. Gyeongju, Republic of Korea: Association for Computational Linguistics, Oct. 2022, pp. 23–34.

[39] B. Fazzinga, A. Galassi, and P. Torroni, "A privacy-preserving dialogue system based on argumentation," *Intelligent Systems with Applications*, vol. 16, p. 200113, 2022.

[40] S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire *et al.*, "Open problems and fundamental limitations of reinforcement learning from human feedback," *unpublished*, 2023.

[41] T. Lewandowski, J. Delling, C. Grotherr, and T. Böhmann, "State-of-the-art analysis of adopting ai-based conversational agents in organizations: A systematic literature review." *PACIS*, p. 167, 2021.

[42] D. Timm, "Evidence Matters," *Journal of the Medical Library Association: JMLA*, vol. 94, no. 4, p. 480, 2006.

[43] M. Ekin Eren, N. Solovyev, E. Raff, C. Nicholas, and B. Johnson, "COVID-19 Kaggle Literature Organization," *arXiv e-prints*, pp. arXiv–2008, 2020.

[44] J. Sauro and E. Kindlund, "A method to standardize usability metrics into a single score," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2005, pp. 401–409.

[45] A. M. Lund, "Measuring usability with the use questionnaire12," *Usability interface*, vol. 8, no. 2, pp. 3–6, 2001.

[46] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving, "Red teaming language models with language models," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 3419–3448.

[47] S. Sato, R. Akama, H. Ouchi, J. Suzuki, and K. Inui, "Evaluating dialogue generation systems via response selection," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 593–599. [Online]. Available: https://aclanthology.org/2020.acl-main.55

[48] J. Deng, H. Sun, Z. Zhang, J. Cheng, and M. Huang, "Recent advances towards safe, responsible, and moral dialogue systems: A survey," *unpublished*, 2023.

[49] C. Song and V. Shmatikov, "Auditing data provenance in text-generation models," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 196–206.