Chatbot Dialog Design for Improved Human Performance in Domain Knowledge Discovery

Roland Oruche, *Member, IEEE*, Xiyao Cheng, Zian Zeng, Audrey Vazzana, MD Ashraful Goni, Bruce Wang Shibo, Sai Keerthana Goruganthu, Kerk Kee, and Prasad Calyam

***Abstract*—The advent of machine learning (ML) has led to the widespread adoption of developing task-oriented dialog systems for scientific applications (e.g., science gateways) where voluminous information sources are retrieved and curated for domain users. Yet, there still exists a challenge in designing chatbot dialog systems that achieve widespread diffusion among scientific communities. In this article, we propose a novel Vidura advisor design framework (VADF) to develop dialog system designs for information retrieval (IR) and question-answering (QA) tasks, while enabling the quantification of system utility based on human performance in diverse application environments. We adopt a socio-technical approach in our framework for designing dialog systems by utilizing domain expert feedback, which features a sparse retriever for enabling accurate responses in QA settings using linear interpolation smoothing. We apply our VADF for an exemplar science gateway, viz. KnowCOVID-19, to conduct experiments that demonstrate the utility of dialog systems based on IR and QA performance, application utility, and perceived adoption. Experimental results show our VADF approach significantly improves IR performance against retriever baselines (up to 5% increase) and QA performance against large language models (LLMs) such as ChatGPT (up to 43% increase) on scientific literature datasets. In addition, through a usability survey, we observe that measuring application utility and human performance when applying VADF to KnowCOVID-19 translates to an increase in perceived community adoption.**

***Index Terms*—Dialog design, diffusion of innovations, information retrieval (IR), publication analytics, utility measurement.**

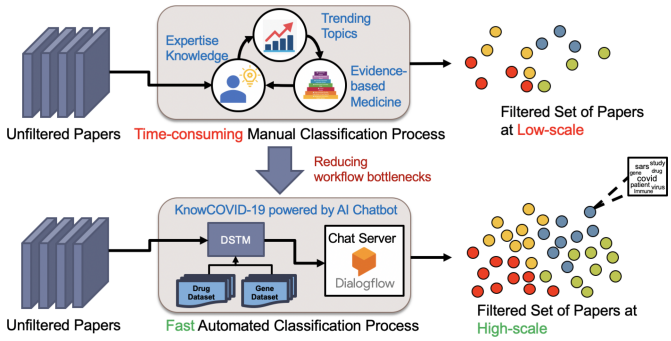


Fig. 1. Improvement of the manual literature search process using a chatbot dialog design for knowledge discovery of COVID-19 publications at high scale.

I. Introduction

THE adoption of chatbot dialog systems that are aided by machine learning (ML) has the potential to be effective for use in various domain-specific applications accessible via science gateways [1]. Science gateways provide easy access to voluminous data sources and computational tools, however they need advanced capabilities for domain users to be able to perform knowledge discovery tasks such as extracting critical insights from data archives or performing scientific simulations. Consequently, domain users seek to find appropriate and supportive technologies, such as chatbot dialog systems, for *prescriptive guidance* that overcome arduous manual steps, and also enable effective knowledge discovery for a given scientific problem.

Fig. 1 showcases an example scenario in which chatbot dialog systems can help alleviate the burden in using manual processes by domain users who analyze literature archives in the context of COVID-19. In this scenario, medical users (e.g., clinicians, researchers) typically face many limitations when relying on expert knowledge and clinical frameworks to manually perform knowledge discovery on papers at high scale. Chatbot dialog systems using ML techniques for automatically filtering and retrieving papers at high scale can provide more rapid literature search outcomes for users, which can yield high adoption/diffusion [2] among scientific communities.

Several works investigate chatbot dialog designs for knowledge discovery tasks within various scientific communities. Sivarathri et al. [3] detailed how context-aware recommender chatbots for science gateways can guide domain-specific users into finding actionable insights on new knowledge. Recent

Received 7 March 2024; revised 1 July 2024, 31 October 2024, and 26 November 2024; accepted 2 December 2024. This work was supported by the National Science Foundation under Award OAC-2006816 and Award OAC-2007100. This article was recommended by Associate Editor Mei-Ling Shyu. *(Corresponding author: Prasad Calyam.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Texas Tech University - Institutional Review Board under Application No. IRB2024-333, and performed in line with the Texas Tech University's Operating Procedures, the Belmont Report, and 45 CFR 46.

Roland Oruche, Xiyao Cheng, Sai Keerthana Goruganthu, and Prasad Calyam are with the Department of Electrical Engineering and Computer Science, University of Missouri-Columbia, Columbia, MO 65211 USA (e-mail: calyamp@missouri.edu).

Zian Zeng is with the Department of Information and Computer Sciences, University of Hawaii, Honolulu, HI 96822 USA.

Audrey Vazzana is with the Department of Computer Science, Truman State University, Kirksville, MO 63501 USA.

MD Ashraful Goni, Bruce Wang Shibo, and Kerk Kee are with the College of Media & Communication, Texas Tech University, Lubbock, TX 79409 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2024.3514742.

Digital Object Identifier 10.1109/THMS.2024.3514742

work leverages pretrained language models such as ChatGPT [4] technology to support medical domain users for knowledge discovery in health care applications such as diagnostic decision-making [5] and scientific document summarization [6]. Despite this, most existing work does not *jointly* consider the socio-technical considerations in chatbot dialog designs for capturing its utility in application settings and achieving widespread adoption/diffusion in scientific communities.

In this study, we develop the *Vidura advisor design framework (VADF)* that enables the design of ML-based dialog systems for knowledge discovery tasks such as information retrieval (IR) and question-answering (QA), as well as the quantification of system utility based on human performance in diverse application environments. We propose a socio-technical approach for design of chatbot-based dialog systems in the scope of diffusion of innovations (DOI) theory [2] by utilizing domain expert feedback to formalize a set of chatbot dialog design requirements for improving response outcomes on downstream IR and QA tasks. The VADF features a natural language understanding (NLU) component that discerns between purposeful query intents and a IR module (i.e., *ViduraQA retriever*) based on a linear interpolation smoothing technique [7] that augments a chatbot language model for providing precise and accurate responses to complex scientific queries. VADF also features a utility measurement methodology based on a set of usability metrics and computes a score of chatbot dialog design merits when integrated on a science gateway using the single usability metric (SUM) [8]. We apply our VADF for an exemplar science gateway viz., KnowCOVID-19 [9], which helps users to perform publication analytics. Through the KnowCOVID-19 application, we detail how VADF can be integrated into scientific applications with CI/CD (continuous improvement/continuous delivery) processes for supporting domain users' knowledge discovery needs as new curated domain data emerges and evolves.

We conduct experiments using the KnowCOVID-19 application and 10 000 medical literature documents from the COVID-19 open research dataset (CORD-19) [10] to demonstrate the utility of our VADF approach based on IR and QA performance, application utility, and perceived adoption. In the IR experiment, we utilize the TREC-COVID [11] and NFCorpus [12] benchmark datasets to evaluate the performance of our ViduraQA retriever against BM25 [13], and when applied to fine-tuned neural models in zero-shot settings [14], [15], [16], [17], [18], [19], [20], [21]. In the QA experiment, we test the performance of our VADF approach against pretrained large language models (LLMs) such as ChatGPT using OpenAI's GPT-3.5 Turbo [22] over the COVID-QA dataset [23]. Following this, we conduct a usability study with 20 (diverse) participants to perform COVID-19 literature search tasks of varying difficulties (i.e., easy, medium, hard) on KnowCOVID-19. Specifically, we quantify human and application performance by using the SUM and System Usability Scale (SUS) [24], respectively, and use these frameworks to compare the performances of KnowCOVID-19 assisted with the VADF against KnowCOVID-19 without an implemented chatbot dialog design. In the final experiment, we conducted a set of application utility measures in the form of survey questionnaires [25], [26], [27] to gather the participants' perceptions for adopting the VADF on KnowCOVID-19.

The contributions summary of this article is as follows.

1) We investigate a socio-technical approach in chatbot dialog design by gathering domain expert feedback to develop the VADF that entails the ViduraQA retriever for downstream IR and QA tasks related to knowledge discovery.
2) We adopt a utility measurement methodology to quantify VADF's utility on scientific applications and detail the integration of VADF on applications (e.g., KnowCOVID-19) with CI/CD and software capabilities to support users for knowledge discovery.
3) We present a set of experiments that evaluate the IR and QA performance of our VADF approach, the application utility based on human performance, and perceived adoption of VADF integrated on KnowCOVID-19. The results provide insights on how VADF, when integrated on KnowCOVID-19, can impact scientific communities in terms of increased performance and perceived adoption.

The rest of this article is organized as follows. Section II discusses the related work. In Section III, we detail the dialog design requirements based on domain expert feedback. In Section IV, we introduce the VADF and detail the ML-based components for knowledge discovery tasks. In Section V, we detail the utility measurement methodology that quantifies VADF's utility integrated in a scientific application. Section VI details the integration of VADF on an exemplar scientific application viz., KnowCOVID-19. In Section VII, we detail the experimental setup of our ViduraQA and VADF approach, and in Section VIII we show the experimental results on the IR and QA performance, utility, and perception of KnowCOVID-19 with VADF against applications without a chatbot dialog design. In Section IX, we provide a discussion on the obtained experiment results and challenges of VADF to scientific applications. Finally, Section X concludes this article.

II. Related Work

A. Chatbot Question-Answering for Information Retrieval

Question-answering (QA) systems have shown the ability to effectively respond to users' questions using ML and natural language processing (NLP) techniques for tasks such as IR [28]. With the increasing popularity of transfer learning and LLMs, several works on neural IR [29] have contributed to the advancement of this field. The work in [14] introduces two variants of BERT [30], namely, monoBERT and duoBERT, which converts the ranking problem into pointwise and pairwise classification problems under a multistage ranking architecture, respectively. ColBERTV2 [31] uses a pretraining and fine-tuning strategy to create an efficient and effective retrieval system that improves the quality and space efficiency of multivector representations for large-scale document ranking tasks. Nogueira et al. [18] showcased the benefits of leveraging the bidirectional nature of text-to-text transfer transformer (T5)-based models for capturing context in document similarity and ranking tasks.

Recent work has shown the success of IR in QA chatbots. The work in [32] proposes a novel IR approach for chatbot

engines viz., DocChat, which leverages unstructured documents to measure the relevance between user utterances and chatbot responses at different levels of granularity. Lommatzsch and Katins [33] proposed a chatbot framework that adapted IR techniques to answer questions related to public administration services. Makhalova et al. [34] used formal concept analysis and pattern structures to build a domain knowledge chatbot in the domain of customer services. Maoro et al. [35] proposed to use a semantic search pipeline and leveraged conversational LLMs such as OpenAI's GPT-3.5 [22] to develop a domain-adaptive QA framework in various domains such as the industrial sector.

Despite the recent advancements, there is a knowledge gap on the socio-technical design approaches of chatbot QA systems that measure application utility for knowledge discovery tasks of scientific applications deployed in the form of science gateways. Our proposed VADF uses a socio-technical approach which features IR techniques for improving knowledge discovery over scientific literature.

B. Utility Measurement of Chatbot Question–Answering

Utility has been long studied for the purposes of disbursement and adoption by users [2], [36], [37]. Early definitions of utility refer to the ability of a system to sufficiently meet the needs of a general user [38]. Adoption frameworks in the context of utility such as the technology acceptance model (TAM) [36] describe the factors behind the adoption of new technology based on concepts such as *perceived usefulness* and *perceived ease of use*. Other frameworks such as unified theory of acceptance and use of technology (UTAUT) [37] include four factors such as performance expectancy, effort expectancy, social influence, and facilitating conditions.

In the scope of chatbot QA systems, previous works have applied measurements to assess its utility for user adoption in various applications. A mixed methods study in [39] identifies relative advantage and information systems infrastructure to be the most influential factors for the adoption of chatbot technology in the German insurance industry. Goli et al. [40] extended the TAM model by adding factors such as perceived enjoyment, perceived innovativeness, perceived information quality, and perceived customization. This work suggests that all factors, with the exception of perceived enjoyment, have a significant impact on user adoption of chatbots. Similarly, De Cicco et al. [41] extended the TAM model by adding factors of compatibility, trust, and perceived enjoyment to understand users' intentions for using e-commerce chatbots. This work finds that user intentions are highly influenced by their attitude toward the technology, which is dependent heavily on the factors of compatibility, perceived usefulness, and perceived enjoyment. Laumer et al. [42] identified additional factors to the UTAUT model such as privacy risk expectancy, trust in provider and system, compatibility, experience in e-diagnosis, and access to the health system to create a framework for the adoption of chatbot technology in health care.

We build upon prior works on the theories of utility measurement and show empirical evidence of factors that influence adoption rates according to DOI theory [2]. Specifically, our work addresses individual-level adoption challenges (e.g., relative advantage, simplicity) by developing VADF to enhance chatbot QA systems in knowledge-intensive science applications by measuring the utility in human performance and perception. Moreover, we devise a utility measurement methodology featuring several utility metrics to quantify the utility of the VADF when applied broadly to applications in diverse environments (e.g., supporting science gateways) based on human performance and perceived adoption.

III. Chatbot Dialog Design Requirements

In this section, we discuss the challenges and design requirements (as shown in Table I) for a chatbot dialog design based on user feedback from a pilot study [43] on KnowCOVID-19, an exemplar science gateway on COVID-19 publication analytics. Herein, we gather the impressions of eight medical users (i.e., clinicians, researchers) and detail the set of common issues participants faced when using the chatbot dialog on the science gateway and subsequent design requirements for our VADF that is aligned with DOI theory.

A. Dialog Personalization

Users of interactive interfaces of complex data systems have varying needs, and demand a personalization of the functions accessible to them. Due to the complexity of interactive interfaces on domain-specific applications such as science gateways, however, certain functions on the application may not be intuitive. For example, medical users in our COVID-19 publication analytics science gateway (KnowCOVID-19) pilot study, often expressed the need to have personalized capabilities with the chatbot dialog such as tailored responses based on their domain proficiency in order to perform effective literature search. This QA chatbot dialog was based on predefined responses and could not generate personalized responses that would uniquely assist them in their literature search methods, making the application difficult to use.

To avoid this issue, a chatbot dialog must provide timely and personalized responses based on interests/needs of domain users on a given application. In cases where conventional and time-intensive literature search methods create a huge workflow bottleneck in performing knowledge discovery, chatbot dialog designs must alleviate this constraint by performing an analysis over a literature corpus and presenting a set of personalized results that uniquely addresses a user's research objective or interest. In addition, previous studies have suggested that designing humanlike conversational features can improve adoption aspects such as satisfaction and other users' experience factors [44], [45]. In light of these works, such designs of text-based dialogs, such as chatbot, must include responses that closely mimic human-to-human interaction such as greetings and follow-up questions in a professional manner.

B. In-Depth Knowledge Discovery

In task-oriented applications, chatbot dialog agents are designed to aid users in completing their request. One major

TABLE I
SET OF CHATBOT DIALOG CHALLENGES AND DESIGN REQUIREMENTS OBTAINED FROM USER FEEDBACK IN A PILOT STUDY [43]

Challenges	Chatbot dialog issue	Design requirement(s)
Dialog Personalization	Complexity of domain-specific application makes it unintuitive for new users and hinders dialog personalization.	(1) Chatbot dialogs must perform an analysis based on users' specific interests/needs to augment their workflow in domain-specific applications. (2) Chatbot dialogs should increase human-likeness by providing responses to users that are polite, professional, and empathetic.
In-depth Knowledge Discovery	Insufficient information for knowledge discovery based on user intents results in the dissatisfaction of the user and low utility of the chatbot.	(1) Design of chatbots must handle complex and domain-specific research questions and provide information that guides users towards knowledge discovery. (2) Dialog designs must not be static, and therefore, should update its domain knowledge based on changes in the real world.
Wide-ranging Chatbot Responses	Inability to comprehend domain-user intentions and respond with correct terminology makes the chatbot ineffective for knowledge-intensive tasks.	(1) Dialog designs must be trained to understand users' intention by discerning between general or scientific-related questions.

challenge among chatbot dialog responses is providing sufficient information that satisfies a domain user's request. Dialog agents can fail due to the functions that over- or underprovision information, leading domain users to feel dissatisfied and discouraged. This issue was manifested among medical users in the KnowCOVID-19 pilot study, when the chatbot dialog had insufficient search capabilities that hindered the participants' ability for performing in-depth knowledge discovery tasks in the form of IR. This functionality in turn did not motivate the medical users to seek out asking the chatbot on specific clinical tasks that involved literature research for in-depth knowledge discovery.

To alleviate this issue, chatbot dialog designs must handle the complexity of domain-specific questions by leveraging ML techniques while training over a large literature corpus with a database of scientific terms—particularly those that are commonly used in the target domain—to generate relevant responses. Furthermore, since complex research queries are generally considered to be open-ended, it is important to consider generating a candidate set of information in the form of articles that converge around the same topic. Besides collecting large amounts of literature and terms for the purpose of generating relevant responses for the users, staying up-to-date with the latest information is crucial to the development of chatbot dialog systems. Thus, it is also important to design a pipeline which integrates a chatbot that continuously collects and re-trains over new and obscure scientific information from disparate sources. A continuous integration and delivery system must be developed and deployed when harnessing the latest batch of text information.

C. Wide-Ranging Chatbot Responses

Due to the high-variability of terminology and objectives among diverse participants, there emerges a design challenge in the chatbot dialog to respond to a wide-range of query intentions. These challenges are critical when user queries are too complex due to domain vocabulary. As an example, the chatbot dialog agent in our KnowCOVID-19 pilot study, failed to respond to complex clinical questions that gave medical users further insight to improve their clinical workflow. As previously mentioned, the chatbot was trained on general questions about the science gateway and instructions for how to use its application functions. Due to the insufficient chatbot design to handle domain-specific query topics with high variability, it was not possible to provide a satisfactory response to clinically related questions.

As researchers are continuously exploring new ideas related to emerging issues in scientific areas such as COVID-19 pandemic response, their intentions for which type of information they seek should be understood. While chatbot dialog agents are designed to provide guidance for solving knowledge discovery tasks, it is crucial to understand domain users' purposeful intentions to handle their particular requests. Specifically, the chatbot must be effective in deciding whether to inform users on general or widely studied knowledge in the field, or to provide suggestions on new knowledge through obscure studies and unverified knowledge. Such a pipeline should handle a set of questions from the user and discern whether it is informative or scientific.

IV. METHOD FOR QUESTION-ANSWERING FOR INFORMATION RETRIEVAL

In this section, we take the expert feedback from our pilot study to develop the *VADF* that improves human performance on knowledge discovery tasks over data archives. Fig. 2 shows the end-to-end QA process between the domain user and chatbot dialog for pertinent response outputs related to scientific IR and QA. In the following, we detail the components of the VADF including the query intent classifier, ViduraQA retriever, and language model for text generation.

A. Query Intent Classifier

The query intent classifier shown in Fig. 2 is a NLU module that discerns between user intents (i.e., queries) using a classification scheme. The goal of the NLU module within our VADF design process is to map the type of question a user is asking to enable the chatbot to provide more pertinent responses.

We build a dataset to allow our NLU to handle a diverse set of query intents in domain specific application settings. Specifically, we build our dataset around five major intent classes: *informative*, *scientific*, *application-specific*, *application-resource*, as well as *chit-chat* (e.g., greetings). An informative intent is a query that seeks known factual information about emerging topics such as the COVID-19 pandemic response. An example of an informative COVID-19 related question could include – *What are the origin and reason for the name of coronavirus disease 2019?* A scientific intent is a question type that yields

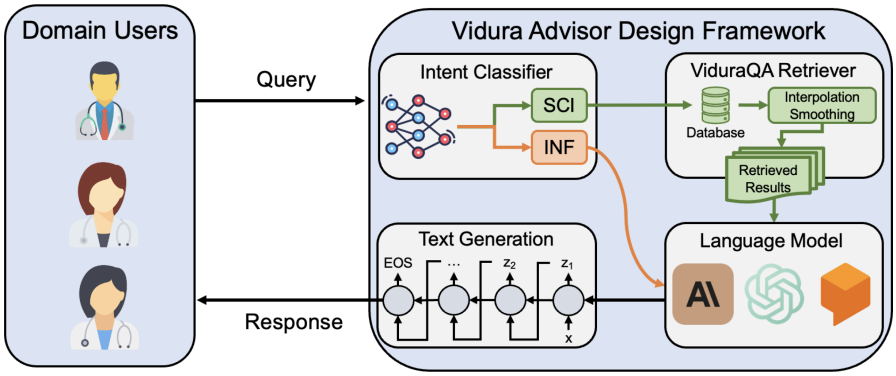


Fig. 2. VADF that includes an intent classifier to discern between scientific (SCI) or informative (INF) queries. A response is generated with candidate publication archives and/or text generated from the language model.

on-going research findings from one or multiple publications on that topic. An example of a scientific COVID-19 query may include—*What is the efficacy of lopinavir in COVID-19 treatment?* Application-specific and resource intents relate to questions about the application such as website navigation and tools/methodologies to assist domain users, respectively.

We construct our multiclass dataset by collecting a set of informative/scientific questions from the COVID question–answering (COVID-QA) dataset [23] and a set of chit–chat questions from the ChatterBot language training corpus [46]. A total of 867 questions were collected from both datasets. The dataset was split 80/20 in terms of 80% of samples (650 sample questions) in the training set and 20% of samples (217 sample questions) in the test set. We evaluated our NLU model by leveraging a pretrained BERT model [30] for intent classification. The BERT model was trained over ten epochs using the AdamW optimizer and learning rate of $5\mathrm{e}-5$. The results report a 92% accuracy (with a standard deviation of ± 0.01) averaged over five runs with different seeds.

B. Question-Answering for Information Retrieval

We propose a technique viz., ViduraQA retriever to apply linear interpolation smoothing for IR tasks. From a user perspective, scientific queries typically are represented as complex research questions on emerging topics and are prevalent in scientific literature. In the context of topics such as COVID-19, a scientific question could include, e.g., *What is the in vitro comparison of antiviral activity of Chloroquine (CQ) and Hydroxychloroquine (HCQ) against COVID-19?* In the following, we describe the process of the smoothing technique motivated by [7] for retrieving relevant articles according to users' queries.

Formally, we define $D = \{d_1, \ldots, d_n\}$ and $Q = \{q_1, \ldots, q_m\}$ as the set of collected documents and queries, respectively. Each document d in the literature corpus has a length of N_d total tokens, and can be represented as a word occurrence using the bag-of-words technique. Based on this representation, each document is constructed as a unigram language model by calculating the probability distribution of terms within a document that is independent from other terms. The aim for IR is to find the probability of generating a query $q \in Q$ given a document d, which can be denoted as $P(q|d)$. The estimation of $P(q|d)$ can be computed as follows:

$$P(q|d) = \prod_{i=0}^{N_q} P(q_i|d) = \prod_{i=0}^{N_q} \frac{tf_{q_i,d}}{N_d} \tag{1}$$

where, $tf_{q_i,d}$ is the term frequency of the ith query term in document d obtained from the bag-of-words representation, N_d is the length of the document, and N_q is the length of the query. A common challenge in (1) occurs when a given set of query terms are not present in a given document. This in turn creates a zero probability of a query success. To alleviate this issue, we apply linear interpolation smoothing, inspired by the work in [7], to calculate both the probability of a word in a given document and the corpus. Linear interpolation smoothing is a technique that mitigates the zero probability issue by building a language model for the whole corpus. Given a corpus C as a collection of documents, we calculate the probability of a query word with respect the corpus as $P(q|C)$. Thus, the calculation of $P(q|d)$ is computed as follows:

$$\begin{aligned} P(q|d) &= \prod_{i=0}^{N_q} \Big[\alpha P(q_i|d) + (1-\alpha) P(q_i|C) \Big] \\ &= \prod_{i=0}^{N_q} \left[\alpha \frac{tf_{q_i,d}}{N_d} + (1-\alpha) \frac{tf_{q_i,C}}{N_C} \right] \end{aligned} \tag{2}$$

where $tf_{q_i,C}$ is the term frequency of the ith query in the corpus, N_C is the total number of tokens in the corpus, and α is a hyperparameter term. This in turn makes the IR technique more robust when handling query terms that are not present within a given document. Hence, this ViduraQA retriever can improve the IR performance of scientific literature documents.

C. Language Modeling for Text Generation

Given the retrieved documents from our ViduraQA retriever, we describe how this evidence can be used as context to enable a language model to provide a response in the form of question–answering. Formally, we denote an input prompt to the language model as $x = (q, E)$, where q is the input query and $E = \{e_1, e_2, \ldots, e_K\}$ is the set of evidence documents of length K. The response generation function of the language model with model parameters θ, denoted as $p(y|x;\theta)$ is conditioned on

the query and evidence documents, where $y = (y_1, y_2, \ldots, y_T)$ is the generated response of length T. The probability of a response given the prompt is calculated as $p(y|x;\theta) = \prod_{t=0}^{T} p(y_t|y_{<t}, x;\theta)$, where $y_{<t} = \{y_1, y_2, \ldots, y_{t-1}\}$ are the previous generated tokens. As indicated in Fig. 2, the VADF can incorporate various language models for text generation such as ChatGPT [4] and Claude [47].

In our usability study, we integrate Google Dialogflow [48], which is a conversational intelligence service that uses NLP techniques to perform QA over user queries. It features four main attributes: *intents*, *entities*, *actions and parameters*, and *fulfillments*. Intents aim to identify the type of query provided by the user, in which variations of these questions are trained to provide one or more predefined responses. Entities create a key-value pair between a word and its corresponding value based on similarity. Actions and parameters are used to filter intents for generating a query format using important query terms. Fulfillments are dynamic and application-specific responses to user queries through API calls. Along with its conversational capabilities, Dialogflow can also integrate various pretrained language models for QA given input prompts through the "Generators" feature. Similar to the above formulation for language models such as ChatGPT and Claude, Dialogflow uses a template prompt to generate responses conditioned on evidence documents. The prompt includes placeholders in the form of word tokens for dynamically inserting queries and retrieved documents using the fulfillment feature. Upon the live execution of Dialogflow Generations, the ViduraQA retriever can use the API endpoint of Dialogflow and replace the placeholder with the query and retrieved documents from the input prompt. Thereby, we enable the language model to generate a response in the form of QA similar to ChatGPT and Claude.

In the scope of query intentions that do not require retrieved evidence, we trained Google Dialogflow over a set of questions from the COVID-QA dataset to enable responses for informative questions. For increasing the human-likeness of our VADF, Google Dialogflow enables a built-in "small talk" feature that initiates conversations with greetings and provides fallback responses to unrecognizable utterances. We further enhance the robustness of the small talk feature by training it over the ChatterBot Language Training Corpus. Google Dialogflow is trained on various intent scenarios to match the users' inputs provided by a rule-based grammar approach capable of restructuring, then categorizing the syntactic structure of a sentence or query, while using ML matching algorithms for pertinent responses. These intended outcomes in the form of the response generation module include answers to general questions on emerging topics such as COVID-19, e.g., *Define the levels of evidence pyramid.* or *In what city did SARS-COV-2 originate?* This enables VADF to effectively respond to users' queries regarding general information on pandemic related topics and the domain-specific science gateway.

V. Measuring Application Utility

Measuring the overall usefulness of chatbot dialog designs for science gateway applications is crucial in understanding human performance, and consequently, the diffusion of these technologies among scientific communities. In this section, we adopt a utility measurement methodology that leverages a set of application usability metrics to calculate a single utility score using the SUM framework [49].

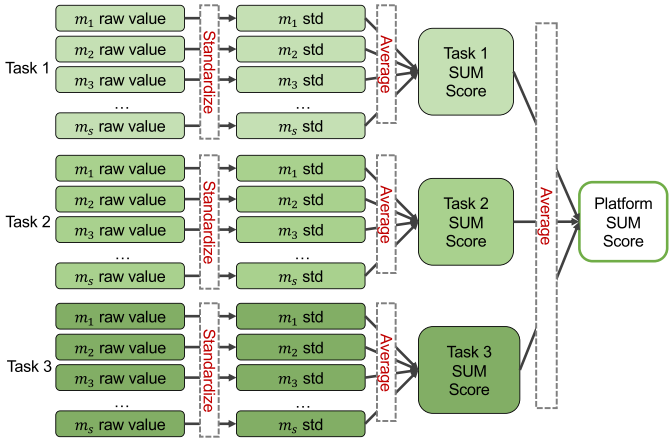


Fig. 3. Single usability measurement framework to calculate z-scores of a given application. The raw metric scores are standardized and then averaged for each task of the application.

Fig. 3 shows the end-to-end process of the SUM framework from raw usability metric scores to task and application scores. The SUM framework examines a set of usability metrics that are mapped to effectiveness, efficiency, and satisfaction. SUM's four major components (time on task, number of errors, completion success, and average satisfaction) have been found to be moderately and significantly correlated across almost 3000 observations, making it a powerful summative tool for reporting these measures in an easy to understand score [8]. These metrics can be represented as either discrete or continuous. Given a set of usability metrics of length s, denoted as $M = \{m_1, m_2, m_3, \ldots, m_s\}$, each metric for a given task is collected in its raw form and converted, or *standardized*, through a z-score calculation. The calculation of z-scores (or normal deviation) pertains to how much a data point deviates from the mean or a specification point. The conversion of these scores allows for quantitative comparisons of application utility between product versions. z-scores can be calculated as

$$z_{m_i} = \frac{x_{m_i} - \bar{x}_{m_i}}{\sigma_{m_i}} \tag{3}$$

where x_{m_i} is the mean value of the ith quantitative metric collected from an application, $\bar{x}_{m_i}$ is the specification limit of ith metric, and σ_{m_i} is the standard deviation of the ith metric.

Based on the framework in Fig. 3, we create three task scenarios for our proposed study with domain users where their performance on those tasks is recorded. These scores get standardized through a z-score calculation and converted to percentage based scores through a z-score table look-up, where all metrics can be averaged for a given task on the product or application. As a result, the SUM tasks scores can then be averaged to calculate the overall product/application score, or the Platform SUM Score.

VI. Implementation of VADF in Real-World Applications

In this section, we detail the process of integrating the VADF on applications in diverse environments. Herein, we show how such applications (e.g., science gateways) with CI/CD pipeline for continuously ingesting data archives such as publications can enable a chatbot agent using the VADF for in-depth knowledge discovery for domain users. We then demonstrate our approach on an exemplar publication analytics science gateway viz., KnowCOVID-19 [9], and detail the website functionality that improves medical user performance on knowledge discovery tasks.

A. Data Pipeline and Orchestration

Facilitating a chatbot dialog design framework such as our VADF on applications in diverse environments requires the continuous collection and processing of data archives in a timely manner. The VADF supports the continuous ingestion of various data archives such as publications, domain-specific QA forums, as well as application code templates for scientific simulations. This is done by using web APIs (e.g., PubMed in the context of medical literature search) or web-crawling tools (e.g., Selenium) over data archives distributed across the web and stored in a cloud server database. Then, we implement a background scheduler that routinely collects new documents at 24 h intervals. Depending on the application and downstream task of the chatbot agent, preprocessing steps of these collected data archives can include but are not limited to: removing punctuation/URLs/stop words, lowering, tokenization, and word stemming.

As shown in Fig. 2, the intent classifier is trained over domain user intents collected from public or private datasets via ML algorithms such as support vector machines and decision trees, as well as DL algorithms such as DNNs and transformer models such as BERT [30]. Depending on the user intent, a generated response can provide either informative information about the scientific application/domain, or scientific information using the ViduraQA retriever. The ViduraQA retriever uses the stored data archives from the database and retrieves the most relevant information based on the user intent. A (pre-trained) language model is utilized on top of the ViduraQA retriever to provide scientific responses to the user. With the rise of transformer-based language models, the VADF can support such popular language models including ChatGPT [4], Claude [47], and Google Dialogflow [48]. These models can be fine-tuned over a specific application to provide pertinent responses to users, while performing in-depth knowledge discovery.

In addition, the VADF can support applications with CI/CD processes for automated deployment. In our use case, we utilize the Jenkins [50] software toolkit, the leading open source automation server for adopting CI/CD. In this process, we use Jenkins to combine both the code and remote service to run the website, as shown in Fig. 4. As the target scientific application and its support features are updated onto a version control application (i.e., GitHub), the Jenkins program will detect the new version code and build a new version code automatically. In this process, if the building result is successful, the code for running the remote service will be updated. On the contrary, if the building fails, it will report errors that can help in debugging. After this step, the online website automatically updates according to the new version code. This orchestration of automatically building and deploying software presents an intuitive and time-saving process in providing medical users faster functionality inputs for knowledge discovery.

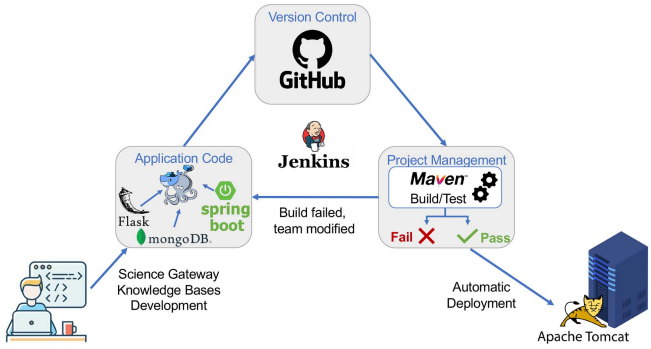


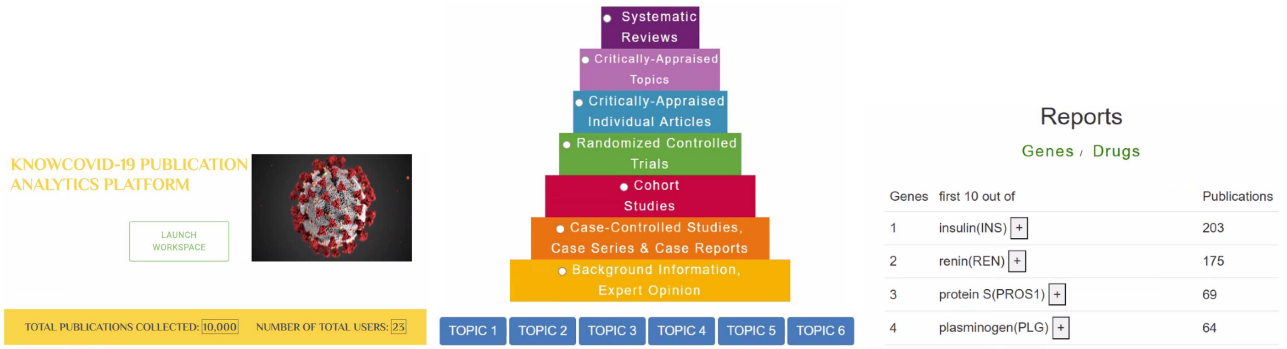
Fig. 4. Jenkins CI/CD architecture over scientific applications such as KnowCOVID-19 science gateway for automated deployment and ingesting new literature archives and knowledge bases of domain-specific terminology for the latest information.

B. KnowCOVID-19 Science Gateway

We exemplify our VADF on an exemplar application viz., KnowCOVID-19 [9], that helps medical users (e.g., clinicians, researchers) to automatically apply filters on the latest articles on the COVID-19 pandemic to obtain reliable information based on the levels of evidence pyramid [51]. The source code and related datasets of VADF implemented for KnowCOVID-19 are openly available at [52]. KnowCOVID-19 ingests and processes documents over the CORD-19 [53], which comprises of more than 1 000 000 scholarly articles from peer-reviewed journals (e.g., New England Journal of Medicine, The Lancet Journal) as well as preprints (e.g., medRxiv, bioRxiv) about the novel coronavirus. The science gateway application uses the Java Spring Boot [54] back-end development framework, which is preconfigured and presugared with a set of technologies that drastically minimize the manual efforts of configuration compared to conventional frameworks. In addition, KnowCOVID-19 is built using a microservices oriented-architecture with Flask [55], which is a lightweight WSGI web application framework for seamless connection to the back-end Java application powered by Spring Boot.

As shown in Fig. 5, the KnowCOVID-19 science gateway includes two important features along with an informative dashboard [see Fig. 5(a)] to enable users in effectively finding literature related to their clinical queries in a timely and automated manner. Herein, we detail each of the main features.

1) Workspace page: This page [shown in 5(b)] enables users to filter literature based on the levels of evidence. Based on our previous work in [9], we used an *evidence-based filtering* method that uses the domain-specific topic model (DSTM) [56] to find latent topics in the CORD-19 corpus and relevant drugs and gene terms.



(a) (b) (c)

Fig. 5. Three main capabilities of the KnowCOVID-19 science gateway application. The dashboard capability introduces the science gateway and all related capabilities for users. The topic filtering capability provides the filtered literature list according to users' input. The reports capability shows the genes and drugs reports for users along with related literature. (a) Dashboard capability. (b) Topic Filtering capability. (c) Reports capability.

2) Reports page: As shown in Fig. 5(c), this page provides users with a text mining report on the amount or frequency of articles for drug and gene term occurrences. We organize the genes and drugs information to present a resultant set of documents to help users find which genes and drugs commonly appear in literature.

The VADF is integrated on the KnowCOVID-19 science gateway interface to further improve medical users performance on literature search tasks. Upon invoking the chatbot dialog agent feature, the VADF starts a conversation with welcoming greetings and humanlike responses to user responses. For knowledge discovery tasks, medical users have the option to either find a publication using the features on KnowCOVID-19 or directly asking the chatbot dialog agent. Selecting the chatbot for literature search can be crucial when users have personalized queries that the KnowCOVID-19 science gateway cannot handle. For example, medical users can inquire about COVID-19 vaccine boosters if the KnowCOVID-19 science gateway does not have that information in the pregenerated topics. In addition, the VADF not only provides publications to medical users but also provides guidance on navigating the science gateway and definitions on terminology.

VII. Experimental Setup

In this section, we provide our experimental setup for the IR performance over our ViduraQA retriever, as well as the QA performance, human performance, and perception study over our VADF. We detail the datasets, metrics, baselines, and implementation details for each experimental study.

A. Datasets

1) Datasets for Information Retrieval: For our IR experiment, we leveraged two widely used large-scale datasets namely, TREC-COVID [11], and NFCorpus [12], specifically designed for IR and zero-shot reranking. The TREC-COVID dataset is a collection of 50 test queries and 171 332 test documents that consists of scientific articles related to COVID-19 and other coronavirus diseases. The dataset includes articles from various sources, such as PubMed Central, bioRxiv, and other relevant preprint archives. The NFCorpus dataset is a collection of 3244 queries and 9964 documents totaling news articles collected from various online news sources. The dataset covers multiple languages and includes rich metadata, such as publication date, source, and topic labels. In our zero-shot retrieval experiment, we use a test set of 325 test queries and 3162 test documents.

2) Datasets for Question-Answering: For our experiments related to QA, we utilized the COVID-QA dataset [23], which consists of 2019 questions created from the CORD-19 dataset and is related to COVID-19 and similar infectious diseases. Each answer in the question-answer pair in the COVID-QA dataset is manually annotated by volunteer biomedical experts.

3) Datasets for Human Performance and Perception: For our human performance and perception study, we collect a set from the CORD-19 [10] dataset. The CORD-19 dataset is a publicly available dataset curated by multiple entities and institutions with abstract and full-text articles on topics such as COVID-19, SARS-CoV-2, and other infectious diseases. This dataset is purposed for publication analytics and text mining using NLP techniques. For our experiments related to human performance and perception, we collected a subset of 10 000 documents for literature search over our KnowCOVID-19 application.

B. Metrics

1) Metrics for Information Retrieval: We evaluate each baseline with the ground truth retrieved documents obtained from the TREC-COVID and NFCorpus datasets. We utilize Precision (P), normalized discounted cumulative gain (NDCG), and mean reciprocal rank (MRR) for the top ten retrieved documents (i.e., $K = 10$) to evaluate the performance of each IR baseline.

2) Metrics for Question-Answering: We utilize a set of metrics commonly used in QA evaluation to test the performance of our VADF which includes: recall-oriented understudy for Gisting Evaluation (ROUGE)-1 and ROUGE-L (longest common subsequence) [57], bilingual evaluation understudy (BLEU) [58], F1, and exact match (EM). To demonstrate the performance on short answers obtained from the ground truth, we denote EM to EM†, where "$\dagger$" represents EM scores for short answers that are 3 words or fewer.

3) Metrics for Human Performance: In the context of usability studies for measuring human performance, the measurements of application utility, standardized by the International Organization for Standardization (ISO), include effectiveness, efficiency, and satisfaction [59], [60]. The usability standards have been adopted to develop benchmark frameworks such as the SUS [24] and SUM. In the following, we provide formal definitions for each metric: success rate, completion time, difficulty score, and number of actions. *Success rate* refers to the user's ability to complete a given task. We simply measure the success rate via binary indication $(Y = 1, N = 0)$ of whether the user successfully completes the task at hand. *Completion time* is the duration (in seconds) of a participant completing a task. We simply calculate this metric by subtracting the recorded end time from the start time of a given task. The specification limit for this measure is calculated by adapting Sauro's and Kinlund's bootstrapping method [61]. While previous studies have utilized SUM to measure satisfaction [8], *Difficulty Score* is measured on a task-by-task basis. We quantify difficulty by using an 8-point scale (rated from 1 to 5 at every 0.5 decimal point) that is reverse-coded when calculating the SUM score, with specification limit of 4. *Number of actions* is the number of events and interactions initiated by a user on an application. We define actions over a given task $A_T = \sum a$, where a is the set of valid actions on a platform as given by $(a_1, a_2, \ldots, a_n)$. These actions include clickable links, completed search queries, switching between tabs, and enabling/disabling the chatbot interface. The specification limit for this measure is calculated by adapting Sauro and Kinlund's bootstrapping method [61].

4) Metrics for Human Perception: To measure human perception for accessing the roles of DOI [2] among the implications of the conditions on the KnowCOVID-19, we leverage the USE questionnaire [25] and DOI metrics [26]. In addition, we also leveraged the SASSI measurement [27] which were utilized for the specific examination of potential relationships between DOI and chatbot usability of VADF. All survey questions for each perception metric are measured on 5-point scales and each adapted measure is described in Table II.

TABLE II
USABILITY QUESTIONNAIRES AND THEIR CONCEPTUAL DEFINITIONS FOR MEASURING THE PERCEPTION OF PARTICIPANTS ACROSS METRICS AND USER ADOPTION AT THE INDIVIDUAL LEVEL

Platform usability measures	
Adapted measures	*Conceptual definition*
USE questionnaire [25]	
Usefulness	The utility of the platform.
Ease of Use	How user-friendly the platform is.
Ease of Learning	How quickly users can learn to use the platform.
Satisfaction	Whether the user is happy with their experience with the platform and would recommend it to others.
DOI measures [26]	
Voluntariness	The degree to which users have a choice in using the platform at work.
Relative Advantage	Whether the platform improves the efficiency of the user's workflow.
Compatibility	Whether the platform would fit into a user's work style and routine.
Image	The degree to which the use of the platform improves one's prestige at work.
Simplicity/Ease of Use	The facility and comprehensibility of the platform.
Result Demonstrability	How clear and explicable the outcomes of using the platform are.
Visibility	To what degree the platform is known and used in one's organization.
Chatbot usability measures	
Adapted measures	*Conceptual definition*
SASSI [27]	
System Response Accuracy	To what degree and how often the chatbot responds correctly.
Likeability	How pleasant and easy the interaction with the chatbot is.
Cognitive Demand	The level of stress and concentration experienced by the user while interacting with the chatbot.
Annoyance	Whether interacting with the chatbot was repetitive, boring, or irritating in any way.
Habitability	The confidence and level of understanding experienced by the user while interacting with the chatbot.

C. Baselines and Implementation Details

1) Information Retrieval: We compare the ViduraQA retriever against BM25 to retrieve a subset of the top N documents from a corpus for downstream IR evaluation. We also test the performance of our ViduraQA retriever as well as the ViduraQA when applied with a set of neural models and compare it to BM25 along with BM25 applied with neural models for zero-shot IR, which has been shown in previous works to improve IR performance via document reranking [62]. Specifically, we utilize a set of dense retrieval models for zero-shot IR such as monoBERT [14] monoT5 [18], Sentence-BERT/SentenceRoBERTa [15], ColBERTv2 [17], dense passage retrieval (DPR) [16], continuous contrastive pretraining (COCO-DR) [19] generative pseudolabeling) (GPL) [20], and Contriever [21].

In terms of the sparse retrievers, we retrieve the top $N = 1000$ documents for each query. In our ViduraQA retriever approach, we experimented with by tuning the hyperparameter $\alpha \in \{0.01, 0.05, 0.10, 0.25, 0.50\}$ and set $\alpha = 0.25$ over the TREC-COVID dataset and $\alpha = 0.10$ over the NFCorpus dataset. For BM25, we set the hyperparameters to $k1 = 1.50$ and $b = 0.75$ over both datasets. Regarding the neural retrievers for zero-shot document reranking, we leverage the checkpoints of each model from HuggingFace Transformers [63] that was fine-tuned over the MS-MARCO dataset [64], which contains a vast collection of real-world queries and corresponding web documents. We utilize the same hyperparameters for each baseline model and utilize BERT-base and T5-base models as the backbone for the document and query encoder. Upon inference, we use the neural retrieval models to compute the relevance scores of the top 1000 documents (provided by the ViduraQA and BM25 sparse retriever) for each query, which is then used to evaluate the performance of the predicted scores based on a set of metrics.

2) Question–Answering: We test our VADF approach against OpenAI's GPT-3.5 Turbo [22], which is commonly used in commercialized products such as ChatGPT. ChatGPT has shown impressive capabilities in answering a wide range of questions in QA settings, including questions on COVID-19, due to pretraining over large text corpora and large parameter size (i.e., 175 B). We use GPT-3.5 Turbo and the ViduraQA retriever as the QA chatbot for our VADF approach. For each query, we leverage our ViduraQA retriever to retrieve a set of candidate articles from the CORD-19 dataset as context to the LLM. We set a constant temperature parameter to $\tau = 1$ and set our ViduraQA retriever parameter to $\alpha =$ 1e-7.

3) Human Performance: The experiment was carried out among 20 participants who were mostly graduate students from a university located in the Midwest and another university located in the Southwest of the United States. The majority of the participants' majors were in the fields of the medical field, media and communication studies, and computer science. Roughly ten participants were assigned to use one search tool out of two conditions, which were KnowCOVID-19 and KnowCOVID-19 with ViduraQA.

All participants of the study involved in the study used the Zoom technology. Consent was obtained from each participant by interviewer(s) of the study and participants' screen activities were recorded and stored in password-protected devices and cloud storage units. Thus, future analysis and coding of research data became more accessible. Participants were first introduced to the concept of the levels of evidence pyramid and given a webpage link where they could go back to, in case they needed a refresher of the levels of evidence to accomplish their tasks. Then participants were asked to perform the same three tasks regardless of the conditions that they were assigned. The difficulty of tasks gradually increased from easy to medium and to hard. For the easy, medium, and hard tasks, participants were given 3, 5, 7 min to complete it, respectively. Failure to complete these tasks on time resulted in an $N = 0$ for completion rate. In turn, the SUM framework defined in Section V were used to calculate the scores of all metrics of the participants for each task and for each condition. The three tasks are shown below.

1) Task 1: Find one article that is labeled as a case-control study. (Easy)
2) Task 2: Find the most recent information regarding COVID-19 vaccine booster where lower level of evidence was accepted. (Medium)
3) Task 3: Find high level of evidence about medicinal treatment of COVID-19 excluding vaccines. (Hard)

4) Human Perception: For our human-perception study, we examine the adoption of the chatbot among its users under the theoretical framework of DOI. Motivated by the DOI challenges among scientific communities, we evaluate the perceived usability and willingness of medical users for adoption of KnowCOVID-19 assisted with the VADF. Therefore, upon finishing the given tasks and interview of the study, a survey questionnaire was sent to each participant where they answered questions about the potential for diffusion of the system for the condition they were assigned during the experiment.

TABLE III
IR RESULTS FOR THE BM25 AND VIDURAQA (VQA) RETRIEVERS, NEURAL MODELS WITH BM25/VQA OVER THE TREC-COVID AND NFCORPUS DATASETS USING PRECISION (P), NORMALIZED DISCOUNTED CUMULATIVE GAIN (NDCG), AND MEAN RECIPROCAL RANK (MRR) AT $K = 10$.

		Evaluation Metrics		
Dataset	**Retriever**	**P@10**	**NDCG@10**	**MRR@10**
TREC-COVID	**BM25** (Sparse)	0.622	0.581	0.790
	ViduraQA (Sparse)	0.550	0.520	0.814
	monoBERT+BM25	0.570	0.519	0.734
	monoT5+BM25	0.682	0.615	0.803
	sBERT+BM25	0.644	0.582	0.799
	sRoBERTa+BM25	0.458	0.406	0.643
	DPR+BM25	0.068	0.145	0.680
	ColBERTv2+BM25	0.494	0.433	0.655
	Contriever+BM25	0.634	0.574	0.782
	COCO-DR+BM25	<u>0.774</u>	**0.735**	**0.947**
	GPL+BM25	0.570	0.503	0.752
	monoBERT+VQA	0.566	0.513	0.721
	monoT5+VQA	0.668	<u>0.607</u>	0.806
	sBERT+VQA	0.652	0.586	0.799
	sRoBERTa+VQA	0.441	0.368	0.628
	DPR+VQA	0.070	0.145	0.700
	ColBERTv2+VQA	0.474	0.429	0.673
	Contriever+VQA	0.612	0.553	0.763
	COCO-DR+VQA	**0.780**	**0.735**	<u>0.945</u>
	GPL+VQA	0.586	0.517	0.750
NFCorpus	**BM25** (Sparse)	0.226	0.295	0.514
	ViduraQA (Sparse)	0.216	0.285	0.497
	monoBERT+BM25	0.230	0.304	0.523
	monoT5+BM25	**0.267**	**0.344**	<u>0.565</u>
	sBERT+BM25	0.241	0.313	0.537
	sRoBERTa+BM25	0.191	0.242	0.457
	DPR+BM25	0.043	0.116	0.431
	ColBERTv2+BM25	0.130	0.156	0.310
	Contriever+BM25	0.259	0.331	0.561
	COCO-DR+BM25	<u>0.266</u>	<u>0.342</u>	0.561
	GPL+BM25	0.213	0.278	0.493
	monoBERT+VQA	0.236	0.318	0.552
	monoT5+VQA	0.262	**0.344**	**0.574**
	sBERT+VQA	0.241	0.313	0.537
	sRoBERTa+VQA	0.191	0.242	0.457
	DPR+VQA	0.042	0.114	0.415
	ColBERTv2+VQA	0.130	0.156	0.310
	Contriever+VQA	0.259	0.331	0.561
	COCO-DR+VQA	0.259	0.331	0.561
	GPL+VQA	0.213	0.278	0.493

Highest scores are boldfaced and second highest scores are <u>underlined</u>.

VIII. EXPERIMENTAL RESULTS

In this section, we present the experimental results to test our VADF and its subsequent components. The following experiments are related to: 1) use of the chatbot design engine for IR; 2) use of the chatbot design engine for IR QA; 3) usability study of our proposed framework implementation and SUM on the KnowCOVID-19 science gateway; 4) perception study of diffusion of innovations.

A. Chatbot Design Engine for Information Retrieval

Table III shows the IR performance results over the TREC-COVID and NFCorpus datasets. In terms of TREC-COVID, the ViduraQA retriever shows comparable performance to the

TABLE IV
QUESTION-ANSWERING RESULTS FOR VADF AND GPT-3.5 TURBO (CHATGPT) OVER THE COVID-QA DATASET USING ROUGE-1, ROUGE-L, BLEU, F1, AND EXACT MATCH (EM†); "†" DENOTES EM SCORES FOR SHORT ANSWERS WITH AT MOST THREE WORDS

Model	ROUGE-1	ROUGE-L	BLEU	F1	EM†
GPT-3.5 Turbo	0.17	0.15	0.03	0.16	0.16
VADF	**0.41**	**0.39**	**0.19**	**0.41**	**0.59**

Bold values under each evaluation metric column represent the highest performing model score.

BM25 in terms of average Precision (BM25: 0.552, VQA: 0.540) and average NDCG (BM25: 0.510, VQA: 0.497) across all baselines. In addition, the results on average MRR across all baselines show that ViduraQA (0.760) marginally outperforms BM25 (0.759), which suggests the linear interpolation smoothing is effective at presenting relevant documents at the top of a limited results set. When applying ViduraQA on COCO-DR for zero-shot retrieval on TREC-COVID, the results show a high increase (↑) in performance compared to BM25 applied with COCO-DR, particularly showing the highest performance gains in P@10 (↑ 0.006). Overall, COCO-DR consistently either outperformed or showed comparable performance against all other baselines across each metric.

Similar to the results on the TREC-COVID dataset, the ViduraQA retriever shows comparable results to BM25 on the NFCorpus dataset in terms average Precision (BM25: 0.210 VQA: 0.205), average NDCG (BM25: 0.272 VQA: 0.271), and average MRR (BM25: 0.495 VQA: 0.497). Upon further inspection of the dataset, the relatively low results compared to the TREC-COVID dataset were due to some query terms that were not present in both the document and corpus. On the other hand, applying ViduraQA on the neural retriever models, mainly monoT5, consistently displayed the highest scores among other baselines,. Specifically, monoT5 applied with the ViduraQA retriever outperformed BM25 in terms of, MRR@10 (↑ 0.060).

On both datasets, neural retrieval models such as sRoBERTa, DPR, and ColBERTv2 displayed lower performance against the ViduraQA and BM25 retriever. Previous studies [62] have indicated that neural retrieval models can perform worse than lexical and sparse retrieval models such as BM25 in zero-shot settings (as shown in Table III). These findings yield further investigation into developing more robust IR-based transfer learning techniques using transformer models. Nevertheless, these results indicate the potential and influence ViduraQA has in performance gain of the neural retrievers for zero-shot IR in scientific applications.

B. Chatbot Design Engine for Question-Answering

The main results are shown in Table IV. Our VADF approach consistently outperforms GPT-3.5 Turbo (i.e., ChatGPT) across all metrics (i.e., ROUGE-1, ROUGE-L, BLEU, F1, and EM†) on the COVID-QA dataset. In terms of short answer evaluation, VADF excels against ChatGPT (EM† of 0.59 vs 0.16), showing its ability to answer complex COVID-19 questions in a direct and precise manner. We also evaluate the performance of each model at different token lengths to measure the accuracy and precision of answers from the CORD-19 dataset. Fig. 6 displays the results

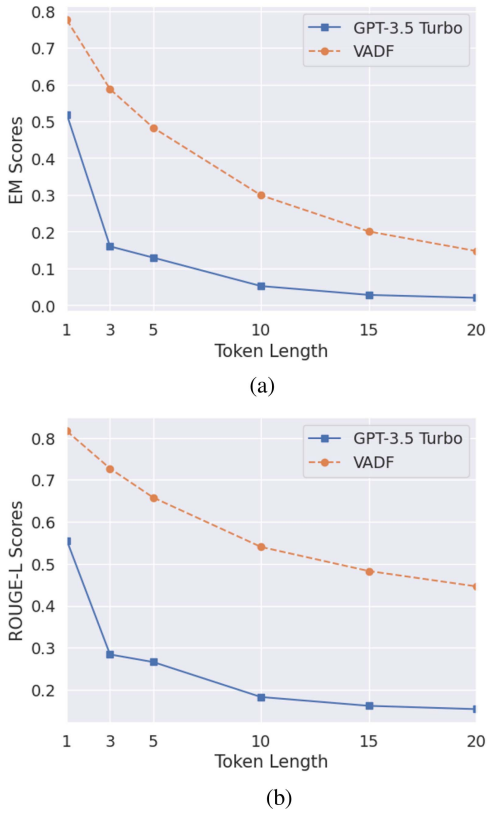


Fig. 6. Performance comparison between GPT-3.5 Turbo (ChatGPT) and VADF using the EM† and ROUGE-L metrics at different word token lengths over the COVID-QA dataset. (a) EM results at different token lengths. (b) ROUGE-L results at different token lengths.

of each model using the EM† [see Fig. 6(a)] and ROUGE-L [see Fig. 6(b)] metrics. VADF consistently outperforms ChatGPT across all token lengths (i.e., $\{1, 3, 5, 10, 15, 20\}$), demonstrating the advantage of our approach in effectively managing complex queries using context retention over the retrieved documents obtain from our ViduraQA retriever. In contrast, ChatGPT exhibits a significant decline for both EM† and ROUGE-L when the answer token length exceeds 1, suggesting challenges in maintaining accuracy and coherence in more extended responses.

Table V provides illustrations on generated responses over given sample queries obtained from the COVID-QA dataset. Similar to results in Table IV, our VADF shows consistent response accuracy, as represented in blue bold text. This is due to our pipeline (shown in Fig. 2) that enhances the responses of language models by retrieving relevant articles as context for answering scientific queries. VADF excels in handling numerical data and historically anchored questions, as seen in its precise answers about the Spanish Influenza timeline and pandemic death tolls. Contrarily, while GPT-3.5 Turbo (ChatGPT) Turbo can provide related responses, it frequently responds with vague or factual errors, as shown in red bold text. This is largely due to the reliance on parametric knowledge, which often causes hallucination [65]. This in turn is insufficient in literature search applications where the precision of results is critical for obtaining accurate information on emerging scientific topics. These results demonstrate the enhanced capability of our VADF approach in delivering factual answers to COVID-related

TABLE V
EXAMPLES OF GENERATED OUTPUTS FROM OUR VADF APPROACH COMPARED TO GPT-3.5 TURBO OVER THE COVID-QA DATASET; THE BLUE BOLD TEXT DISPLAYS CORRECT RESPONSES BASED ON THE GROUND TRUTH, WHILE THE RED BOLD TEXT SHOWS INCORRECT RESPONSES

Question	Ground Truth	GPT-3.5 Turbo (ChatGPT)	VADF
What has been correlated with the pathogenicity of the H5N1 infection?	caspase-dependent apoptosis in airway cells	The presence of **specific genetic mutations**.	**Caspase-dependent apoptosis in airway cells** has been correlated with the pathogenicity of the H5N1 infection.
When did the first wave of the H1N1 swine flu (Spanish Influenza) occur?	a first or spring wave began in March 1918 and spread unevenly through the United States, Europe, and possibly Asia over the next 6 months.	The first wave of the H1N1 swine flu (Spanish Influenza) occurred in the **spring of 2009.**	The first wave of the H1N1 swine flu (Spanish Influenza) occurred in **1918**.
When was the first tissue culture system developed?	1907	The first tissue culture system was developed in the **early 20th century**.	In **1907**
Approximately how many people died during the 1918-1919 influenza pandemic?	During the 1918-1919 influenza pandemic, between 21 and 25 million people died of influenza worldwide.	An estimated **50 million people died during** the 1918-1919 influenza pandemic.	Between **21 and 25 million people died during** the 1918-1919 influenza pandemic.

TABLE VI
SUM AND SUS SCORES (IN %) BY PLATFORM AND TASK FOR THE KNOWCOVID-19 AND KNOWCOVID-19 + VADF APPLICATIONS, WITH DIFFICULTY REVERSE-CODED (-) TO ALIGN WITH SATISFACTION; UP ARROWS (↑) AND DOWN ARROWS (↓) FOR EACH METRIC INDICATE DESIRABLE TASK SCORES

	KnowCOVID-19		**KnowCOVID-19+VADF**	
Tasks	**Scores**	**SUM**	**Scores**	**SUM**
Task 1		**58.1**		52.0
Success ↑	0.78 (±0.4)	**77.8**	0.60 (±0.5)	60.0
Difficulty (-) ↑	3.50 (±1.0)	31.2	3.85 (±1.2)	**45.2**
Time ↓	248.8 (±92.8)	**71.9**	262.9 (±103)	64.8
Actions ↓	14.34 (±9.6)	**51.6**	17.78 (±9.8)	37.8
Task 2		10.1		**38.0**
Success ↑	0.00 (±0.0)	0.00	0.60 (±0.5)	**60.0**
Difficulty (-) ↑	1.56 (±0.5)	0.00	2.65 (±1.1)	**11.1**
Time ↓	361.4 (±109)	31.9	287 (±130)	**57.1**
Actions ↓	44.0 (±23.1)	8.53	32.8 (±28.6)	**23.9**
Task 3		65.0		**70.3**
Success ↑	0.67 (±0.5)	55.6	0.80 (±0.4)	**80.0**
Difficulty (-) ↑	3.72 (±0.7)	**33.7**	2.70 (±0.8)	5.59
Time ↓	269.6 (±124)	94.8	248.1 (±126)	**96.2**
Actions ↓	21.56 (±15.1)	76.1	13.60 (±7.3)	**99.5**
Total SUM		44.4		**53.4**
SUS Platform	62.2 (±0.2)	N/A	**65.8** (±0.2)	N/A
SUS Chatbot	N/A	N/A	76.0 (±0.1)	N/A

Bold values under the "SUM" columns represent the highest performing platform between KnowCOVID-19 and KnowCOVID-19+VADF using the SUM metric.

questions, especially for scientific literature domains for guided knowledge discovery.

C. Usability Study of Implementation of Chatbot and SUM in the KnowCOVID-19 Science Gateway

As shown in Table VI, the total SUM scores were consistently higher in the KnowCOVID-19 with VADF condition when compared to the KnowCOVID-19 condition without the chatbot. This was especially true in Task 2 where participants were unable to successfully complete the task with the platform alone, whereas 60% were able to complete the task with the help of the chatbot. Exceptions to this trend are apparent in Task 1, where participant usability was higher in the platform-only condition, perhaps due to the simple nature of the first task (i.e., the help of the chatbot was not always necessary). Further, while overall usability was higher with the VADF condition in Task 3, the KnowCOVID-19 only condition was rated as comparatively less difficult.

Looking more specifically at the SUM measures, t-tests revealed participants who used KnowCOVID-19 with VADF ($M = 0.6$, $SD = 0.52$) outperformed KnowCOVID-19 ($M = 0$, $SD = 0$) significantly in terms of successfully accomplishing the task, $t(17) = -3.48$, $p < 0.001$. Participants who used KnowCOVID-19 with VADF ($M = 21.56$, $SD = 0.52$) also outperformed participants who used KnowCOVID-19 ($M = 13.6$, $SD = 0.52$) by having significantly less number of actions, $t(17) = 1.49$, $p < 0.01$. These results are promising, as they appear to show that the use of a chatbot aided in the completion of more complex tasks and reduced the amount of actions necessary in completing these tasks.

Regarding the SUS scores shown in Table VI, no statistically significant difference was found between the two conditions. This can be attributed to inadequate sample sizes of each condition, as the general trend of the chatbot condition having a higher usability mean than the KnowCOVID-19 only condition continued to manifest at face value. Examining the SUS scores themselves, it has been found that 68% is the benchmark, with anything above this score being above average, and anything below being below average [66]. Bearing this in mind, the SUS scores for the KnowCOVID-19 platform in both conditions (62.2% and 65.8%) would then be below average, whereas the SUS score for the chatbot at 76% is above average.

Lastly, based on comments during the usability trials, participants' attitudes towards the chatbot, VADF, were generally positive. Once users grasped how the chatbot worked, they were generally pleased by VADF's ability to offer ten relevant articles in response to their queries. A couple of errors were noted,

TABLE VII
AVERAGE RESULTS OF USER PERCEPTION MEASUREMENTS (WITH STANDARD DEVIATIONS) CALCULATED OVER THE CANDIDATE APPLICATIONS: KNOWCOVID-19 AND KNOWCOVID-19 WITH VADF

	KnowCOVID-19	**KnowCOVID-19 + VADF**
Platform Measurements	***Platform Scores***	
USE - Usefulness	3.42 (±1.01)	**3.88** (±1.01)
USE - Ease of Use	2.86 (±1.15)	**3.15** (±1.10)
USE - Ease of Learning	3.44 (±1.11)	**3.47** (±1.21)
USE - Satisfaction	2.78 (±1.09)	**2.88** (±1.07)
DOI - Voluntariness	**4.72** (±0.57)	3.95 (±1.04)
DOI - Relative Advantage	3.11 (±1.44)	**3.42** (±1.22)
DOI - Compatibility	2.82 (±1.44)	**2.93** (±1.28)
DOI - Image	2.30 (±0.75)	**2.60** (±0.84)
DOI - Simplicity/Ease of Use	3.03 (±1.24)	**3.25** (±1.13)
DOI - Result Demonstrability	**3.63** (±0.74)	3.58 (±0.87)
DOI - Visibility	**2.50** (±0.79)	2.45 (±1.01)
Chatbot Measurements	***Chatbot Scores***	
SASSI - Response Accuracy	N/A	3.03 (±1.08)
SASSI - Likeability	N/A	3.71 (±0.83)
SASSI - Cognitive Demand	N/A	3.92 (±0.60)
SASSI - Annoyance	N/A	2.44 (±0.98)
SASSI - Habitability	N/A	2.88 (±1.00)
USE - Satisfaction	N/A	3.26 (±1.13)

Bold values represent the highest platform score between KnowCOVID-19 and KnowCOVID-19+VADF given the corresponding platform measurements in each row.

however. First, VADF has shown to be case and plural-sensitive to a number of query utterances. Secondly, several of the article links retrieved by the chatbot were broken due to data preprocessing issues with the archive files, requiring participants to copy and paste the article information into Google Scholar manually in order to view the articles in full. It is also worth noting that users often had to be reminded a few times that the chatbot was available for them to use, with many initially choosing to use only the platform functions. This could be because the first task was designed to be the easiest task, thus apparently not necessitating the extra help of the chatbot. This may have given participants an inflated sense of confidence in using only the platform in the next two tasks where the chatbot's help was predicted to be more vital to complete the task quickly and effectively.

D. Perception Study of Diffusion of Innovations

The perception study results are shown in Table VII, in which we present the mean and standard deviation calculations from the measurements of the questionnaire. Cronbach's Alpha was used to check the internal consistency of each measurement. While most measures passed, trialability and speed did not meet the recommended alpha coefficient, which needs to be 0.65 or higher [67], and thus were excluded from the results and further analysis.

Independent t-tests demonstrated no statistically significant differences of measurements of USE and DOI between KnowCOVID-19 and KnowCOVID-19 with VADF. As mentioned in previous sections, however, these results should be interpreted as pilot due to a small number of participants in each section. As for SASSI measures of the chatbot, it has been argued that a score of 4/5 is sufficient for a new technology to gain acceptance among most people [27]. Among the scores of SASSI for KnowCOVID-19 with VADF, although no mean of any measure was higher than four, the score of the cognitive demand did approach this number (3.92). This is promising as it indicates that the chatbot's functionality was not mentally difficult for participants to operate when entering their queries.

IX. DISCUSSION

In this section, we provide a detailed discussion on our VADF capabilities. First, we provide a detailed analysis on the KnowCOVID-19 usability study when applying our VADF in the form of a correlation analysis to show the relationship between the usability performance and the perception study. Lastly, we detail the design challenges of the VADF when integrating our framework into a real-world application.

A. Correlation Study

To examine the relationships between the usability performance and the user perceptions, we performed a correlation study among SUM, USE, DOI, and SASSI. When the correlation coefficient r is greater than 0.5 in value, the relationships are considered strong [68]. Pearson correlations revealed the Difficulty usability metric was also strongly and negatively correlated with SUS. Indeed, difficulty was also significantly and negatively correlated with most attributes of the DOI such as perceived relative advantage $[r(17), = -0.708,\ p < 0.01]$, simplicity/ease of use $[r(17), = -0.813,\ p < 0.01]$, and result demonstrability $[r(17), = -0.697,\ p < 0.01]$. The strong yet negative correlations suggest that a system that is hard to use hinders usability of the technology and the prospect of such technology becoming well accepted and diffused among its users [2].

Contrarily, compatibility from DOI perspective was strongly and positively correlated with attributes of USE viz., perceived usefulness $[r(17), = 0.817,\ p < 0.01]$, ease of use $[r(17), = 0.909,\ p < 0.01]$, ease of learning $[r(17), = 0.815,\ p < 0.01]$, and satisfaction $[r(17), = 0.886,\ p < 0.01]$. Most metrics of USE had positive and significant correlations with some attributes of DOI, namely, relative advantage, simplicity, and result demonstrability. Furthermore, SUS was strongly and positively correlated with some attributes of the DOI, including compatibility $[r(17), = 0.866,\ p < 0.01]$, relative advantage $[r(17), = 0.837,\ p < 0.01]$, simplicity $[r(17), = 0.968,\ p < 0.01]$, and result demonstrability $[r(17), = 0.903, p < 0.01]$. These strong and positive correlations indicated that when improving usability of a system, designers may need to put more considerations regarding if their system is compatible with potential users' prior experience and how simple their systems are when compared to other designs [26].

When adding metrics from SASSI into the correlation analysis of KnowCovid-19 with VADF, system response accuracy of the chatbot was strongly and positively correlated with metrics of USE, specifically, perceived usefulness $[r(8), = 0.664,$

$p < 0.05$], ease of use [$r(8), = 0.769$, $p < 0.01$], ease of learning [$r(8), = 0.668$, $p < 0.05$], and satisfaction [$r(8), = 0.687$, $p < 0.05$]. System response accuracy was also strongly and positively correlated with two attributes of DOI, simplicity [$r(8), = 0.799$, $p < 0.01$], and result demonstrability [$r(8), = 0.772$, $p < 0.01$]. Among the correlations between SASSI and DOI, most metrics of SASSI were strongly and positively correlated with simplicity and result demonstrability. Lastly, likability from SASSI also had strong and positive correlations with USE and DOI. Such correlation patterns further bolster the fact that ensuring a system is simple to use and its results are easy to comprehend [2] is crucial for usability improvement [25]. More specifically for chatbots, how accurately the chatbot can respond to its users and the frequency of such accuracy [69] could have positive impacts on how users perceive the simplicity and clarity of the chatbot, which in part, could partially influence how well the system is received and adopted by its users [2].

B. Socio-Technical Challenges in Our VADF

1) Scalability: Scalability can become a significant challenge in chatbot dialog design when recruiting participants with domain expertise for design considerations and usability studies, especially in niche scientific areas. Such participants may be unavailable due to numerous time conflicts or inaccessible due to the limited number of members/practitioners of a niche scientific community. In our target user phase for collecting dialog design considerations (see Section III), the number of participants was not abundant in the target population (8) and the research/work background of participants was not coherent. While having diverse input is necessary for building chatbot dialog systems, the diversity of input should be met with sufficient knowledge around the application domain. The lack of design inputs from an insufficient number of participants may hinder the full capabilities of the chatbot. Furthermore, in the context of DOI, this can also hinder the utility and widespread user adoption of a chatbot in a real-world application.

2) Understanding Scientific Community Needs: Scientific communities are immersed in specialized knowledge and intricate concepts. For instance, medical users on KnowCOVID-19 have distinct needs, and the VADF can be purposefully customized when integrated with KnowCOVID-19 to assist medical users for knowledge discovery by catering to those specific requirements. In the context of DOI theory and widespread chatbot adoption, previous work has studied the role of chatbot designs in science gateways [70], where the scientific community seeks references to support their inquiries rather than seeking straightforward answers from the chatbot [1]. This necessitates chatbot designers to take into account the technical and subject knowledge of the users. However, a significant challenge lies in the lack of comprehensive knowledge about the scientific community and their specific technical and domain expertise. To ensure widespread adoption of the chatbot, it must effectively address the unique concerns of the scientific community. A potential solution could be allowing the chatbot to ask users specific questions and make assumptions about the scientists' needs prior to usage. By doing so, the chatbot can offer more precise and task-oriented services to its users.

3) Verifiable Scientific Evidence: With the new advancements and discoveries that constantly change the conventional knowledge in scientific domains, there presents a challenge for designing chatbots that can identify and convey verifiable evidence based on scientific claims to further improve knowledge discovery for domain users. In addition, while chatbot dialogs are able to provide responses for domain users that guide them in knowledge discovery tasks, generative models such as LLMs have shown to be prone to hallucination [65] by providing falsehoods or misspeaking on user intents. In this work, we have demonstrated how the VADF can to support LLMs such as ChatGPT [4] and Claude [47] and can reduce hallucination by updating their internal states with world knowledge and retrieved information using the ViduraQA retriever. However, LLM-based chatbots are still challenged with providing users with the latest and most truthful information on a constantly evolving topic such as the COVID-19 pandemic. In the context of DOI theory, this can result in low human perception, and hence, a lack of adoption in the scientific communities.

X. Conclusion

In this article, we present an ML-based chatbot dialog design, viz., VADF, that enhances human performance on knowledge discovery tasks and quantifies application utility applications in diverse environments. We formalized a set of chatbot design requirements based on expert feedback to develop VADF that features ML-based models to distinguish between query intents and provide domain users knowledge from relevant publication archives. The VADF framework also features a utility measurement methodology based on a set of usability metrics and computes a utility score when integrated into applications in diverse environments using the single usability metric. Our socio-technical approach demonstrates the effectiveness of the VADF in engaging a community of users in adopting innovative technologies through the diffusion of innovations theory for continuous knowledge discovery.

As part of future work, we aim to study human-in-the-loop neural dialog design methods that enable our VADF to foster a smoother dialog between medical users and the chatbot to improve the adoption at the individual and community levels. In particular, we seek to handle dynamic dialog state tracking and out of domain utterances.

Acknowledgment

Any opinions, findings, 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. The authors would like to thank the following students who have contributed in various parts of this project: E. Milman and C. Kulkarni.

References

[1] R. Oruche et al., "Science gateway adoption using plug-in middleware for evidence-based healthcare data management," *Concurrency Comput.: Pract. Experience*, vol. 35, no. 18, 2023, Art. no. e7195.

[2] E. M. Rogers, *Diffusion of innovations*. Simon and Schuster: New York, NY, USA, 2010.

[3] S. S. Sivarathri et al., "Chatbot guided domain-science knowledge discovery in a science gateway application," in *Proc. 14th Gateway Comput. Environ. Conf.*, 2019, pp. 1–4.

[4] OpenAI, "ChatGPT introduction," 2023. [Online]. Available: https://openai.com/blog/chatgpt

[5] A. Lecler, L. Duron, and P. Soyer, "Revolutionizing radiology with GPT-based models: Current applications, future possibilities and limitations of chatgpt," *Diagn. Interventional Imag.*, vol. 104, no. 6, pp. 269–274, 2023.

[6] D. Su, Y. Xu, T. Yu, F. B. Siddique, E. Barezi, and P. Fung, "CAiRE-COVID: A question answering and query-focused multi-document summarization system for COVID-19 scholarly information management," in *Proc. 1st Workshop NLP COVID-19*, 2020, pp. 1–11.

[7] F. Jelinek, "Interpolated estimation of Markov source parameters from sparse data," in *Proc. Workshop Pattern Recognit. Pract.*, 1980, pp. 381–397.

[8] J. Sauro and E. Kindlund, "Using a single usability metric (sum) to compare the usability of competing products," in *Proc. Hum. Comput. Interaction Int. Conf.*, 2005, pp. 1–9.

[9] R. Oruche et al., "Evidence-based recommender system for a COVID-19 publication analytics service," *IEEE Access*, vol. 9, pp. 79400–79415, 2021.

[10] L. L. Wang et al., "CORD-19: The COVID-19 open research dataset," in *Proc. ACL Workshop Natural Lang. Process. COVID-19*, 2020, pp. 1–12.

[11] E. Voorhees et al., "Trec-covid: Constructing a pandemic information retrieval test collection," *SIGIR Forum*, vol. 54, no. 1, pp. 1–12, 2021.

[12] V. Boteva, D. Gholipour, A. Sokolov, and S. Riezler, "A full-text learning to rank dataset for medical information retrieval," in *Proc. 38th Eur. Conf. Inf. Retrieval*, 2016, pp. 716–722.

[13] S. Robertson and H. Zaragoza, *The Probabilistic Relevance Framework: BM25 and Beyond*. Now Publishers Inc: Norwell, MA, USA, 2009.

[14] R. Nogueira, W. Yang, K. Cho, and J. Lin, "Multi-stage document ranking with BERT," 2019, *arXiv:1910.14424*.

[15] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 3982–3992.

[16] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 6769–6781.

[17] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, and M. Zaharia, "ColBERTv2: Effective and efficient retrieval via lightweight late interaction," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2022, pp. 3715–3734.

[18] R. Nogueira, Z. Jiang, R. Pradeep, and J. Lin, "Document ranking with a pretrained sequence-to-sequence model," *Findings Assoc. Comput. Linguistics*, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, pp. 708–718, Nov. 2020.

[19] Y. Yu, C. Xiong, S. Sun, C. Zhang, and A. Overwijk, "Coco-dr: Combating distribution shift in zero-shot dense retrieval with contrastive and distributionally robust learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 1462–1479.

[20] K. Wang, N. Thakur, N. Reimers, and I. Gurevych, "Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2022, pp. 2345–2360.

[21] G. Izacard et al., "Unsupervised dense information retrieval with contrastive learning," *Trans. Mach. Learn. Res.*, pp. 1–21, 2022.

[22] OpenAI, "GPT-3.5 turbo fine-tuning and API updates," 2023. [Online]. Available: https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates

[23] T. Möller, A. Reina, R. Jayakumar, and M. Pietsch, "COVID-QA: A question answering dataset for COVID-19," in *Proc. ACL Workshop Natural Lang. Process. COVID-19*, 2020, p. 1.

[24] J. Brooke et al., "SUS-A quick and dirty usability scale," *Usability Eval. Ind.*, vol. 189, no. 194, pp. 4–7, 1996.

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

[26] G. C. Moore and I. Benbasat, "Development of an instrument to measure the perceptions of adopting an information technology innovation," *Inf. Syst. Res.*, vol. 2, no. 3, pp. 192–222, 1991.

[27] K. S. Hone and R. Graham, "Towards a tool for the subjective assessment of speech system interfaces (SASSI)," *Natural Lang. Eng.*, vol. 6, no. 3/4, pp. 287–303, 2000.

[28] O. Kolomiyets and M.-F. Moens, "A survey on question answering technology from an information retrieval perspective," *Inf. Sci.*, vol. 181, no. 24, pp. 5412–5434, 2011.

[29] Y. Zhu et al., "Large language models for information retrieval: A survey," *CoRR*, vol. abs/2308.07107, 2023.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., Vol. 1 (Long Short Papers)*, 2019, pp. 4171–4186.

[31] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, and M. Zaharia, "ColBERTv2: Effective and efficient retrieval via lightweight late interaction," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2022, pp. 3715–3734.

[32] Z. Yan et al., "DocChat: An information retrieval approach for chatbot engines using unstructured documents," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Volume 1: Long Papers)*, 2016, pp. 516–525.

[33] A. Lommatzsch and J. Katins, "An information retrieval-based approach for building intuitive chatbots for large knowledge bases," in *Proc. LWDA*, 2019, pp. 343–352.

[34] T. Makhalova, D. Ilvovsky, and B. Galitsky, "Information retrieval chatbots based on conceptual models," in *Proc. 24th Int. Conf. Conceptual Struct.*, 2019, pp. 230–238.

[35] F. Maoro, B. Vehmeyer, and M. Geierhos, "Leveraging semantic search and LLMS for domain-adaptive information retrieval," in *Proc. Int. Conf. Inf. Softw. Technol.*, 2023, pp. 148–159.

[36] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quart.*, vol. 13, pp. 319–340, 1989.

[37] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quart.*, vol. 27, pp. 425–478, 2003.

[38] J. Nielsen, *Usability Engineering*. Burlington, MA, USA: Morgan Kaufmann, 1994.

[39] D. Cardona, O. Werth, S. Schönborn, and M. Breitner, "A mixed methods analysis of the adoption and diffusion of chatbot technology in the German insurance sector," in *Proc. Americas Conf. Inf. Syst.*, 2019, pp. 1–10.

[40] M. Goli, D. A. Sahu, S. Bag, and P. Dhamija, "Users' acceptance of artificial intelligence-based chatbots: An empirical study," *Int. J. Technol. Hum. Interact.*, vol. 19, pp. 1–18, 2023.

[41] R. De Cicco, S. Iacobucci, A. Aquino, F. R. Alparone, and R. Palumbo, "Understanding users' acceptance of chatbots: An extended tam approach," *Lecture Notes Comput. Sci.*, vol. 13171, pp. 3–22, 2022.

[42] S. Laumer, C. Maier, and F. T. Gubler, "Chatbot acceptance in healthcare: Explaining user adoption of conversational agents for disease diagnosis," in *Proc. Eur. Conf. Inf. Syst.*, 2019, pp. 1–13.

[43] R. Oruche et al., "Measurement of utility in user access of COVID-19 literature via ai-powered chatbot," in *Proc. IEEE Appl. Imagery Pattern Recognit. Workshop*, 2021, pp. 1–13.

[44] M. Xuetao, F. Bouchet, and J.-P. Sansonnet, "Impact of agent's answers variability on its believability and human-likeness and consequent chatbot improvements," in *Proc. AISB*, 2009, pp. 31–36.

[45] M. Skjuve, I. M. Haugstveit, A. Følstad, and P. Brandtzaeg, "Help! is my chatbot falling into the uncanny valley? an empirical study of user experience in human–chatbot interaction," *Hum. Technol.*, vol. 15, no. 1, pp. 30–54, 2019.

[46] G. Cox, "Chatterbot language training corpus," Accessed: Oct. 11, 2022. [Online]. Available: https://github.com/gunthercox/chatterbot-corpus

[47] Anthropic, "Introducing claude," 2023. [Online]. Available: https://www.anthropic.com/news/introducing-claude

[48] N. Sabharwal and A. Agrawal, "Introduction to Google dialogflow," in *Cognitive Virtual Assistants Using Google Dialogflow*. Berlin, Germany: Springer, 2020, pp. 13–54.

[49] J. Sauro and E. Kindlund, "A method to standardize usability metrics into a single score," in *Proc. Conf. Hum. Factors Comput. Syst.*, 2005, pp. 401–409.

[50] "Jenkins," Accessed: Sep. 16, 2022. [Online]. Available: https://www.jenkins.io/

[51] D. Timm, "Evidence matters," *J. Med. Library Assoc.: JMLA*, vol. 94, no. 4, 2006, Art. no. 480.

[52] R. Oruche, "KnowCOVID-19 github," 2024. [Online]. Available: https://github.com/rro2q2/KnowCOVID-19

[53] "The COVID-19 open research dataset (CORD-19)," Accessed: Nov. 5, 2021. [Online]. Available: https://www.semanticscholar.org/cord19

[54] "Spring boot introduction," 2021. [Online]. Available: https://spring.io/projects/spring-boot

[55] "Flask introduction," 2021. [Online]. Available: https://www.palletsprojects.com/p/flask/

[56] Y. Zhang, P. Calyam, T. Joshi, S. Nair, and D. Xu, "Domain-specific topic model for knowledge discovery in computational and data-intensive scientific communities," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 2, pp. 1402–1420, Feb. 2023.
[57] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, 2004, pp. 74–81.
[58] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
[59] E. Frøkjær, M. Hertzum, and K. Hornbæk, "Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated?," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2000, pp. 345–352.
[60] J. Jeng, "Usability assessment of academic digital libraries: Effectiveness, efficiency, satisfaction, and learnability," *Int. J. Libraries Inf. Stud.*, vol. 55, no. 2-3, 2005, pp. 96–121.
[61] J. Sauro and E. Kindlund, "How long should a task take? Identifying specification limits for task times in usability tests," in *Proc. Hum. Comput. Interact. Int. Conf.*, 2005, pp. 1–5.
[62] R. Ren et al., "A thorough examination on zero-shot dense retrieval," in *Proc. Findings Assoc. Comput. Linguistics*, 2023, pp. 15783–15796.
[63] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations*, 2020, pp. 38–45.
[64] T. Nguyen et al., "MS MARCO: A human generated machine reading comprehension dataset," 2016, *arXiv:1611.09268*.
[65] Z. Ji et al., "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, 2023.
[66] J. Sauro, *A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices*, Measuring Usability LLC: Denver, CO, USA, 2011.
[67] C. Goforth, "Using and interpreting Cronbach's alpha," *Univ. Virginia Library Res. Data Serv. Sci.*, 2015.
[68] L. A. Baxter and E. R. Babbie, *The Basics of Communication Research*. Cengage Learning: Boston, MA, USA, 2003.
[69] K. Hone, "Usability measurement for speech systems: Sassi revisited," in *Proc. SIGCHI Conf.*, 2014, pp. 1–4.
[70] K. F. Kee, P. Calyam, and H. Regunath, "The role of Vidura chatbot in the diffusion of KnowCOVID-19 gateway," *Hum.- Mach. Commun.*, vol. 3, pp. 47–63, 2021.

Roland Oruche (Member, IEEE) received the B.S. degree in information technology in 2019 from the University of Missouri-Columbia, Columbia, MO, USA, where he is currently working toward the Ph.D. degree in computer science.

His research interests include machine learning, natural language processing, human-centered AI, and dialog systems.

Xiyao Cheng is currently working toward the Ph.D. degree in computer science with the College of Engineering, University of Missouri, Columbia, MO, USA.

Before coming back to the graduate school, she worked as a Data Analyst with an IT technology company that provides map navigation. Her research interests include knowledge graph-based data modeling and recommendation system.

Zian Zeng is currently working toward the undergraduate degree with the University of Hawaii at Manoa, Honolulu, HI, USA.

His research interests include machine learning, natural language processing, human-centered artificial intelligence (AI), and AI for health care.

Audrey Vazzana received the B.S. degree in computer science from the Truman State University, Kirksville, MO, USA, in 2024. She is currently working toward the Ph.D. degree in computer science with the University of Nebraska-Lincoln, Lincoln, NE, USA.

Her main research interests include machine learning and wireless communications.

MD Ashraful Goni is currently working toward the Ph.D. degree in Media & Communication with the College of Media & Communication, Texas Tech University, Lubbock, TX, USA.

His research interests include generative AI and its applications in media, particularly in supporting ethnic media and enhancing digital journalism.

Bruce Wang Shibo is currently working the Ph.D. degree with the College of Media and Communication, Texas Tech University, Lubbock, TX, USA.

Before coming back to graduate school, he worked in the cattle industry where he cowboyed while producing media content. His research interests include the intersections among artificial intelligence, presumed media influence, and drought in media.

Sai Keerthana Goruganthu received the B.Tech. degree in computer science from B. V. Raju Institute of Technology, Tuljaraopet, India, in 2022. She is currently working toward the graduate degree in computer science with the University of Missouri-Columbia, Columbia, MO, USA.

Her research interests include working with machine learning techniques, data analytics, and artificial intelligence.

Kerk Kee received the Ph.D. degree in organizational communication, with an emphasis on workplace technologies, from the Department of Communication Studies, University of Texas at Austin, Austin, TX, USA, in 2010.

He is currently an Associate Professor with the College of Media & Communication, Texas Tech University, Lubbock, TX, USA. His research interests include the diffusion of innovations in the scientific, health, and environmental contexts.

Dr. Kee's work has been funded by the Bill & Melinda Gates Foundation, the Robert Wood Johnson Foundation, and US National Science Foundation (NSF), include an NSF CAREER award in 2015.

Prasad Calyam received the M.S. and Ph.D. degrees in electrical and computer engineering from the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA, in 2002 and 2007, respectively.

He is currently a Professor with the Department of Computer Science, University of Missouri-Columbia, Columbia, MO, USA. Previously, he was a Research Director with the Ohio Supercomputer Center. His research interests include distributed and cloud computing, computer networking, and cybersecurity.