A Survey on the Recent Advancements in Human-Centered Dialog Systems

ROLAND ORUCHE, University of Missouri, Columbia, United States
SAI KEERTHANA GORUGANTHU, University of Missouri, Columbia, United States
RITHIKA AKULA, University of Missouri, Columbia, United States
XIYAO CHENG, University of Missouri, Columbia, United States
ASHRAFUL MD GONI, Texas Tech University, Lubbock, United States
BRUCE W. SHIBO, Texas Tech University, Lubbock, United States
KERK KEE, Texas Tech University, Lubbock, United States
MARCOS ZAMPIERI, George Mason University, Fairfax, United States
PRASAD CALYAM, University of Missouri, Columbia, United States

Dialog systems (e.g., chatbots) have been widely studied, yet related research that leverages artificial intelligence (AI) and natural language processing (NLP) is constantly evolving. These systems have typically been developed to interact with humans in the form of speech, visual, or text conversation. As humans continue to adopt dialog systems for various objectives, there is a need to involve humans in every facet of the dialog development life cycle for synergistic augmentation of both the humans and the dialog system actors in real-world settings. We provide a holistic literature survey on the recent advancements in *human-centered dialog systems* (HCDS). Specifically, we provide background context surrounding the recent advancements in machine learning-based dialog systems and human-centered AI. We then bridge the gap between the two AI sub-fields and organize the research works on HCDS under three major categories (i.e., Human-Chatbot Collaboration, Human-Chatbot Alignment, Human-Centered Chatbot Design & Governance). In addition, we discuss the applicability and accessibility of the HCDS implementations through benchmark datasets, application scenarios, and downstream NLP tasks.

This work is supported by the National Science Foundation under awards: OAC-2006816 and OAC-2007100. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
Authors' Contact Information: Roland Oruche, University of Missouri, Columbia, Missouri, United States; e-mail: rro2q2@mail.missouri.edu; Sai Keerthana Goruganthu, University of Missouri, Columbia, Missouri, United States; e-mail: sgmhz@missouri.edu; Rithika Akula, University of Missouri, Columbia, Missouri, United States; e-mail: rahvk@missouri.edu; Xiyao Cheng, University of Missouri, Columbia, Missouri, United States; e-mail: xcheng@mail.missouri.edu; Ashraful Md Goni, Texas Tech University, Lubbock, Texas, United States; e-mail: mgoni@ttu.edu; Bruce W. Shibo, Texas Tech University, Lubbock, Texas, United States; e-mail: shibobruce.wang@ttu.edu; Kerk Kee, Texas Tech University, Lubbock, Texas, United States; e-mail: kerk.kee@ttu.edu; Marcos Zampieri, George Mason University, Fairfax, Virginia, United States; e-mail: mzampier@gmu.edu; Prasad Calyam, University of Missouri, Columbia, Missouri, United States; e-mail: CalyamP@missouri.edu.

ACM 0360-0300/2025/05-ART258
https://doi.org/10.1145/3729220

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Human-centered computing** → *Human computer interaction (HCI)*; • **Computing methodologies** → **Artificial intelligence**; **Natural language processing**; **Discourse, dialogue and pragmatics;**

Additional Key Words and Phrases: Chatbot dialogs, human-centered artificial intelligence, natural language processing, machine learning, human-computer interaction

ACM Reference Format:
Roland Oruche, Sai Keerthana Goruganthu, Rithika Akula, Xiyao Cheng, Ashraful Md Goni, Bruce W. Shibo, Kerk Kee, Marcos Zampieri, and Prasad Calyam. 2025. A Survey on the Recent Advancements in Human-Centered Dialog Systems. *ACM Comput. Surv.* 57, 10, Article 258 (May 2025), 36 pages. https://doi.org/10.1145/3729220

1 Introduction

The advancement of dialog systems has influenced multiple research fields in adopting principles of **artificial intelligence** (**AI**) and **natural language processing** (**NLP**), and has benefits for various applications such as healthcare, customer service, and education [27, 40]. A dialog system aims to interact with humans through many forms such as speech, visual, or text conversation for purposes including open-ended conversations or targeted task accomplishment. Early works in this field showed an intelligent machine engaging in conversations with humans using pattern-matching techniques [197].

The emergence of big data, increased commodity computation power, and robust **machine learning** (**ML**) and **deep learning** (**DL**) algorithms, have transformed the development of chatbot dialogs from rule-based systems to neural dialog systems [20]. This has engendered two types of dialog systems, as shown in Figure 1 (left): *task-oriented* and *open-domain systems*. Task-oriented dialog systems are developed to guide users in achieving specific goals [105]. When a user queries the chatbot, the question is interpreted through semantic understanding and keyword extraction through the **Natural Language Understanding** (**NLU**) module. Then, the Dialog State Tracker tracks the conversation history with the user and the Dialog Policy module creates an optimal policy to form a prediction. Finally, a response is made through the **Natural Language Generation** (**NLG**) module. Open-domain, general dialog systems are designed to engage users in conversations on various of topics such as chit-chat and open-world knowledge, while demonstrating empathy [149]. These systems are often developed using an encoder-decoder architecture. Despite these advancements, dialog system development often lack the consideration of potential opportunities and risks for human stakeholders, which ultimately poses challenges to societal trust and safety in human-AI interactions [161]. As more dialog systems continue to scale with emerging capabilities, and as user adoption rates continue to increase, there is a need to incorporate humans in the related AI life cycle [208].

Human-centered AI (**HCAI**) [161], which has been evolving at the intersection between AI and **human-computer interaction** (**HCI**) research, aims to leverage human and machine intelligence for synergistic interaction to develop more human-aligned models. This area of study has since matured in recent years and has permeated across a spectrum of fields including AI, psychology, and cognitive science [201]. Previous work has extensively studied HCAI approaches and has leveraged techniques such as NLP that underpin the technology for text-based dialog systems. Existing works such as in [20, 195, 201] provide a broad survey of HCAI techniques for various areas in ML such as computer vision, and NLP by summarizing works based on their tasks, goals, human interactions, as well as feedback learning mechanisms. These works, however, provide a limited view of ML-based dialog system development using human-centered approaches. Therefore, a connection between the

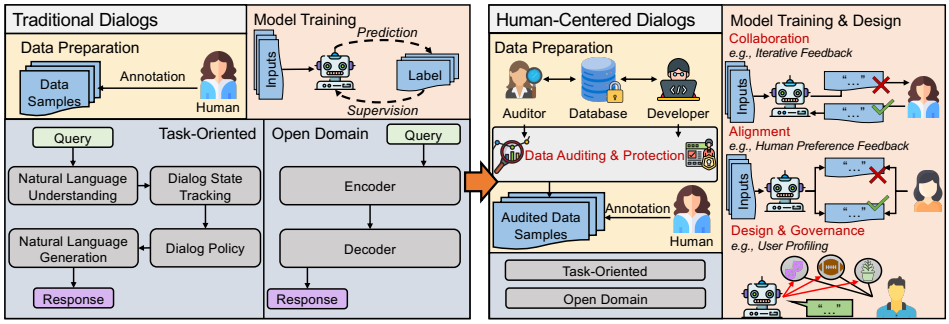

Fig. 1. Illustration showing paradigm shift from traditional dialog systems to human-centered dialog systems. We display discernible differences between the two paradigms as the architecture on the left shows how traditional systems do not feature any human involvement. The architecture on the right shows how human-centered dialog system development involves humans in various ways in the development life cycle.

two widely studied areas must be investigated and established to characterize the recent progress in this latent sub-area of study.

In this article, we provide a holistic view of the recent advancements in ***human-centered dialog systems*** (**HCDS**) through a holistic literature survey. HCDS refers to the design of conversational agent systems that considers humans stakeholders (e.g., individuals, communities, society) as an intricate part of the development life cycle. Figure 1 demonstrates the foundational differences between traditional dialogs and HCDS through developmental areas such as data preparation and model training. Specifically, HCDS leverages human input for privacy protection and auditing during data preparation, as well as human supervision, values, and personal preferences through the dialog model training and design. To the best of our knowledge, we are the first to provide a comprehensive survey on dialog system modeling with a human-centered approach. Specifically, we employ a literature review method to collect a set of peer-reviewed publications from notable conference and journal venues from databases such as the ACM Digital Library, ScienceDirect, IEEE Xplore, OpenReview, and ACL Anthology. Due to vast amounts of literature on HCDS, we solely focus on the scope of *text-based* HCDS. Based on our article collection, we provide a background on both ML-based dialog systems and HCAI; and detail the core concepts within recent advancements. We then establish the connection between the two AI sub-fields and outline a comprehensive scope on HCDS with details on how humans are considered in the dialog system development life cycle. Specifically, we categorize HCDS into the sub-groups of *Human-Chatbot Collaboration*, *Human-Chatbot Alignment*, and *Human-Centered Chatbot Design & Governance*. Next, we detail the recent advancements of each sub-group and discuss its applicability and accessibility through benchmark datasets, application scenarios, and downstream NLP tasks. Lastly, we list the research challenges from a socio-technical perspective and provide future research directions as well as open-ended discussions for improving human involvement and increasing trustworthiness in the dialog system development life cycle.

The taxonomy presented in Figure 2 describes key elements and recent advancements in HCDS, application tasks, as well as the challenges and future directions. The remaining sections are as follows: Section 2 presents the background of ML-based dialog systems and HCAI. In Section 3, we detail the recent advancements of HCDS and discuss notable datasets. Section 4 describes HCDS under different application scenarios (e.g., open-domain, task-oriented), and downstream NLP tasks. In Section 5, we detail the research challenges in HCDS from a socio-technical perspective. In Section 6, we provide a set of future research directions. Finally, Section 7 concludes the article.

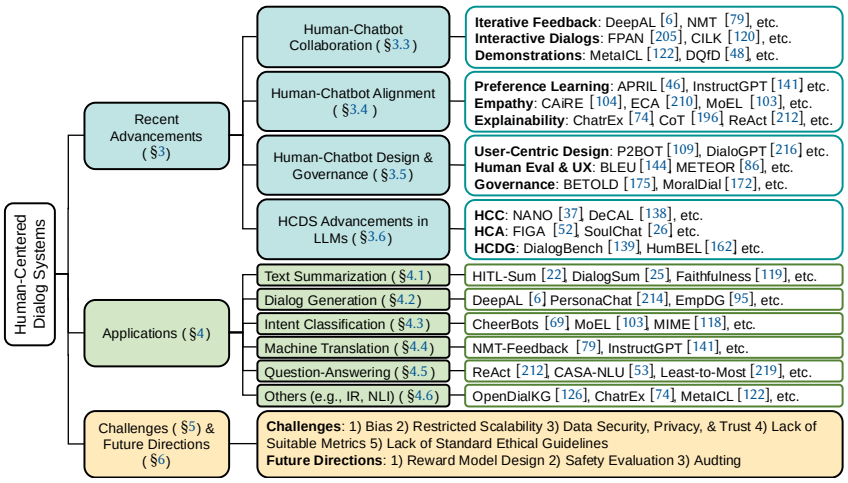

Fig. 2. HCDS survey taxonomy showing the scope of recent advancements, applications, challenges, and future directions.

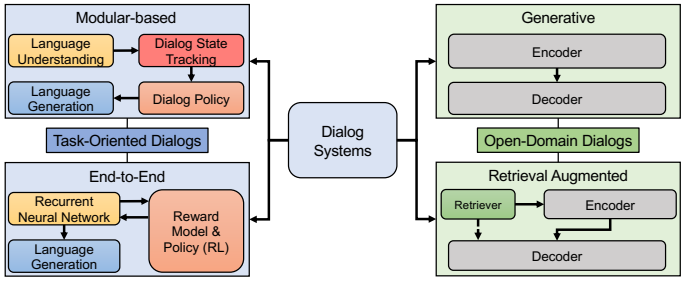

Fig. 3. Illustration showing the classification of the dialog system architectures including task-oriented (e.g., pipeline, end-to-end) and open-domain (e.g., generative retrieval-based) that are examples of common models and methods for each category.

2 Background

In this section, we provide background on various aspects of dialog systems and human-centered AI. Specifically, we detail the core methodologies and variants behind each concept.

2.1 Dialog Systems Concepts and Advancements

A dialog system, also known as a conversational agent or chatbot, is an AI program designed to interact with humans through natural language. Originating in the 1960s with Joseph Weizenbaum's ELIZA at MIT, which utilized rule-based and pattern-matching techniques [197], dialog systems have evolved to mimic human-like communication and provide engaging responses. They are now widely applied in areas such as customer service, personal assistance, education, and entertainment [25, 27, 40]. With the advancement of ML/DL approaches, the taxonomy of dialog systems broadly categorized under task-oriented and open-domain dialog systems. As illustrated in Figure 3, common task-oriented architectures include *pipeline* and *end-to-end*, while open-domain architectures feature *generative-based* and *retrieval-based* approaches. In the following, we briefly describe each of the core concepts and advancements within ML-based dialog systems.

2.1.1 Pipeline Approach. Pipeline-based dialog systems employ a series of ML models arranged in a modular, sequential format to support natural conversational exchanges. These systems are organized into essential components: NLU, dialog management, and NLG. The NLU component interprets human utterances by categorizing them into semantic slots. Dialog management, through a dialog state tracker, logs each conversational turn and maintains the dialog history, managing the conversation's flow and decision-making. Subsequently, based on the current state of the conversation, the dialog policy dictates the next steps. The NLG module then transforms these decisions into textual responses, creating outputs that maintain the natural flow of human speech. Notable implementations such as RASA NLU [11], promote automatic learning and are applied in areas like finance and education to reduce labor costs, especially in university admissions. However, these systems face challenges like complex feedback integration and module interdependencies that require extensive retraining on new data, highlighting the pipeline's tightly coupled nature [218].

2.1.2 End-to-End Approach. End-to-end dialog systems simplify the traditional multi-component architecture into a single model, using a **recurrent neural network** (**RNN**) to encode the conversational context into a vector representation [218]. This approach eliminates separate NLU and dialog state tracking modules, integrating them into a unified process. These systems are highly flexible, easily retrained on larger datasets, and free from the constraints of domain-specific semantic slots. Advances include using pre-trained word embeddings to enhance task success rates, demonstrated in Reference [33], and incorporating **reinforcement learning** (**RL**) to adapt via user feedback, despite potential challenges like mismatched state distributions during offline training and online application [106]. Innovations such as the gated memory networks for hospital reservations [176] and relational dialogue systems for medical diagnostics [206] show improved generalization with fewer labels. However, the opacity of these "black box" systems often limits their interpretability compared with more modular, pipeline-based approaches.

2.1.3 Generative-Based Methods. Generative-based dialog systems leverage the robust computational power and extensive memory of neural networks to generate responses that emulate human conversation from a wide array of examples. The shift to **sequence-to-sequence** (**Seq2Seq**) learning [29], utilizing an encoder-decoder RNN network, marked a significant advancement in how user inputs are transformed into chatbot responses. This method was further enhanced with the development of transformer architectures [182], which introduced self-attention mechanisms to replace recurrent layers, enabling parallel processing and better management of long-range dependencies. The progression continued with the introduction of **Large Language Models** (**LLMs**) such as OpenAI's **Generative Pre-trained Transformer** (**GPT**) [1], which advanced response generation through extensive self-supervised training on large text corpora. This training allows models to grasp complex aspects of language such as structure, syntax, semantics, and context. Despite their capabilities, these models face challenges such as limited context understanding and the tendency to produce misleading responses, known as "hallucination" [102], underscoring the ongoing need for meticulous development and enhancement of these sophisticated NLP tools.

2.1.4 Retrieval-Based Methods. Retrieval-based dialog systems select the most suitable response from a predefined database rather than generating new responses from scratch. Traditionally, these systems utilized keyword matching combined with ML techniques to identify relevant responses to user inputs, as seen in the sequential matching network by the work in Reference [202], which integrates pattern matching with an RNN to maintain important contextual information. However, these methods often require extensive labeled data. Recent advancements have incorporated more sophisticated neural models such as convolutional neural networks and RNNs, which remember past interactions and significantly improve dialog handling [51]. LLMs have further revolutionized

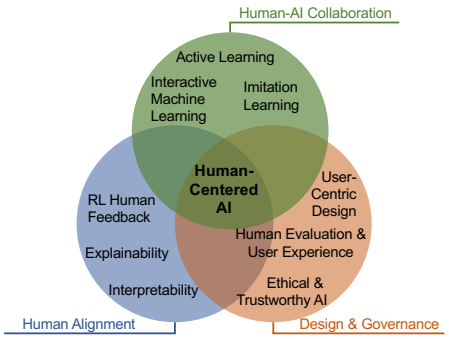

Fig. 4. HCAI taxonomy that includes three major concepts which are Human-AI Collaboration, Human Alignment, and Design & Governance. We detail the most common methods within each concept and show their spatial relation in the triple Venn diagram.

retrieval-based systems by minimizing reliance on predefined responses and enhancing context understanding through pre-trained knowledge. These models either add a separate retrieval component to an LLM or integrate a retrieval layer within the LLM architecture itself, as demonstrated by newer developments [90], enabling more dynamic and contextually aware interactions.

2.2 Human-Centered AI Concepts and Advancements

HCAI represents an essential intersection between human innovative thinking and the logical capabilities of AI. Early works of HCAI have focused on human-in-the-loop approaches where the learner (agent) is used to query or generalize the knowledge distilled from the trainer (human) [156]. As shown in Figure 4, we segment HCAI into three categories—(a) human-AI collaboration, (b) human alignment, and (c) human-centered chatbot design & governance. These categories serve as the foundation for a harmonious synergy of humans and their use of AI, promoting innovation and advancement while ensuring AI adoption meets responsible and ethical standards.

2.2.1 Human-AI Collaboration. Human-AI collaboration refers to a dynamic partnership and interaction between humans and AI systems across a broad spectrum of tasks and domains. Early collaborations between human and AI models/system included *active learning* (AL) [158] approaches, which allows the learner (i.e., agent) to select unlabeled samples and query them to the oracle (i.e., human) to annotate for faster convergence of training a model with fewer data samples. *Imitation Learning* (IL) [64] involves the process by which an AI agent learns from human-provided demonstrations, effectively replicating human behavior within specific scenarios. This multifaceted sub-field in human-AI collaboration includes various categories such as behavioral cloning, direct policy learning, inverse RL, and reward-based imitation. ***Interactive Machine Learning*** **(IML)** [127], is a human-AI collaborative technique where humans provide feedback to the agent through an interactive learning process or setting. In essence, this approach advocates for a collaborative synergy between humans and computers, wherein each entity interactively leverages its respective strengths at any given moment within a specific task or workflow. Variants of IML include: (i) rapid feedback loops for users [4] for swift generation of models, which can continually refine based on user input, and (ii) collaborative IML [55] for enables humans to manipulate algorithms in real-time, contributing to model improvement as it is generated.

2.2.2 Human-AI Alignment. In the context of HCAI, human-AI alignment pertains to the crucial goal of making sure that AI systems are created, developed, and implemented in a way that

is consistent with human values, requirements, and goals. Methods such as RL has emerged successfully when incorporating human values in the feedback process. Early works on RL for human alignment have included frameworks for training RL agents to learn a policy based on human input using both positive and negative reward signals [75]. In recent years, ***reinforcement learning from human feedback*** (**RLHF**) extends RL for human-AI alignment to optimizing language models [223]. It encompasses a structured approach consisting of four main phases: pre-training the language model, collecting human feedback based on text-generated outputs, fitting the reward model based on human judgments, and learning an RL policy over the reward model. Alternatively, **Direct Preference Optimization** (**DPO**) [147] aims to simplify RLHF by eliminating the reward model and optimizing only over binary human preferences. In the context of HCAI, ***Explainabilty* and *interpretability*** (**XAI**) are human-AI alignment methods that pertains to alleviating the core "black box" issue of AI systems by implementing clear and transparent explanations for their predictions [208]. This has engendered notable research work including: (i) LIME (Local Interpretable Model-Agnostic Explanations) [152], which is designed to explain the predictions of black-box ML models, and (ii) SHAP (SHapley Additive exPlanations) [112] a method based on cooperative game theory that assigns SHAP values to each feature in a prediction to explain its contribution to the final prediction. Other works on XAI have developed ensemble techniques, which are model-agnostic schemes that enhances transparency across a spectrum of ensemble-based AI systems [21].

2.2.3 Human-Centered Design & Governance. Human-centered design and governance refers to a collection of comprehensive design frameworks and policies established to safeguard the development, deployment, and use of AI technologies while upholding an unwavering commitment to people's welfare, security, morality, and societal values. In the aspect of human-centric design, *user-centric design* aims to develop AI systems that are not only technically capable but also inherently user-friendly, highly effective, and precisely consistent with the values and expectations of the humans they serve. Existing work has shown to create AI solutions that enhance and empower user experience through a dynamic and iterative process [142, 208]. Other human-centered design techniques such as *human evaluation & user experience* describe how humans can provide input when assessing the performance of an AI technology [142]. Existing work has investigated how these techniques can deliver unique insights into how users interact with and perceive an AI system [208]. Furthermore, AI governance methods such as ethical and trustworthy AI refers to the creation and implementation of AI systems that address individual and social concerns around safety, privacy, and trust at the core of the entire AI life cycle [161]. Existing work on ethical AI have put emphasis on fairness, transparency, accountability, and the protection of user privacy, and have fostered the development of ethical frameworks within HCAI [180]. In addition, previous works in trustworthy AI have investigated the ethical and legal dimensions of AI systems in various domains [145].

3 Human-Centered Dialog Systems

HCDS are conversational agent systems (or chatbot systems) that are designed in a way that keeps humans at the center of the development cycle. In this section, we present a taxonomy on HCDS and detail the recent advancements under three sub-groups: **Human Chatbot-Collaboration** (**HCC**), **Human-Chatbot Alignment** (**HCA**), and **Human-Centered Chatbot Design & Governance** (**HCDG**).

3.1 HCDS Review Methodology and Taxonomy

Given the broad spectrum of existing work in HCAI and dialog systems, we present a literature review methodology over HCDS concepts and methods. Our methodology is split into three stages,

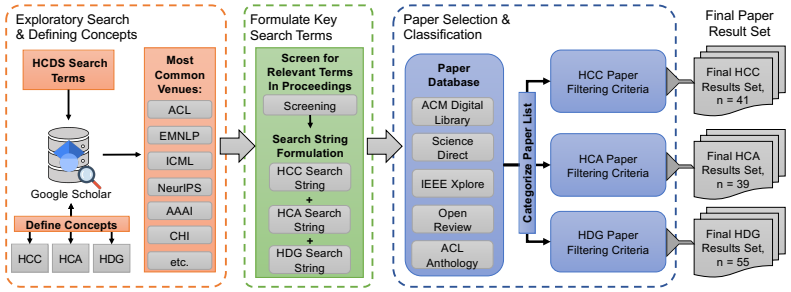

Fig. 5. Paper search and selection process for HCDS taxonomy categories: Human-Chatbot Collaboration (HCC), Human-Chatbot Alignment (HCA), and Human-Centered Chatbot Design & Governance (HCDG).

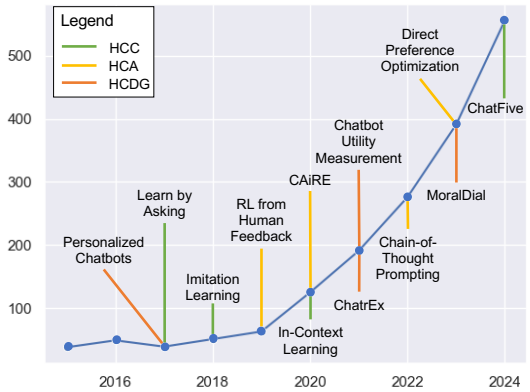

Fig. 6. Statistics representing the number of papers collected per year. Representative milestones of HCDS technologies are included in the years from 2015–2024.

as shown in Figure 5. We first perform an exploratory search on Google Scholar and randomly sampled 1,000 published articles with abstracts and then identified a common set of non-preprint venues (e.g., conferences, journals, books) that are related to HCDS. As shown in Figure 5, the majority of the published articles from this initial search are from venues which include but are not limited to: ACL, AAAI, EMNLP, ICML, NeurIPS, and ACM SIGCHI. In the second stage, we formalize key search terms for our literature review that will be used to search over multiple digital libraries. Specifically, we screen the proceedings from the above venues to find the most relevant set of key terms that fit within the HCDS concepts (i.e., HCC, HCA, HCDG). In our final stage, we utilized our formulated key search terms to perform literature search across multiple digital libraries and academic search engines including ACM Digital Library, ACL Anthology, Science Direct, IEEE Xplore, and OpenReview. We filtered our result set removing papers without full text and narrowing our search to papers published from January 1st, 2015 until December 31st, 2024. This results in 637 papers. In addition, we filter our final result set based on the criteria scope of HCC, HCA, HCDG. Thus, we identified a total of 135 papers for the purposes of this literature survey contributions.

In Figure 6, we illustrate the increase of papers submitted using our HCDS review methodology and identify recent advancements and techniques from 2015–2024. The plot demonstrates notable milestones throughout the timeline for each HCDS category, where more human-centered techniques have been introduced. In the scope HCC, dialog systems were developed through the years starting with interactive ML techniques such as *Learn by Asking* (2017) and *ChatFive* (2024), to

Table 1. Collection of HCDS-Related Datasets with their Respective Domains and HCDS Taxonomy Category

Datasets	Domain	HCDS Category	HCDS Methods
MultiWOZ [14]	Restaurant, Hotel, Attraction, Taxi, Train, Hospital, Police	Human-Chatbot Collaboration, Human-Chatbot Alignment	Reinforcement Learning, Interactive Learning, Imitation learning, Active Learning, Explainable AI
IMDB [117]	Movie Reviews	Human-Chatbot Collaboration	Reinforcement Learning
SST-2 [166]	Movie Reviews	Human-Chatbot Alignment	Explainable AI
AGNEWS [215]	News	Human-Chatbot Alignment, Human-Chatbot Collaboration	Explainable AI, Active Learning
QNLI [189]	Wikipedia	Human-Chatbot Collaboration, Human-Chatbot Alignment	Explainable AI, Active Learning, Reinforcement Learning
QQP [65]	Social QA Questions	Human-Chatbot Alignment	Explainable AI
Cornell Movie-Dialogs Corpus [31]	Online teaching, E-Learning	Human-Chatbot Collaboration	Reinforcement Learning, Active Learning
BookCorpus [222]	Free Novel Books	Human-Chatbot Collaboration	Interactive Learning, Active Learning
EmpatheticDialogues [149]	Open-Domain	Human-Chatbot Collaboration	Reinforcement Learning, Active Learning
DailyDialog [99]	Daily human-to-human Communication	Human-Centered Chatbot Design & Governance	Active Learning
Dialog State Tracking Challenge Dataset [35]	Reddit	Human-Centered Chatbot Design & Governance	Interactive Learning, Active Learning, Imitation Learning, Reinforcement Learning from Human Feedback
Schema-Guided Dialogue Dataset [150]	Banking, Media, Calendar, Travel, Weather	Human-Centered Chatbot Design & Governance	Reinforcement Learning
PersonaChat [214]	Textual description-based user profiles	Human-Chatbot Collaboration, Human-Chatbot Alignment, Human-Centered Chatbot Design & Governance	Interactive Learning, Reinforcement Learning, Trustworthy AI, User-centric Design
Stanford Multi-Turn [36]	Weather Information Retrieval, Calendar Scheduling, and Point-of-Interest Navigation for in-car Assistant	Human-Chatbot Collaboration, Human-Centered Chatbot Design & Governance	Interactive Learning, Active Learning

(Continued)

learning from human (expert) demonstrations in imitation learning (2018) and in-context learning. In Human-Chatbot Alignment, a notable breakthrough in alignment was **reinforcement learning from human feedback** (**RLHF**) (2019) and DPO (2023), followed by empathizing with users (2019) and incorporating interpretability techniques via Chain-of-Thought (2022). Lastly, personalized chatbots (2017) became more popular in the scope of Human-Centered design and Governance as well as more sophisticated utility measurements (2021) and benchmarks such as MoralDial (2023).

3.2 Datasets

Given the diversity of subtopics within recent advancements in HCDS, we provide a list of datasets and benchmarks from highly cited publication venues in Table 1. We detail each dataset in terms of their domain and categorize them by their HCDS taxonomy category, and the particular HCDS methods that were used for experimentation or evaluation. Then, we detail the recent advancements in HCDS through our proposed taxonomy that is built around these datasets.

Table 1. Continued

Datasets	Domain	HCDS Category	HCDS Methods
WikiMovies [121]	Movies	Human-Chatbot Collaboration	Interactive Learning
bAbI/Personalization [12, 72]	Open-domain QA	Human-Chatbot Collaboration, Human-Centered Chatbot Design and Governance	Reinforcement Learning, Interactive Learning, User-Centric Design
TriviaQA [73]	Wikipedia	Human-Chatbot Collaboration	Imitation Learning
TruthfulQA [102]	Open-domain QA	Human-Chatbot Alignment, Human-Centered Chatbot Design & Governance	Imitation Learning, Reinforcement Learning
CoQA [151]	Open-domain	Human-Chatbot Collaboration	Imitation Learning
Natural Questions [83]	Wikipedia	Human-Chatbot Collaboration	Imitation Learning
DialogSum [25]	Education, Work, Medication, Shopping, Leisure, Travel	Human-Chatbot Collaboration	Reinforcement Learning
Reddit Causal Conversations [47]	Open-domain	Human-Chatbot Collaboration, Human-Chatbot Alignment	Reinforcement Learning, Imitation Learning
Webis-TLDR-17 [185]	News	Human-Chatbot Collaboration, Human-Chatbot Alignment	Reinforcement Learning, Imitation Learning
CNN/Daily Mail [60]	News	Human-Chatbot Collaboration, Human-Chatbot Alignment	Reinforcement Learning, Imitation Learning
EmpatheticDialogue [149]	Emotion Dialogue	Human-Chatbot Alignment, Human-Centered Chatbot Design & Governance	Reinforcement Learning, User-centric Design, Interactive Learning
Emotional Dialogues in Open Subtitles (EDOS) [198]	Emotion Dialogue, Movies	Human-Chatbot Alignment	Reinforcement Learning
OpenDialKG [126]	Manually annotated human-to-human role-playing dialogs	Human-Chatbot Alignment, Human-Centered Chatbot Design & Governance	Explainable AI, User-centric Design
FEVER: Fact Extraction and VERification [177]	Wikipedia	Human-Chatbot Alignment, Human-Centered Chatbot Design & Governance	Explainable AI, User-centric Design

3.3 Human-Chatbot Collaboration

The collaboration between humans and dialog systems involves the process in which both actors team up together (i.e., human-AI teaming) using their skill sets to achieve some common goal. Therefore, human-chatbot collaboration includes various human-the-loop concepts such as iterative feedback, interactive dialog systems, and learning from demonstrations. Herein, we dive into three concepts and detail their respective human-chatbot collaboration methods.

3.3.1 Iterative Feedback. Iterative feedback is a form of collaboration in which humans provide their inputs for chatbot models to learn from. Two forms of iterative feedback include *offline* and *online* feedback. The primary difference between the two depends on whether humans provide feedback during model training or inference on the application. In the following, we describe the recent advancements in online and offline iterative feedback in human-chatbot collaboration.

Recent work on offline iterative feedback has focused primarily on human-in-the loop methods such as active learning and RL for efficient supervised training and reward function optimization, respectively. Authors in Reference [79] improve **neural machine translation** (**NMT**) models by developing an offline RL approach that utilizes human feedback from logged user activities on an e-commerce platform. The authors showed dialog models can perform well in NMT by iteratively learning from explicit feedback (i.e., user ratings) and implicit feedback (e.g., search behavior). Similarly work in Reference [67] demonstrates the effective use of an offline RL approach that

takes advantage of explicit and implicit human feedback for producing desirable dialog responses. Authors in Reference [22] use RL for offline feedback on downstream abstractive summarization tasks. This includes *local feedback*, in which humans provide feedback on salient information needed for appropriate summaries, and *global feedback* which compares summaries based on a summarization criterion. They use RL to build reward models and a summarization policy for local and global feedback. Other works such as in Reference [170], employ different iterative feedback techniques such active learning, which enables humans to provide offline feedback to dialog systems for improving tasks related to dialog generation. Despite this, offline feedback approaches are often limited with training fixed and static datasets which can affect the dialog system's performance due to the lack of real-time interactions and inevitable distribution shifts in real-world settings [184].

Online iterative feedback has shown to improve such challenges of fixed datasets for more robust dialog performance on downstream tasks. For instance, the work in Reference [93] showed that training over online interactions by using reward-based imitation yields improves the the dialog's question-answering ability. Authors in Reference [6] alleviate the challenges of offline supervision for dialog generation by developing a novel end-to-end interactive dialog system. This novel interaction method features online active learning that enables humans to provide dialog response feedback during conversations. The work in Reference [105] develops a hybrid interactive feedback model using both offline and online approaches in which a supervised task-oriented dialog is further trained by interacting users online to improve its response quality. While common iterative feedback approaches such as AL and RL show potential in optimizing a model over human feedback, such algorithmic policies can fail to produce high-quality responses if human feedback is noisy and/or adversarial or if the model cannot properly interpret human judgments as a useful feedback signal [80]. In addition, dialogs can fail from *bandit feedback*, where the model does not make any assumptions about that data and relies on sparse rewards from human feedback.

3.3.2 Interactive Dialog Systems. Dialog systems can be designed with an interactive component such as a user interface or command line interface. These designs promote the collaboration between humans and chatbots, as they allow humans to be effective in collaborative tasks such as providing feedback and dialogs to ask questions or optimize over human feedback. We describe the recent advancements in interactive learning for building collaborative systems with humans in the loop.

Interactive dialog systems with user interfaces are commonly built for human annotators to improve training by correcting the dialog response output [70]. For example, the work in [54] proposes a self-feeding chatbot that extracts new human-chatbot conversation samples for training. It utilizes an interface to interact with the end-user and improve its dialog conversation through estimating conversation quality metrics (i.e., user satisfaction) and the prediction of feedback provided by unsatisfied users. The development of user interfaces can often be used to rest the robustness of dialog models [187]. Authors in this proposed work develop a system that features a user interface for enabling humans to generate adversarial examples in the task of question-answering. These visualizations on UI can help end-users better understand the failure points of dialog systems in QA-related tasks. Other tasks where user interfaces are developed as a medium for human-chatbot interaction are text summarization [46] information retrieval [204]. Despite this, providing detailed feedback on user interfaces can be time-consuming and costly, which could deter online users or annotators from long feedback sessions and ultimately effect the dialog performance [207]. Allowing the chatbot dialogs to learn which feedback is crucial. Asking or probing the end-user can help improve its training and response quality [54].

Interactive feedback can be in the form of an inquiry, specifically when a chatbot asks users questions and uses them as samples to train over. Authors in Reference [94] alleviate the aforementioned interactive feedback limitations by enabling an end-to-end dialog to be interactive during

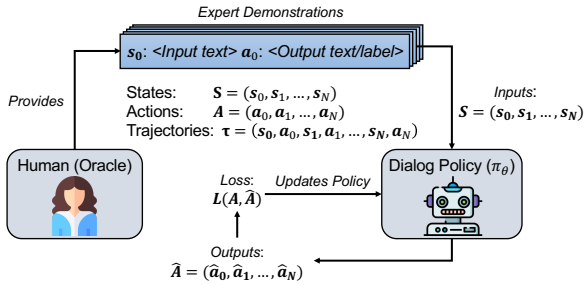

Fig. 7. Example imitation learning pipeline in which human (oracles) provides a set of expert demonstrations, denoted as trajectories τ to the dialog agent. The dialog agent learns a policy π_Θ over the state-action pairs and updates the policy from the given loss function.

the conversation by allowing dialog learners to *learn* how to ask questions and receive feedback from human teachers. The work in Reference [120] leverages continual learning techniques to build a question-answering dialog with fixed knowledge bases that interacts with the user by asking questions to unknown knowledge queries by the user. Authors in Reference [61] enable dialogs to query users to clarify ambiguous questions using RL to improve NLU and response quality. Authors in Reference [46] aim at alleviating the need for voluminous interaction rounds for summarization feedback by developing an active learning method that queries samples to the user. In the scope of conversational recommenders, the work in Reference [205] leverages online user feedback to estimate user preferences by inquiring users about attributes in a multi-round fashion.

3.3.3 Learning from Demonstrations. Dialog systems can collaborate with humans by learning from human (or teacher) demonstrations. These concepts include methods such as imitation learning, behavior cloning, and in more recent advancements, prompt learning and in-context learning. Figure 7 displays a general imitation learning problem setup for which the dialog agent has a set of inputs states $S = (s_0, s_1, \ldots, s_N)$ and actions $A = (a_0, a_1, \ldots, a_N)$ and produces a set of trajectories $\tau = (s_0, a_0, s_1, a_1, \ldots, s_N, a_N)$ of size N based on demonstrations provided by the human oracle. The dialog agent then learns a parameterized policy π_Θ that is updated based on a loss objective $L(A, \hat{A})$, where $\hat{A} = (\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_N)$ are the predicted action outputs. In the following, we detail the recent advancements in which dialogs learn from human demonstrations.

The work in Reference [199] proposes a dialog-based language learning model using an end-to-end memory network (MemN2N), that receives natural and implicit supervision during human-chatbot conversation. The authors employ imitation learning and are able to leverage human demonstrations of utterances to train the MemN2N model for effective response generation during conversation. Authors in Reference [106] develop a dialogue imitation learning technique that enables the dialog agent to obtain lessons from human teaching. During the conversation, the dialog converses with the user. When the agent made mistakes, they requested users to fix them and show what predictions and actions were expected of the agent. The work in Reference [160] shows how human demonstrations can enable RL-based policy algorithms in effective learning persuasion in dialog system conversation. *S^2Agent* [190] uses policy shaping and reward shaping to learn how to leverage human demonstrations to learn a dialog policy in task-oriented settings. Authors in Reference [100] create ImitKD (imitation-based knowledge distillation) for autoregressive dialogs in NLG tasks that treat the teacher model as the oracle and corrects the student model at every generation step as the student model explores its generation strategy during training. Authors in Reference [49] show that using expert demonstrations from human oracles is crucial for effectively training an RL-based dialog policy algorithm with large state and action spaces.

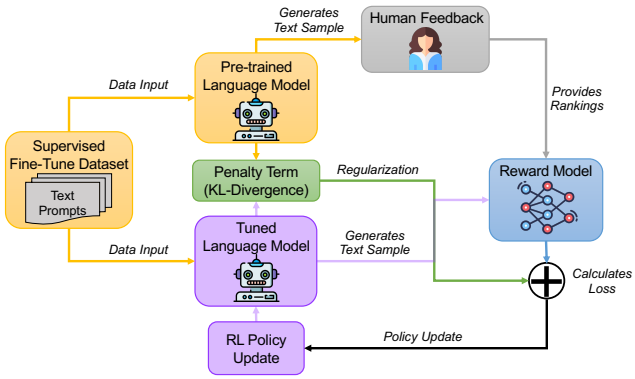

Fig. 8. Architecture for reinforcement learning from human feedback. After the language is pre-trained on a large corpus, a reward model is built based on preference feedback from humans over the model's generated outputs. An RL policy is built to optimize the reward of a newly generated output.

Similarly, authors in [48] question using expensive expert demonstrations that are rule-based and develop a deep-Q learning method that shows that an accurate dialog manager can be learned over weaker and cheaper demonstrations.

The advent of LLMs has given rise to processes that leverage humans to provide examples - thereby helping with the overall training process for various downstream NLP tasks. The majority of these works include methods such as prompt learning and **in-context learning (ICL)** [108]. The article in Reference [13] was one of the first works to demonstrate the powerful capabilities of pre-trained LLMs in various NLP task scenarios using very little expert demonstrations (or few-shot learning). The work on few-shot learning capabilities continued to advance in terms of improving in-context learning from self-supervised training [23] and applying few-shot learning methods to small(er) language models [44]. While authors in Reference [23] show that pre-trained models are not suited to optimize its training for ICL, the authors in Reference [122] propose MetaICL, a meta-learning framework that enables LLMs to learn new in-context tasks at test time. They show that by simply conditioning the LLM on a few expert demonstrations, it yields comparable performance to models that are fully fine-tuned on the target domain as well as bigger models with significantly more parameters. Previous studies have investigated why LLMs perform well when conditioning on a few samples. In one example, authors in Reference [123] show that labels are not required for in-context examples. They suggest a huge contribution toward the success of ICL and include examples of the label space, distribution of the input text, and the format of the input sequence.

3.4 Human-Chatbot Alignment

Human alignment is a key aspect in the field of HCAI, which aims to ensure that AI systems are meticulously designed, developed, and implemented in harmony with human values, preferences, and goals. Recent works have employed human alignment approaches to build dialog systems that more closely align with their users' values. Thus, we detail state-of-the-art human alignment strategies such as preference learning, empathy, and explainability.

3.4.1 Human Preference Learning. Human preference learning exemplifies how conversational agents can more closely align toward human values via human feedback using their judgments and values. This has become more evident as LLMs emerge as the preferred dialog system in academic research and commercial applications. In Figure 8, we present an example architecture that shows

how human preference learning is integrated to train the pre-trained LLM for using RL on a specific task. In the following, we present preference learning techniques, mainly RLHF, in recent studies.

Early works showed success in using RL from human-bandit feedback based on cardinal 5-point ratings and pairwise preference [81] for machine translation. Authors describe how RL can show improvement over automated metrics scores such as BLEU [144] with small amounts of human feedback. The work in Reference [46] develops APRIL (Active Preference-based RL) for NLP systems in the context of text summarization. APRIL aligns with human preferences over generated summaries and creates a policy for efficiently searching for the near-optimal summary. Other work such as [213] show how explicit turn-level human feedback in dialog response generation leads to high correlations with human evaluations. While receiving preference feedback does show improvements in various dialog tasks, small-scale human feedback may not be feasible for the chatbot model to minimize the perplexity of their target responses. The work in Reference [45] alleviates this challenge by creating a large-scale human feedback prediction dataset via crowdsourcing. A set of pre-trained language models are used with a ranking method that showed a stronger correlation with human preferences than previous baseline models.

Large pre-trained models such as LLMs have shown significant performance on various NLP tasks using RLHF. The work in Reference [171] uses a large collection of human comparisons on LLM-generated summaries to create a reward model that produces a scalar to closely mimic human preference. Authors then develop an RL policy that trains over the reward model given a newly generated summary. InstructGPT [141] uses RLHF on an LLM to demonstrate that larger-size models (in terms of parameter size) do not always follow a user's intent and generate appropriate responses. The work in Reference [8] considers leveraging diverse human feedback to develop a reward model that quantifies and ranks a consensus of preferences that are aligned with the overall group of human judges. These works show how using reward modeling from human feedback as proxies for human judgment outperforms automated metrics such as ROGUE [101] and BLEU [144] and baseline language models during human evaluation.

Despite the significant improvement in LLMs through human alignment techniques, LLMs struggle to provide truthful answers and are prone to hallucination due to their reliance on internal representation (or weights) that are limited to the updating or uncovered knowledge in the real world [119, 141]. Authors in [76] propose to incorporate human feedback in the pre-training phase where LLMs typically train over the unfiltered internet. They present multiple objectives to study the tradeoff between LLM performance and alignment. Specifically, they show that the integrated human feedback in the pre-training phase leads to increased preference satisfaction. Authors in Reference [110] propose SENSEI, a novel RL algorithm that focuses on learning an embedding of human preferences at each generative step. SENSEI employs an Actor-Critic framework, where the Critic distributes the rewards, mimicking the assignment procedure of humans, and the Actor guides the model generation toward the maximum reward.

3.4.2 Empathetic Dialog Systems. Empathy has become significantly acknowledged as a key element in enhancing mutual understanding and alignment between two agents. This awareness has spurred the inclusion of empathetic elements within conversational systems, now referred to as empathetic dialog systems. Previous work demonstrates that integrating elements such as emotional reasoning, ancillary knowledge, and the alignment of emotions within response generation models significantly improves their performance and user experience [116].

Authors in Reference [69] develop CheerBots, a deep RL to generate responses that align with users' desired emotional tones, emphasizing empathy and appropriateness. The authors conducted a user study, which validated the model's effectiveness in real-world scenarios, highlighting its ability to provide empathetic and emotionally suitable responses in dialog conversations. CAiRE

[104], an empathetic chatbot, incorporates empathy into generative dialogue systems for engaging in immersive conversations with users. A feedback loop is created from continuous user feedback to enhance the quality of its responses while identifying and eliminating undesirable response patterns through active learning and negative training, which ultimately reduces the occurrence of unethical or inappropriate responses.

The study in Reference [210] introduces a structural approach to developing empathy-capable embodied conversational agents, addressing a key challenge in the realm of ECAs (Embodied Conversational Agents). This structural approach is designed to serve as a foundational component for the addition of further empathetic and behavioral functions while maintaining a conversational agent that is both efficient and quick to respond. Authors in Reference [103] approach empathy in dialog systems by putting emphasis on understanding user emotions to provide a suitable response. MoEL generates a distribution of emotions in response to a user's emotional state and seamlessly integrates the output from specific decoders, denoted as *listeners*, that are trained to generate the correct empathetic response. Other works investigate techniques to improve empathetic dialogs such as capturing nuanced user emotion through textual conversations [95], emotion mimicry [118], and leveraging external knowledge via context graphs to learn implicit emotions [96].

Furthermore, this has sparked initiatives for building benchmarks and datasets. EmpatheticDialogues [149] is a notable dataset that includes 25,000 samples of emotional dialogue. Experiments in this work show that generative dialogs are perceived to be more empathetic by human evaluators. Authors in Reference [198] introduced a substantial dialogue dataset called **Emotional Dialogues in Open Subtitles** (**EDOS**), comprising 1 Million emotionally rich dialogues sourced from movie subtitles. Notably, a model fine-tuned on this dataset achieved top performance in terms of diversity metrics. This dataset holds significant promise for the development of empathetic conversational emotion analysis within dialog systems.

3.4.3 Explainable Dialog Systems. Enabling transparency and comprehensive decision tracing in dialog systems is vital to ensure that humans understand the successes and failures of their decision-making processes and constructive feedback, respectively. **Explainable AI** (**XAI**) has emerged in early work as a useful technique/tool for increasing transparency, trust, and alignment between humans and chatbot systems [133]. These techniques include explainable techniques and interfaces, as well as generating human-like reasoning steps in the model's thought process.

The work in Reference [126] develops a conversational reasoning model based on a knowledge graph that was built from a new dialog dataset called, *Open-ended Dialog Knowledge Graph (OpenDialKG).* Authors in this work develop a model that learns symbolic dialog context transitions through the knowledge graph and show that using self-attention over sentences provides a robust and explainable prediction. Authors in Reference [62] propose to learn natural language actions that represent conversational utterances in spanned words, which enables an explainable dialog generation process. Authors in Reference [221] alleviate the need for expensive human annotations in NLG settings by developing a variational EM framework that treats natural language as latent variables that learn the reasoning steps that underpin the neural language model. The study in Reference [98] proposes a two-stage framework for explainable dialog response generation that enables users to adjust and interpret the interaction pattern between generating a response first, and then instantiating the final response.

Recent literature has investigated prompting methods, namely **chain-of-thought** (**CoT**) prompting, to better develop reasoning steps in LLMs [193, 196, 219]. Authors in Reference [196] demonstrate that adding CoT prompting as intermediate steps in LLMs makes models interpretable while yielding large performance gains on complex tasks related to arithmetic, commonsense, and symbolic reasoning compared with baseline LLMs. Authors in Reference [193] propose a new decoding

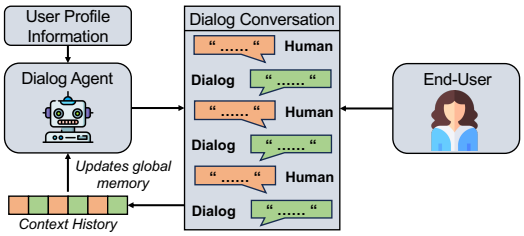

Fig. 9. Example design for user-centric dialog system in which the chatbot is developed to be more personalized and engaged with the end-user. The end-user's profile information is integrated with the chatbot to provide personalized responses, and the context history saves previous conversations and updates the chatbot's global memory.

strategy in interpretable LLM reasoning viz., *self-consistency*. This strategy is superior to the naive greedy decoding strategy by sampling a diverse set of reasoning paths and marginalizing the reasoning paths to select the most consistent answers. Although CoT allows developers and users to view the intermediate steps of LLMs, it still struggles to avoid hallucination [196]. Authors in Reference [212], overcome this issue by proposing ReAct, an interactive web-based LLM that generates logical reasoning paths and task-specific actions. They show that by interacting with a simple web-based API, the chatbot generates human-like task-solving abilities that are more interpretable than previous CoT reasoning baseline models.

3.5 Human-Centered Chatbot Design & Governance

Building dialog systems and making them easily available and accessible to humans of varied needs and expertise is crucial for user engagement and widespread adoption. In addition, safe and ethical considerations must be put in place to build trust and continual adoption of dialog systems. In this section, we detail the recent advancements in human-centered chatbot design and governance through user-centric design, human evaluation, and user experience as well governance.

3.5.1 User-Centric Design. User-centric design is a form of human-centered chatbot design in which HCI designs are developed for a target user group. In the context of conversational agents, user-centric design has been implemented as the personalization of model responses by conditioning a generative dialog model on the responses of their respective speakers. This personalization is highly dependent on the approaches for harnessing information such as *user profiling* and *context history* in conversation. Figure 9 displays an example HCDS design that improves personalization and style in dialog conversations by considering personal attribution from the user's profile information and retaining the context history in previous conversations. In the following, we detail recent advancements in user-centric design approaches in HCDS.

Authors in Reference [211] segment user profile information into the following: (i) factual knowledge [214], and (ii) stylistic modifiers [157], which covers personal, emotional, and situational choices. To boost user engagement, it is crucial to incorporate factual information that is unique to a given user. User trivia [214] includes personal data like the user's name or location, or user attributes like occupation; whereas, user preferences include opinions, values, and beliefs [157]. Situational stylistic modifiers view the usage of language as a function of the situation, the genre, and the target audience, which includes formal/informal [194], professional/colloquial, personal/impersonal, or polite/impolite [42]. The works in References [7, 39, 69] feature chatbots that generate responses that are in tune with users' desired emotional tones, placing a strong emphasis on empathy and appropriateness. The work in Reference [7] presents three different approaches to include emotional

content in encoder-decoder neural conversation models: affect-based loss functions, affect-based word embeddings, and effectively varied beam search for decoding.

Another important factor in personalization is maintaining the context histories in human-chatbot conversations [53, 78, 216]. For example, authors in Reference [53] built a **context-aware self-attentive NLU** (**CASA-NLU**) model that employs the current user utterance such as prior intentions and dialog acts. The work in Reference [114] builds a Profile Model, which uses distributed profile representation to understand user personalities, and a global memory to store conversation context from similar users. This allows recommendation policy changes and appropriate language style selection to be made based on the user profile. A study in Reference [109] uses cognitive science approaches for inducing high-quality chit-chat conversations to propose, P2BOT, which enhances conversation generation via mutual persona perception and openly models interlocutors' comprehension. In the context of LLMs, authors in Reference [216] proposed the novel DialoGPT—a large, tunable neural conversational response generation model - that aims to improve conversational systems by generating more relevant, contentful, and context-consistent responses.

3.5.2 Human Evaluation and User Experience. Human evaluation and user experience are a crucial part for designing dialog systems with human-centered approaches. These concepts enable conversational agents to be improved by allowing users to display their performance and perception when interacting with agents. Herein, we detail the common approaches used for human evaluation (e.g., *utility-driven*, *user-driven*) and user experience (e.g., *satisfaction*) for HCDS design and development.

Dialog systems are evaluated depending on their type and intended application [163]. Common evaluation approaches include: (i) utility-driven, which focuses on the content and performance of the chatbot, and (ii) user-driven, which measures the quality based on the generated chatbot response. *Utility-driven evaluation* concentrates on how much the generated response deviates from the matching ground truth and uses representative metrics such as BLEU [144] and METEOR [86]. *User-driven evaluation* encompasses the quality of the outcome responses of the chatbot with a user-centric approach. In quality-based criteria, fluency and diversity (e.g., distinct diversity [92]) are among some of the metrics that can be used to successfully measure the dialog system's response. The work in Reference [173] suggests a model-driven statistic called Hits@1/N, which determines how well the given answer can be automatically categorized to the appropriate user or user group. These two types of metrics together give a good idea of the performance of the dialog system by measuring its user-centric behavior. Authors in Reference [183] propose an automatic evaluation process for conversational AIs through metrics as proxies for human judgment that granularly analyze the conversational agents that are not captured in human ratings. Although there are several methods for automatically evaluating the quality of responses produced, humans still play a crucial part in evaluating user systems due to the lack of a strong correlation between automated and human evaluations [107].

In user experience, users' interaction with a conversational agent can dictate the success in utility and perceptions in adoption. User satisfaction is a common aspect of user experience that gauges the chatbot's performance, human performance, and perceived adoption. A article in Reference [41] develops a qualitative framework for chatbots that provides insights on some of the key drivers for user experience and satisfaction. Authors in Reference [137] evaluate user performance and perception of elements like user interface design, functionality, and derived information insights, through a usability study [136] and identify challenges of user adoption and satisfaction of publication analytics at the individual level. Using annotations on six dialogue aspects—relevance, interestingness, understanding, task completion, efficiency, and interest arousal—the work in

Reference [163] offers a detailed examination of user satisfaction in task-oriented dialog systems. In the domain of customer service, the work in Reference [82] examines customer satisfaction reports and demonstrates that user experience varies significantly based on the challenges of encouraging users to engage with the chatbot. Other works show how human-like responses and chatbot personality can have significant positive effects on user experience and satisfaction [66, 134, 165].

3.5.3 Governance. Incorporating governance into the life cycle of dialog systems is crucial to ensure that practitioners make ethical decisions as they move through the AI development and deployment life cycle. Recent human-centered methods have been incorporated for the assurance of designing safe and responsible dialog systems such as establishing *ethical guidelines*, *transparency*, and *privacy protection*.

The article [58] discusses the ethical concerns in dialogue systems research encompassing a range of critical issues. These issues include implicit biases inherent in data-driven systems, the emergence of adversarial examples such as misspelled words and paraphrased sentences that can potentially manipulate these systems, sources of privacy violations stemming from the use of learned models, safety concerns, especially in the context of RL systems, and challenges related to reproducibility. This has raised the initiative to discuss the control of ethical design and focus on eliminating biases and promoting fairness and accountability in dialogue conversations with users [9, 58]. Previous works employ ethical design principles and strategies to study how human perception is affected by dialogue conversation [188] and language style [181] have on user perception. These considerations have engendered efforts in creating datasets and building frameworks that address some important ethical issues such as social bias [220], and morality [172, 224].

Works in transparency, another key facet in governance, have mainly covered algorithmic techniques such as explainability and interpretability for the purposes of building users' trust in dialog systems. The Algorithmic Transparency Framework [59] aims to leverage "black box" models to provide users context that influences verifiable human reasoning toward their decision-making processes. ChatrEx [74], demonstrates the improvement of transparency and trust in chatbot interfaces using explainable techniques. Authors in Reference [200] empower humans to have more control in the inspection, chaining, and modification of prompts in LLMs for improving the quality of dialog-related tasks. Other works develop frameworks for developers that establish principles in designing conversational agents to promote dialog system transparency [186].

Privacy in data management and human-chatbot conversations is important to ensure governance in dialog system design. This has led to the investigation of privacy issues such as data leakage in conversations [85, 91] and data storage [56]. Authors in Reference [207] develop a detection scheme that resolves such privacy concerns and ensures the security of users' data in dialog interaction. Authors in Reference [175] create, BETOLD, a privacy-preserving dataset for conversational breakdown detection, and propose a detection approach for potential breakdowns. The work in Reference [10] investigates users' privacy concerns and proposes a list of features that are important to build trust in services mediated by chatbots such as decoupling the individuals that disclose their data and improving chatbot response quality and grammatical correctness. Similarly, the work in Reference [38] proposes a dialog system architecture that is capable of preserving the privacy of the user by leveraging argumentation as a framework for non-monotonic reasoning and explainability. It also uses the latest European data protection regulations to implement data minimization, purpose limitation, and storage limitation principles.

3.6 HCDS Advancements in Large Language Models

LLM-based chatbots today have become an integral part in current HCDS systems due to its emerging capabilities and adaptations over dialog tasks. In this section, we detail recent advancements

and the impact of LLMs within the scope HCC, HCA, and **Human-Centered Chatbot Design and Governance** (**HCDG**).

3.6.1 LLMs for Human-Chatbot Collaboration. HCDS systems often rely on the performance of dialog system architectures to foster collaboration and teaming with humans for creating synergistic outcomes. For instance, the work in Reference [37] demonstrates that providing human-in-the-loop iterative feedback improves the unquantified distributions of text generated in LLM-based dialogs for better capturing human preferences and personalization. Authors in Reference [168] suggest the increase in relevancy matching performance of recommender dialogs are often attributed to user feedback in the form of prompt engineering (i.e., reprompting). Multiple studies have even utilized knowledge acquisition strategies such as active learning and interactive task learning and have shown that LLMs are suitable for enabling human-AI teaming when training in low data regimes [50, 87, 138, 153]. Apart from training, ChatFive [88] demonstrates the ability of LLMs to better assess user personality traits compared with traditional Likert-scale tests through live conversations on a user interface. Similarly, the work in Reference [15] allows humans to revise the reasoning process of planning LLMs via flowcharts on a user interface. In addition, other previous works leverage the context understanding capabilities of LLMs to develop methods that ask clarifying user questions during conversations [24, 128, 178].

3.6.2 LLMs for Human-Chatbot Alignment. Recent HCDS studies leverage the contextual understanding and adaptation abilities of LLMs to capture human emotion and optimize alignment objectives with human values and preferences. The work in Reference [52] develops an alignment scheme that imitates the good and bad behavior of humans. Compared with RLHF which enables holistic human feedback, FIGA implements a fine-grained quality reward signal to allow the LLM to better understand human behaviors. This has engendered recent evaluation benchmarks namely Chatbot Arena [28], to assess the performance of LLMs for aligning to human preferences. In the scope of empathetic LLMs, empirical studies such as in References [57, 115, 146] suggest zero and few-shot LLMs are able to align with human emotion in conversational settings. However, the performance is increased using fine-tuning techniques for existing dialogue datasets [26]. Furthermore, aligning dialog models by increasing algorithmic transparency is often challenged due to large-scale "black-box" models [217]. Despite this, recent work has used HCA techniques LLMs to provide transparent explanations to better align with the users. For example, LLMCheckup [192] aims to further improve the users' understanding of dialog predictions by integrating a set of black-box and white-box models and provides explanations over a conversational user interface.

3.6.3 LLMs for Human-Centered Chatbot Design & Governance. LLMs have also significantly shaped the design methodologies and developmental frameworks employed in modern HCDS. User-centric approaches in HCDS often identify the needs of target users to develop a dialog system that serves those users on an application use-case. For example, existing studies design user-centric frameworks that entail human-centered techniques in various parts of the LLM development life cycle which include the data management, model prompting and training, and application utility [2, 135]. Authors in Reference [139] develop DialogBench, an evaluation benchmark that tests the human-likeness of LLMs for multi-turn dialogue. The results suggest that while instruction tuning helps LLMs to be more user-centric by establishing personable connections with users, LLMs still struggle to perceive emotions and user personalities. Moreover, additional evaluation studies are performed to assess whether LLMs reach the satisfactory performance and compatibility of target users pertaining to various demographics and task-specific needs [162, 191]. In the scope of governance, substantial efforts in LLM developmental practices aim at addressing challenges

Table 2. Collection of Peer-Reviewed Papers that belong to a Specific Task (Rows)

Specified Task	Human-Chatbot Collaboration	Human-Chatbot Alignment	Human-Centered Chatbot Design & Governance
Text Summarization	Interactive DS [46]; Iterative Feedback [22]	Human Preference Learning [46, 119, 141, 171]	HE/UX [119]
Dialog Generation	Interactive DS [54, 70]; Iterative Feedback [6, 67, 93, 105, 170]; Learning from Demonstrations [13, 48, 49, 100, 106, 190]	Human Preference Learning [8, 76, 141, 213]; Empathetic Dialogs [69, 95, 96, 103, 104, 118, 210]; Explainable Dialogs [62, 133, 193, 196, 219]	User-Centric Design [7, 39, 42, 78, 109, 114, 194, 214, 216]; HE/UX [41, 92, 107, 134, 173, 183]; Governance [181, 200]
Classification/Natural Language Understanding	Learning from Demonstrations [23, 44, 122]	Human Preference Learning [76, 141]; Empathetic Dialogs [69, 96, 103, 118]; Explainable Dialogs [98, 221]	User-Centric Design [39, 53, 69]; HE/UX [173, 183]; Governance [38, 59, 91]
Machine Translation	Iterative Feedback [79]; Learning from Demonstrations [13]	Human Preference Learning [81, 141]	HE/UX [92]; Governance [200]
Question-Answering	Interactive Dialogs [61, 94, 120, 187, 205]; Learning from Demonstrations [13, 122]	Human Preference Learning [76, 110, 141]; Explainable Dialogs [193, 196, 212, 219]	User-Centric Design [53]; HE/UX [41, 137]
Others (e.g., Information Retrieval, Recommender System, Natural Language Interface)	Interactive Dialogs [204]; Learning from Demonstrations [122, 160]	Human Preference Learning [45, 110, 141]; Explainable Dialogs [74, 126, 212]	HE/UX [82, 136, 163]; Governance [207]

We categorize papers based on the HCDS taxonomy category (columns). We then further categorize each paper based on their respective HCDS taxonomy subtopic.

related to privacy [3, 63, 209]. These works address the vulnerability of LLMs in leaking confidential information (i.e., prompt leaks) through various attack strategies and defense mechanisms.

4 Human-Centered Dialog System Applications

In this section, we discuss the applications of HCDS. Similar to traditional dialog systems, HCDS tasks typically depend on what the chatbot being built is purposed for (e.g., task-oriented, open-domain). Based on the aforementioned literature, the chatbot type includes NLP-based tasks that can vary for each application setting. Specifically, we detail the related tasks for task-oriented and open-domain task scenarios shown in Table 2.

4.1 Text Summarization

Text summarization is a NLP task that involves condensing a given document or body of text while retaining its essential information. Reference [22] proposes a human-in-the-loop i.e., *Human-Chatbot Collaboration* conversation summarization model trained on the DialogSum [25] dataset to improve coherence, faithfulness, and overall quality of text summaries generated by chatbots. Concepts such as active learning and reinforcement learning are employed in making effective text summarization models, with active learning helping dialog systems by selecting diverse examples for human annotation, and reinforcement learning enhancing models by training them to generate optimal summaries through interactions with the environment [46]. In the scope of

HCA, notable HCDS models such as ChatGPT [1] and LLaMA 2 [179] have been widely used for text summarization tasks using RLHF produce human desired summaries. The study in Reference [119] focuses on improving the *Design & Governance* of HCDSs in open-ended tasks, observing significant instances of hallucinated content in model-generated summaries and confirming that pre-trained models produced more faithful and factual summaries through human evaluation.

4.2 Dialog Generation

Dialog generation is the process of creating coherent and contextually relevant responses in a conversational setting, allowing for effective communication and interaction with users, thereby improving HCI. Works such as Reference [54] improve performance on datasets like PersonaChat [214] through self-feeding chatbots capable of extracting new training examples from conversations. Approaches like [6] generate interesting and semantically meaningful responses through online human-in-the-loop active learning and offline two-phase supervised learning. Pipeline-based architectures employed in Dialog Generation, as in References [49] and [190], train dialog systems to learn dialog policies with expert examples via RL, policy shaping, and reward shaping for domains like taxi reservation, restaurant reservation, and movie ticket booking. Popular open-domain chatbot dialogs such as ChatGPT [1], LLaMA 2 [179], Gemini [174], and Claude [5] are developed using RLHF to provide appropriate and human-aligned responses in various conversational settings. Empathetic dialog generation is vital for human-centered systems, ensuring responses reflect understanding and sensitivity, and fostering natural and effective interactions that enhance user satisfaction and engagement, as addressed in works like [95] which ensures empathetic dialog generation using *HCA* via a learning framework involving interactive adversarial methods that leverage user feedback to determine the level of emotional awareness in generated responses.

4.3 Intent Classification and Natural Language Understanding

Intent classification, or NLU, is crucial in NLP for deciphering the user's intent behind a given natural language input. In the realm of *Human-Chatbot Collaboration*, works like [23] and [44] focus on enhancing NLU through self-supervision and refining language models with limited annotated examples. The *HCA* category encompasses contributions like [69, 76, 96, 98, 103, 118, 141, 221], which underscore the importance of tuning chatbot behavior to align with human expectations and social norms, such as modeling the underlying reasoning process of neural models as latent variables [221] and using deep RL to craft emotionally resonant responses [69]. Lastly, the *Design & Governance* category, with works like [59] and [38], highlights the broader ethical, privacy, and governance considerations in chatbot development, including enhancing intent classification explicability through integration with semantic knowledge graphs and designing task-oriented chatbots with ethical considerations in mind.

4.4 Machine Translation

Machine translation, focused on automatically translating text or speech between languages while preserving meaning and context, is a critical NLP task for bridging language barriers. In the *Human-Chatbot Collaboration* sphere, [13] introduces GPT-3, a massive 175 billion parameter autoregressive language model showcasing prowess in open-domain NLP tasks like machine translation, enhancing human-chatbot interaction, while [79] demonstrates using Reinforcement Learning with minimal human feedback for tasks like German-to-English translation, underscoring adaptive, learning-oriented chatbots. Regarding *HCA*, [81] ventures into enhancing neural machine translation with human reinforcement using explicit and implicit user feedback, aligning outputs closer to human expectations, complemented by [141]'s training frameworks harmonizing chatbot behaviors with user intentions. In the *Design & Governance* context, [92] addresses diversifying machine translation

outputs to align with the multifaceted nature of human language, while [200] empowers users to inspect, chain, and modify prompts in LLMs, elevating transparency, controllability, and trust in AI interactions, aligning chatbot design and governance with ethical standards.

4.5 Question-Answering

Question-answering, an NLP task involving the synthesis of contextual responses to queries by understanding user input and retrieving/summarizing information from diverse knowledge sources, has seen significant advancements in the realm of *Human-Chatbot Collaboration* through works like [13]'s GPT-3, which enhanced NLP tasks like question-answering with its vast scale, Reference [120] on lifelong learning for continuous chatbot improvement, Reference [61] on refining interaction quality by clarifying ambiguous questions, References [205] and [122] presumably contributing to chatbot adaptivity and learning, and Reference [53] addressing context management in task-oriented dialogues. Regarding *HCA*, Reference [212]'s ReAct enhances open-domain dialog systems with interleaved reasoning traces and task specifications, improving performance and human interpretability. In *Design & Governance*, Reference [41] introduces a framework for qualitatively analyzing chatbot dialogues to refine customer service, Reference [137] proposes an interface for efficient literature search in medical archives, highlighting chatbot versatility and the importance of thoughtful design and governance for reliable, ethical, and effective chatbot technologies. Similarly, MediTron [27] ensures safe and ethical responses related to medical literature for numerous users.

4.6 Other

Other applications of human-centered dialog systems in task-based domains include [74]'s ChatrEx, providing visual explanations for complex spreadsheet tasks, Reference [133] augmenting task-oriented dialogs with explanations to foster human-computer trust in incomprehensible situations, and Reference [136]'s KnowCOVID-19 evidence-based recommender system aiding scientific knowledge discovery. Regarding open-domain dialog systems, *Governance* is employed in Reference [207] to ensure data privacy by warning users of suspicious sentences, while *User-Centric Design* is addressed in Reference [204] by improving internet-retrieval models through incorporating human feedback during deployment, enabling chatbots to adapt to new information and improve concurrently. Other applications such as code generation, which employ widely used LLMs (e.g., StarCoder [97], Code LLaMA [154]), use *Governance* to ensure safety and privacy of deployed systems.

5 Research Challenges

In this section, we discuss research challenges from a practical design and socio-technical perspective. We discuss such topics on challenges related to bias, restricted scalability, data security, privacy and trust, lack of suitable metrics, and lack of standard ethical guidelines. In Table 3, we detail each of the research challenges under the HCDS taxonomy and highlight/checkmark areas within the taxonomy where such research challenges can arise.

5.1 Bias

In the development of HCDS, bias stands out as a critical research challenge, an issue underscored by key research in the field. In the scope of human-chatbot collaboration, CAiRE [104] notes the scarcity and imbalance of data in crucial areas such as emotion recognition and empathetic response generation. This lack of a diverse dataset that reflects the target population of human collaborators hinders the chatbot's ability to accurately perceive and interpret human emotions, a fundamental requirement for any human-centric interaction. Consequently, the sparsity of data leads to a practical challenge that not only impacts the system's current capabilities but allows

Table 3. Research Challenges that Relate to Human-Chatbot Collaboration, Human-Chatbot Alignment, and Design & Governance for the Development of Human-Centered Dialog System Applications

	Human-Centered Dialog System Taxonomy		
Research Challenge	**Human-Chatbot Collaboration**	**Human-Chatbot Alignment**	**Design & Governance**
Bias	Data imbalance and data scarcity [104]	–	Escalation of errors [106]
Restricted Scalability	Expensive human annotations [54, 169]	Limited diversity judgments and unscalable human feedback [18, 93]	–
Data Security, Privacy & Trust	–	–	Privacy leakage and user trust [17, 19, 68, 113]; insufficient anthropomorphism [130]
Lack of Suitable Metrics	–	Undermining chatbot limitations [30, 111, 129, 164]	Defining appropriate evaluation metrics [132]
Lack of Standard Ethical Guidelines	Defining roles for responsibility and accountability [16]	Personalizing chatbots over diverse perspectives and morals [141, 155]	Lack of uniformity of governance practices [89]

We briefly describe the types of pitfalls and potential risks that are exemplified within the HCDS taxonomy.

the dialog to perpetuate stereotypes against target sub-populations [220]. From a socio-technical perspective, this also limits the potential for developing new learning techniques and evolving in a way that is truly empathetic and nuanced in human interactions. These combined factors underscore the need for more robust and diverse datasets, as well as enhanced error tracking and mitigation strategies in the development of dialogue systems that can genuinely understand and empathize with human users. Bias can also arise in HCDG, particularly in the development life cycle. Authors in Reference [106] highlight a particularly concerning aspect where errors, once introduced into the system, are not simply passed along the pipeline but also amplified. This escalation of errors significantly complicates the process of tracing back to their origins, thereby impeding effective rectification and refinement of the system. If not caught or mitigated by developers, this can cause a system to lack transparency and truthfulness in the deployment of HCDS.

5.2 Restricted Scalability

Scaling up HCDSs to manage large datasets or accommodate numerous users presents a significant challenge, primarily due to the inherent limitations in the system's ability to process data swiftly and efficiently owing to the reliance on human experts or operators. In the context of human-chatbot collaboration, interactive dialog systems require humans to provide feedback continuously, which is time-consuming for users to provide their inputs on chatbot response [54, 169]. From a practical sense, these constraints could render such systems impractical for real-world applications. For example, in task-oriented settings, the absence of expert annotators and datasets for models to improve training could both hinder the reproducibility and deployment performance of dialog systems [48, 49, 79, 94]. In more recent advancements, restricted scalability can arise in HCA, specifically when failing to acquire a diverse set of human judges to give their preference for the dialog system in effective learning human values [18]. The scalability of HCDSs is significantly challenged by the human-in-the-loop training process [93, 141], as well as by defining values that involve consensus among various stakeholders [8]. This approach requires a large number of human evaluators to provide feedback, which increases resource demands and poses consistency issues. As

the system expands, the need for human input creates a bottleneck, limiting the rapid development and widespread deployment of these advanced dialogue systems. This highlights the necessity for more efficient training methods that balance automation with empathetic understanding.

5.3 Data Security, Privacy, and Trust

When it comes to the personalization of dialog systems in human-centered chatbot design and governance, a large amount of data is required even for every single user. This brings in concerns regarding trust and privacy. Authors in Reference [17] demonstrated how a potential attacker may retrieve particular training examples from pre-trained language models using an extraction technique. Sensitive data such as names, phone numbers, emails, and even deleted material may be present in these samples. This raises significant privacy concerns, especially in the context of user-focused pre-training, where models could inadvertently retain and expose training data, potentially making it vulnerable to malicious use. Although progress has been made for chatbots to interact with users more easily, trust toward chatbots remains a strong social challenge that hinders the technology from being further diffused into society [113]. In addition, distrust toward chatbots becomes more salient when it comes down to people's concerns over privacy [130]. Other factors that influence people's trust toward chatbots can fall under the category of insufficient anthropomorphism which is often shown via chatbots' lack of language abilities [68], intelligence for solving problems and communicating with humans, emotions, and personalities that necessitate relationship building with users [19].

5.4 Lack of Suitable Metrics

The suitability of metrics used for a dialog system outcome is dependent on the application scenario. In the scope of human-centered alignment, it has been found that a chatbot's ability to pose as a human and reflect their preferences and values does lead to more positive assessments by users overall [129]. Furthermore, users are more likely to make an effort to be understood when they believe the chatbot is more human-like [30]. However, this does present the risk of users underestimating the limitations of a chatbot [111]. Flaws such as a lack of social smarts, odd responses, and slow response times are viewed as strange by users [164]. For human-centered chatbot design and governance, users might not often expect personalization in their responses in task-oriented dialog systems. Whereas, in open-domain systems, understanding the user's queries properly and profiling the user is important to generate user-specific responses and increase user engagement. There is a need for an evaluation process that explores the dimension users expect to encounter and determines the objectives they have for the dialog system's responses. Currently, there is an open challenge in identifying and utilizing the appropriate set of metrics for different stakeholders [86, 137, 144]. Since no metric is ubiquitous, identifying the appropriate metric is crucial for getting useful feedback for improving HCDS and assessing widespread user adoption in real-world applications. To build trustworthy and reliable dialog systems, a metric to assess the reliability and explainability of the user-centric responses is necessary [132]. This is different from data-driven systems since they focus on "learning" the training data and "mimicking" the responses. Defining appropriate evaluation metrics for HCDSs is still an evolving area, and existing benchmarks may not fully capture the nuances of natural, interactive conversations.

5.5 Lack of Standard Ethical Guidelines

Although previous studies have investigated governance techniques for ensuring safe and ethical practices of HCDS design, there is a lack of standard guidelines and uniform practices to be adopted by practitioners and organizations. In terms of human-chatbot collaboration, a lack of establishing roles for co-supervision (e.g., accountability, individual responsibility) between the human and

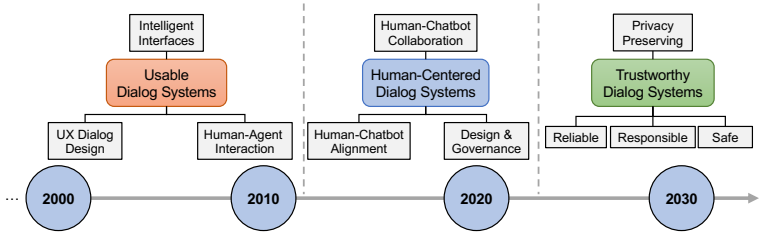

Fig. 10. Past and future phases of dialog system development based on existing literature. The past phases span *Usable Dialog Systems* and recently *Human-Centered Dialog Systems*. The future phase (i.e., 2024 and beyond) encompasses *Trustworthy Dialog Systems* involves important facets including reliable, responsible, safe, and privacy-preserving.

chatbot agent can cause divergence in the practical and socio-technical development of dialog systems [16]. Thus, roles must be established between chatbots and humans to create synergy in performing well on various task scenarios. In addition, collaborations between dialog systems and humans for the augmentation of humans often present tradeoff challenges between machine automation and human control during HCDS training processes. Thus, these facets of HCDS development must be taken into account when considering human stakeholders (e.g., individuals, communities, society) and application needs [161]. In the context of human-chatbot alignment, making personalized chatbots that align more with human values can spark many ethical issues such as the legal responsibilities of AI, users being misled or even mentally manipulated, and reducing the meanings of users' social reality with real people [155]. Furthermore, human alignment techniques such as RLHF can lead to conflict and misinterpretation from the dialog system during alignment training when there is a disagreement of values among a large group of human labelers [141]. Lastly, the adoption of such practices and standards in the scope of governance across organizations is scarce [89], and the establishment of ethical design practices is isolated and lacks uniformity. Such research challenges can arise in governance when audits on data management and algorithms are performed internally in an organization. This can lead to a loss of truthfulness and transparency in HCDS design if organizations do not integrate ethical standards properly.

6 Future Research Directions

Future research directions in HCDSs are poised to prioritize the development of trustworthy AI. In Figure 10, we show the future trends in dialog system development that show the upcoming stage, namely, *Trustworthy Dialog Systems*. This entails the development and deployment of reliable, responsible, safe, and privacy-preserving dialog systems. In the following, we detail some promising future directions related to *reward model design*, *safety evaluation*, and *auditing* from a HCDS perspective.

6.1 Reward Model Design

Modeling user feedback and preferences using reward-based techniques is not always straightforward or accurate. Poor reward proxies for modeling user input can happen when they fail to capture the full spectrum of actual rewards, causing *misspecification* [143]. In addition, these inaccurate rewards can cause *goal misgeneralization* [34] when faced with distribution shifts. Both issues can lead to what is known as *reward hacking* [43], where the agent exploits a reward system to gain high rewards without optimizing over human feedback. As more complex HCDS continue to integrate into the daily activities of human society, bad outcomes can emerge if agents are not properly aligned with human values [18].

Resolving these challenges entails the redesign of reward systems to not reduce misalignment from the dialog models, but also capture the values of diverse human judges to mitigate algorithmic biases. An example study in Reference [8] addresses the issue of heterogeneous human judgments, particularly when there is a lack of diversity in human feedback. In their solution approach, authors consider heterogeneity in human preference modeling and address the diversity in human judgments by employing a pre-trained LLM to generate the expected outcomes that form consensus among human groups to avoid reward misspecification and goal misgeneralization. Human judges then rate the agreement and quality of these generated statements, and then a reward model is built to predict humans' individual preferences. Other reward redesign strategies include eliminating holistic and ambiguous rewards provided by humans. This has been exemplified in Reference [203], where they address the challenge of coarse-grained feedback by proposing a fine-grained RLHF framework in LLMs. Authors in this work show that incorporating human feedback on individual sentences as well as multiple reward models with different feedback types (e.g., factual correctness, irrelevance, information completeness) improves LLM performance via automated and human evaluation. Furthermore, authors in Reference [84] show that modeling a reward system based on user feedback can be expressed through explicit prompts of the desired behavior or preference that are fed into LLMs, which can serve as a "cheap" proxy for a reward function. As a result, these developments can impact current HCA technologies by mitigating algorithm biases and misalignment related challenges in HCDS to ensure more trustworthy dialog systems.

6.2 Safety Evaluation

Building HCDS that are safeguarded from producing harmful and offensive content is crucial for widespread adoption once deployed in real-world applications. Avoiding such practices or not properly assessing the amount of toxicity in chatbots can, and has resulted in negative consequences and experiences for companies and consumers, respectively. Safety evaluation [32] of HCDS can help ensure they are critically evaluated before being deployed. This can come in various forms such as inspecting the model inputs and outputs or the dataset that can cause biases and/or harmful responses from the model.

According to the survey in Reference [32] that addresses the safety and trust concerns of LLM-based chatbots, two core safety evaluation approaches can be employed: First, developers and practitioners must inspect the model and the input data before generation. This is done by computing the probability of the given data input samples which helps check the amount of bias within a model. Second, practitioners must inspect the model after generation. This can be done by inputting context into the model that could trigger an unsafe response and measuring the safety by the success of that response based on the input context. Safety evaluations measures can also be extended to measuring stereotypical biases in dialog architectures. Authors in Reference [159] implemented this approach in pre-trained language models by diversifying their prompts based on human demographics. The results showed the amount of bias that pre-trained models inherited when calculating sentiment scores were based on the generated outputs of prompts from different groups. Other works such as StereoSet [131] and [140] suggested suitable metrics and datasets to help measure the amount of bias and toxicity that models exhibit during generation, and in pre-training-like objectives, respectively. Implementing such practices can ensure more responsible model deployment and enhance user trust of dialogs in practical applications.

6.3 Auditing

Auditing efforts on HCDS systems have been put in place through internal company practices and external entities. There are two types of audits: algorithmic [77, 148] and ethics-based [124]. The work in Reference [77] surveys the key areas necessary to perform auditing and assurance in

algorithms. Authors in Reference [148] introduce internal algorithmic audits as a way to verify that the engineering procedures used in the development and implementation of AI systems adhere to established ethical norms and expectations, such as organizational AI principles. The work in Reference [124] states that ethics-based auditing must be a constructive, ongoing process that approaches ethical alignment from an organizational perspective, and it must be in line with public laws and incentives for morally righteous behavior in order to be practical and successful.

Based on this, various auditing approaches can be enacted in the development life cycle of dialog systems to improve transparency and standard ethical practices. For example, the work in Reference [125] proposes a multi-layered approach for auditing an AI governance framework for LLMs. This framework comprises of three layers that complement and inform each other: (i) audits performed by the technology providers that develop and deploy LLMs among practitioners and communities, (ii) audits performed after the LLM has been pre-trained but before its release, and (iii) audits of the applications the LLMs are applied to. Authors in Reference [71] perform auditing in dialogs by casting them into LLMs as an optimization that eliminates derogatory and nonfactual completions on famous celebrities. The work proposes **Autoregressive Randomized Coordinate Ascent** (**ARCA**), a discrete optimization algorithm that jointly optimizes over token inputs and generated text outputs. In addition, previous studies such in Reference [167] have demonstrated the ways to measure unethical development and training. Specifically, authors developed auditing tools to detect the misuse of the training on unauthorized user data. In the scope of governance, these tools can lead to more ethical developmental practices for building trustworthy HCDS.

7 Conclusion

In conclusion, we presented a holistic overview on HCDS through a literature survey. In our work, we aimed to bridge the gap between ML-based dialog systems and HCAI approaches. We developed a taxonomy that categorizes HCDS under the sub-groups of Human-Chatbot Collaboration, HCA, and HCDG. We described the recent advancements in each sub-group and discussed notable benchmark datasets, application scenarios, and downstream NLP-tasks. We presented common research challenges and identified that challenges such as bias, ethical practices, and data security/privacy/trust present a major role socially for end-users just as much as it does technically for practitioners. We observed that while fundamental issues in dialog systems such as misalignment, safety, and lack of auditing present a serious risk to society, emerging works offer promising research directions toward building a more safe, ethical, and trustworthy dialog system. We aim at informing researchers and practitioners about the salient gaps within HCDS to yield breakthrough research discoveries and create momentum in fostering the next generation of dialog systems and their applications.

Acknowledgments

We thank the following students who have contributed in various parts of this project: Eric Milman and Chaitra Kulkarni.

References

[1] Josh Achiam, Steven Adler, et al. 2023. Gpt-4 technical report. (2023), 1–100.

[2] Anum Afzal, Alexander Kowsik, Rajna Fani, and Florian Matthes. 2024. Towards optimizing and evaluating a retrieval augmented QA chatbot using LLMs with human-in-the-loop. In *Proceedings of the 5th Workshop on Data Science with Human-in-the-Loop (DaSH'24)*. 4–16.

[3] Divyansh Agarwal, Alexander Fabbri, Ben Risher, Philippe Laban, Shafiq Joty, and Chien-Sheng Wu. 2024. Prompt leakage effect and mitigation strategies for multi-turn LLM applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 1255–1275.

[4] Saleema Amershi, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. ModelTracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 337–346.

[5] Anthropic. 2023. *Model Card and Evaluations for Claude Models*.

[6] Nabiha Asghar, Pascal Poupart, Xin Jiang, and Hang Li. 2017. Deep active learning for dialogue generation. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*. 78–83.

[7] Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective neural response generation. In *Proccedings of the 40th European Conference on IR Research*. 154–166.

[8] Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. In *Proceedings of the 36th Conference on Neural Information Processing Systems*. 38176–38189.

[9] Junseong Bang, Sineae Kim, Jang Won Nam, and Dong-Geun Yang. 2021. Ethical chatbot design for reducing negative effects of biased data and unethical conversations. In *Proceedings of the 2021 International Conference on Platform Technology and Service*. 1–5.

[10] Rahime Belen Saglam, Jason R. C. Nurse, and Duncan Hodges. 2021. Privacy concerns in Chatbot interactions: When to trust and when to worry. In *Proceedings of the 23rd HCI International Conference*. 391–399.

[11] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management. (2017), 1–9.

[12] Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *Proceedings of the 5th International Conference on Learning Representations*. 1–15.

[13] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th Conference on Neural Information Processing Systems*. 1877–1901.

[14] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 5016–5026.

[15] Yuzhe Cai, Shaoguang Mao, Wenshan Wu, Zehua Wang, Yaobo Liang, Tao Ge, Chenfei Wu, WangYou WangYou, Ting Song, Yan Xia, Nan Duan, and Furu Wei. 2024. Low-code LLM: Graphical user interface over large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 12–25.

[16] José J Cañas. 2022. AI and Ethics when human beings collaborate with AI agents. *Frontiers in Psychology* 13 (2022), 1–9.

[17] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium*. 2633–2650.

[18] Stephen Casper, Xander Davies, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research* (2023), 1–42.

[19] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758.

[20] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter* 19, 2 (2017), 25–35.

[21] Hugh Chen, Scott Lundberg, et al. 2021. Explaining models by propagating Shapley values of local components. *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability* 914 (2021), 261–270.

[22] Jiaao Chen, Mohan Dodda, and Diyi Yang. 2023. Human-in-the-loop abstractive dialogue summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 9176–9190.

[23] Mingda Chen, Jingfei Du, et al. 2022. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3558–3573.

[24] Yue Chen, Chen Huang, et al. 2024. STYLE: Improving domain transferability of asking clarification questions in large language model powered conversational agents. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024*. 10633–10649.

[25] Yulong Chen, Yang Liu, et al. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. 5062–5074.

[26] Yirong Chen, Xiaofen Xing, et al. 2023. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023*. 1170–1183.

[27] Zeming Chen, Alejandro Hernández Cano, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. (2023), 1–38.

[28] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*. 1–30.

[29] Kyunghyun Cho, Bart van Merriënboer, et al. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 1724–1734.

[30] Kevin Corti and Alex Gillespie. 2016. Co-constructing intersubjectivity with artificial conversational agents: People are more likely to initiate repairs of misunderstandings with agents represented as human. *Computers in Human Behavior* 58 (2016), 431–442.

[31] Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*. 76–87.

[32] Jiawen Deng, Hao Sun, and Zothers. 2023. Recent advances towards safe, responsible, and moral dialogue systems: A survey. (2023), 1–18.

[33] Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2017. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 484–495.

[34] Lauro Langosco Di Langosco, Jack Koch, et al. 2022. Goal misgeneralization in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*. 12004–12019.

[35] Luis Fernando D'Haro, Koichiro Yoshino, et al. 2020. Overview of the seventh dialog system technology challenge: DSTC7. *Computer Speech & Language* 62 (2020), 1–21.

[36] Mihail Eric, Lakshmi Krishnan, et al. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 37–49.

[37] Xiang Fan, Yiwei Lyu, et al. 2023. Nano: Nested human-in-the-loop reward learning for few-shot language model control. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023*. 11970–11992.

[38] Bettina Fazzinga, Andrea Galassi, et al. 2022. A privacy-preserving dialogue system based on argumentation. *Intelligent Systems with Applications* 16 (2022), 1–17.

[39] Mauajama Firdaus, Naveen Thangavelu, et al. 2023. I enjoy writing and playing, do you?: A personalized and emotion grounded dialogue agent using generative adversarial network. *IEEE Transactions on Affective Computing* 14, 3 (2023), 2127–2138.

[40] Asbjørn Følstad and Marita Skjuve. 2019. Chatbots for customer service: User experience and motivation. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. 1–9.

[41] Asbjørn Følstad and Cameron Taylor. 2021. Investigating the user experience of customer service chatbot interaction: A framework for qualitative analysis of chatbot dialogues. *Quality and User Experience* 6, 1 (2021), 1–17.

[42] Liye Fu, Susan R. Fussell, et al. 2020. Facilitating the communication of politeness through fine-grained paraphrasing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Nov. 2020), 5127–5140.

[43] Leo Gao, John Schulman, et al. 2023. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 10835–10866.

[44] Tianyu Gao, Adam Fisch, et al. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 3816–3830.

[45] Xiang Gao, Yizhe Zhang, et al. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 386–395.

[46] Yang Gao, Christian M. Meyer, et al. 2020. Preference-based interactive multi-document summarisation. *Inf. Retr.* 23, 6 (Dec 2020), 555–585.

[47] Asma Ghandeharioun, Judy Hanwen Shen, et al. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. *Advances in Neural Information Processing Systems* 32 (2019), 13665–13676.

[48] Gabriel Gordon-Hall, Philip John Gorinski, et al. 2020. Learning dialog policies from weak demonstrations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1394–1405.

[49] Gabriel Gordon-Hall, Philip John Gorinski, et al. 2020. Show us the way: Learning to manage dialog from demonstrations. *In Proceedings of the 8th Dialog System Technology Challenge (DSTC-8) at AAAI 2020* (2020), 1–10.

[50] Sai Keerthana Goruganthu, Roland R. Oruche, et al. 2024. Adaptive open-set active learning with distance-based out-of-distribution detection for robust task-oriented dialog system. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 357–369.

[51] Jia-Chen Gu, Tianda Li, et al. 2020. Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. ACM, 2041–2044. arXiv:2004.03588

[52] Geyang Guo, Ranchi Zhao, et al. 2024. Beyond imitation: Leveraging fine-grained quality signals for alignment. In *Proceedings of the Twelfth International Conference on Learning Representations*. 1–16.

[53] Arshit Gupta, Peng Zhang, et al. 2019. CASA-NLU: Context-aware self-attentive natural language understanding for task-oriented chatbots. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 1285–1290.

[54] Braden Hancock, Antoine Bordes, et al. 2019. Learning from dialogue after deployment: Feed yourself, chatbot!. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3667–3684.

[55] Tom Hanika, Marek Herde, et al. 2019. Collaborative interactive learning–a clarification of terms and a differentiation from other research fields. arXiv:1905.07264. Retrieved from https://arxiv.org/abs/1905.07264

[56] Martin Hasal, Jana Nowaková, et al. 2021. Chatbots: Security, privacy, data protection, and social aspects. *Concurrency and Computation: Practice and Experience* 33, 19 (2021), e6426.

[57] Md Rakibul Hasan, Md Zakir Hossain, et al. 2024. LLM-GEm: Large language model-guided prediction of people's empathy levels towards newspaper article. In *Proceedings of the Findings of the Association for Computational Linguistics: EACL 2024*. 2215–2231.

[58] Peter Henderson, Koustuv Sinha, et al. 2018. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 123–129.

[59] Sam Hepenstal, Neesha Kodagoda, et al. 2019. Algorithmic transparency of conversational agents. In *Proceedings of the 2019 Workshop on Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies*.

[60] Karl Moritz Hermann, Tomas Kocisky, et al. 2015. Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems* 28 (2015), 1693–1701.

[61] Xiang Hu, Zujie Wen, et al. 2020. Interactive question clarification in dialogue via reinforcement learning. In *Proceedings of the 28th International Conference on Computational Linguistics*. 78–89.

[62] Xinting Huang, Jianzhong Qi, et al. 2020. Generalizable and explainable dialogue generation via explicit action learning. In *Proceedings of the Association for Computational Linguistics*. 3981–3991.

[63] Bo Hui, Haolin Yuan, et al. 2024. Pleak: Prompt leaking attacks against large language model applications. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 3600–3614.

[64] Ahmed Hussein, Mohamed Medhat Gaber, et al. 2017. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)* 50, 2 (2017), 1–35.

[65] Shankar Iyer, Nikhil Dandekar, et al. 2017. Quora Question Pairs.

[66] Mohit Jain, Pratyush Kumar, et al. 2018. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 895–906.

[67] Natasha Jaques, Judy Hanwen Shen, et al. 2020. Human-centric dialog training via offline reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 3985–4003.

[68] Liss Jenneboer, Carolina Herrando, et al. 2022. The impact of chatbots on customer loyalty: A systematic literature review. *Journal of Theoretical and Applied Electronic Commerce Research* 17, 1 (2022), 212–229.

[69] Jiun-Hao Jhan, Chao-Peng Liu, et al. 2021. Cheerbots: Chatbots toward empathy and emotionusing reinforcement learning. arXiv:2110.03949. Retrieved from https://arxiv.org/abs/2110.03949

[70] Megha Jhunjhunwala, Caleb Bryant, et al. 2020. Multi-action dialog policy learning with interactive human teaching. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 290–296.

[71] Erik Jones, Anca Dragan, et al. 2023. Automatically auditing large language models via discrete optimization. arXiv:2303.04381. Retrieved from https://arxiv.org/abs/2303.04381

[72] Chaitanya K. Joshi, Fei Mi, et al. 2017. Personalization in goal-oriented dialog. *31st Conference on Neural Information Processing Systems (NIPS 2017)* (2017), 1–15.

[73] Mandar Joshi, Eunsol Choi, et al. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1601–1611.

[74] Anjali Khurana, Parsa Alamzadeh, et al. 2021. ChatrEx: Designing explainable chatbot interfaces for enhancing usefulness, transparency, and trust. In *Proceedings of the 2021 IEEE Symposium on Visual Languages and Human-Centric Computing*. IEEE, 1–11.

[75] W. Bradley Knox and Peter Stone. 2008. Tamer: Training an agent manually via evaluative reinforcement. In *Proceedings of the 2008 7th IEEE International Conference on Development and Learning*. IEEE, 292–297.

[76] Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. 2023. Pretraining language models with human preferences. *International Conference on Machine Learning* abs/2302.08582 (2023), 17506 – 17533.

[77] Adriano Koshiyama, Emre Kazim, et al. 2022. Algorithm auditing: Managing the legal, ethical, and technological risks of artificial intelligence, machine learning, and associated algorithms. *Computer* 55, 4 (2022), 40–50.

[78] Satwik Kottur, Xiaoyu Wang, et al. 2017. Exploring personalized neural conversational models. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 3728–3734.

[79] Julia Kreutzer, Shahram Khadivi, et al. 2018. Can neural machine translation be improved with user feedback?. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 92–105.

[80] Julia Kreutzer, Stefan Riezler, et al. 2021. Offline reinforcement learning from human feedback in real-world sequence-to-sequence tasks. In *Proceedings of the 5th Workshop on Structured Prediction for NLP*. 37–43.

[81] Julia Kreutzer, Joshua Uyheng, et al. 2018. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. In *Proceedings of the Association for Computational Linguistics*. 1777–1788.

[82] Knut Kvale, Eleonora Freddi, et al. 2020. Understanding the user experience of customer service chatbots: What can we learn from customer satisfaction surveys?. In *Proceedings of the International Workshop on Chatbot Research and Design*. Springer, 205–218.

[83] Tom Kwiatkowski, Jennimaria Palomaki, et al. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics* 7 (2019), 452–466.

[84] Minae Kwon, Sang Michael Xie, et al. 2023. Reward design with language models. In *Proceedings of the 11th International Conference on Learning Representations*. 1–18.

[85] Martha Larson, Nelleke Oostdijk, et al. 2021. Not directly stated, not explicitly stored: Conversational agents and the privacy threat of implicit information. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 388–391.

[86] Alon Lavie and Michael J. Denkowski. 2009. The METEOR metric for automatic evaluation of machine translation. *Machine Translation* 23 (2009), 105–115.

[87] Lane Lawley and Christopher Maclellan. 2024. Val: Interactive task learning with gpt dialog parsing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[88] Jungjae Lee, Yubin Choi, et al. 2024. ChatFive: Enhancing user experience in likert scale personality test through interactive conversation with LLM agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. 1–8.

[89] Tom Lewandowski, Jasmin Delling, et al. 2021. State-of-the-art analysis of adopting AI-based conversational agents in organizations: A systematic literature review. *PACIS* (2021), 167.

[90] Patrick Lewis, Ethan Perez, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[91] Haoran Li, Yangqiu Song, et al. 2022. You don't know my favorite color: Preventing dialogue representations from revealing speakers' private personas. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 5858–5870.

[92] Jiwei Li, Michel Galley, et al. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 110–119.

[93] Jiwei Li, Alexander H. Miller, et al. 2017. Dialogue learning with human-in-the-loop. In *Proceedings of the 2017 International Conference on Learning Representations*. 1–23.

[94] Jiwei Li, Alexander H. Miller, et al. 2017. Learning through dialogue interactions by asking questions. In *Proceedings of the 2017 International Conference on Learning Representations*. 1–16.

[95] Qintong Li, Hongshen Chen, et al. 2020. EmpDG: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*. 4454–4466.

[96] Qintong Li, Piji Li, et al. 2022. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, Vol. 36. 10993–11001.

[97] Raymond Li, Loubna Ben allal, et al. 2023. StarCoder: May the source be with you! *Transactions on Machine Learning Research* (2023), 1–55.

[98] Shaobo Li, Chengjie Sun, et al. 2023. Toward explainable dialogue system using two-stage response generation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 22, 3, Article 68 (Mar 2023), 18 pages.

[99] Yanran Li, Hui Su, et al. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957* (2017), 1–10.

[100] Alexander Lin, Jeremy Wohlwend, et al. 2020. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 6121–6133.

[101] Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. 605–612.

[102] Stephanie Lin, Jacob Hilton, et al. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3214–3252.

[103] Zhaojiang Lin, Andrea Madotto, et al. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 121–132.

[104] Zhaojiang Lin, Peng Xu, et al. 2020. Caire: An end-to-end empathetic chatbot. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, Vol. 34. 13622–13623.

[105] Bing Liu and Ian Lane. 2018. End-to-end learning of task-oriented dialogs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 67–73.

[106] Bing Liu, Gokhan Tur, et al. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. (June 2018), 2060–2069.

[107] Chia-Wei Liu, Ryan Lowe, et al. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. (Nov. 2016), 2122–2132.

[108] Pengfei Liu, Weizhe Yuan, et al. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* 55, 9, Article 195 (Jan 2023), 35 pages.

[109] Qian Liu, Yihong Chen, et al. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1417–1427.

[110] Ruibo Liu, Ge Zhang, et al. 2022. Aligning generative language models with human values. In *Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics, 241–252.

[111] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA" The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5286–5297.

[112] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30 (2017), 4768–4777.

[113] Bei Luo, Raymond Y. K. Lau, et al. 2022. A critical review of state-of-the-art chatbot designs and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, 1 (2022), e1434.

[114] Liangchen Luo, Wenhao Huang, et al. 2019. Learning personalized end-to-end goal-oriented dialog. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, Vol. 33. 6794–6801.

[115] Man Luo, Christopher J. Warren, et al. 2024. Assessing empathy in large language models with real-world physician-patient interactions. In *Proceedings of the 2024 IEEE International Conference on Big Data (BigData'24)*. 6510–6519.

[116] Yukun Ma, Khanh Linh Nguyen, et al. 2020. A survey on empathetic dialogue systems. *Information Fusion* 64 (2020), 50–70.

[117] Andrew Maas, Raymond E. Daly, et al. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 142–150.

[118] Navonil Majumder, Pengfei Hong, et al. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 8968–8979.

[119] Joshua Maynez, Shashi Narayan, et al. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1906–1919.

[120] Sahisnu Mazumder, Bing Liu, et al. 2019. Lifelong and interactive learning of factual knowledge in dialogues. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, 21–31.

[121] Alexander Miller, Adam Fisch, et al. 2016. Key-value memory networks for directly reading documents. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Nov. 2016), 1400–1409.

[122] Sewon Min, Mike Lewis, et al. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2791–2809.

[123] Sewon Min, Xinxi Lyu, et al. 2022. Rethinking the role of demonstrations: What makes in-context learning work?. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 11048–11064.

[124] Jakob Mökander and Luciano Floridi. 2021. Ethics-based auditing to develop trustworthy AI. *Minds and Machines* 31, 2 (2021), 323–327.
[125] Jakob Mökander, Jonas Schuett, et al. 2023. Auditing large language models: A three-layered approach. *AI and Ethics* (2023), 1–31.
[126] Seungwhan Moon, Pararth Shah, et al. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 845–854.
[127] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, et al. 2022. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review* (2022), 1–50.
[128] Fangwen Mu, Lin Shi, et al. 2024. ClarifyGPT: A framework for enhancing LLM-based code generation via requirements clarification. *Proceedings of ACM Software Engineering* 1, FSE (2024), 1–23.
[129] Alessandro Murgia, Daan Janssens, et al. 2016. Among the machines: Human-bot interaction on social Q&A websites. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA'16)*. Association for Computing Machinery, 1272–1279.
[130] Morten Johan Mygland, Morten Schibbye, et al. 2021. Affordances in human-chatbot interaction: A review of the literature. In *Proceedings of the 20th IFIP WG 6.11 Conference on e-Business, e-Services and e-Society*. Springer, 3–17.
[131] Moin Nadeem, Anna Bethke, et al. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 5356–5371.
[132] Florian Nothdurft, Gregor Behnke, et al. 2015. The interplay of user-centered dialog systems and AI planning. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 344–353.
[133] Florian Nothdurft, Felix Richter, et al. 2014. Probabilistic human-computer trust handling. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 51–59.
[134] Shereen Oraby, Lena Reed, et al. 2018. Controlling personality-based stylistic variation with neural natural language generators. *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue* (July 2018), 180–190.
[135] Roland Oruche, Rithika Akula, et al. 2024. Holistic multi-layered system design for human-centered dialog systems. In *Proceedings of the 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS'24)*. 1–8.
[136] Roland Oruche, Vidya Gundlapalli, et al. 2021. Evidence-based recommender system for a covid-19 publication analytics service. *IEEE Access* 9 (2021), 79400–79415.
[137] Roland Oruche, Eric D. Milman, et al. 2021. Measurement of utility in user access of COVID-19 literature via AI-powered Chatbot. In *Proceedings of the IEEE Applied Imagery Pattern Recognition Workshop*. IEEE, 1–13.
[138] Roland Oruche, Marcos Zampieri, et al. 2024. Deep contrastive active learning for out-of-domain filtering in dialog systems. In *2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA'24)*. 1–10.
[139] Jiao Ou, Junda Lu, et al. 2024. DialogBench: Evaluating LLMs as human-like dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 6137–6170.
[140] Nedjma Ousidhoum, Xinran Zhao, et al. 2021. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 4262–4274.
[141] Long Ouyang, Jeffrey Wu, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
[142] Ozlem Ozmen Garibay, Brent Winslow, et al. 2023. Six human-centered artificial intelligence grand challenges. *International Journal of Human–Computer Interaction* 39, 3 (2023), 391–437.
[143] Alexander Pan, Kush Bhatia, et al. 2022. The effects of reward misspecification: Mapping and mitigating misaligned models. In *Proceedings of the 2022 International Conference on Learning Representations*. 1–19.
[144] Kishore Papineni, Salim Roukos, et al. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 311–318.
[145] Rob Procter, Peter Tolmie, et al. 2023. Holding AI to account: Challenges for the delivery of trustworthy AI in healthcare. *ACM Transactions on Computer-Human Interaction* 30, 2 (2023), 1–34.
[146] Yushan Qian, Weinan Zhang, et al. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023*. 6516–6528.
[147] Rafael Rafailov, Archit Sharma, et al. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the 37th Conference on Neural Information Processing Systems*.
[148] Inioluwa Deborah Raji, Andrew Smart, et al. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 33–44.

[149] Hannah Rashkin, Eric Michael Smith, et al. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5370–5381.
[150] Abhinav Rastogi, Xiaoxue Zang, et al. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, Vol. 34. 8689–8696.
[151] Siva Reddy, Danqi Chen, et al. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics* 7 (2019), 249–266.
[152] Marco Tulio Ribeiro, Sameer Singh, et al. 2016. "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.
[153] Hamidreza Rouzegar and Masoud Makrehchi. 2024. Enhancing text classification through LLM-driven active learning and human annotation. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*. 98–111.
[154] Baptiste Roziere, Jonas Gehring, et al. 2023. Code llama: Open foundation models for code. *Meta AI* (2023), 1–48.
[155] Arleen Salles, Kathinka Evers, et al. 2020. Anthropomorphism in AI. *AJOB Neuroscience* 11, 2 (2020), 88–95.
[156] Claude Sammut and Ranan B. Banerji. 1983. Learning concepts by asking. *Machine Learning: An Artificial Intelligence Approach, Volume II* 2 (1983), 167.
[157] Thomas Scialom, Serra Sinem Tekiroğlu, et al. 2020. Toward stance-based personas for opinionated dialogues. In *Proceedings of the Association for Computational Linguistics*. 2625–2635.
[158] Burr Settles. 2009. Active learning literature survey. *University of Wisconsin-Madison Department of Computer Sciences* (2009), 1–67.
[159] Emily Sheng, Kai-Wei Chang, et al. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 3407–3412.
[160] Weiyan Shi, Yu Li, et al. 2021. Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. In *Proceedings of the Association for Computational Linguistics*. 3478–3492.
[161] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.
[162] Anthony Sicilia, Jennifer Gates, et al. 2024. HumBEL: A human-in-the-loop approach for evaluating demographic factors of language models in human-machine conversations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1127–1143.
[163] Clemencia Siro, Mohammad Aliannejadi, et al. 2022. Understanding user satisfaction with task-oriented dialogue systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2018–2023.
[164] Marita Skjuve, Ida Maria Haugstveit, et al. 2019. Help! Is my chatbot falling into the uncanny valley? An empirical study of user experience in human–chatbot interaction. *Human Technology* 15, 1 (2019), 30–54.
[165] Tuva Lunde Smestad and Frode Volden. 2019. Chatbot personalities matters: Improving the user experience of chatbot interfaces. In *Proceedings of the 2018 International Conference on Internet Science*. 170–181.
[166] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1631–1642.
[167] Congzheng Song and Vitaly Shmatikov. 2019. Auditing data provenance in text-generation models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 196–206.
[168] Kyle Dylan Spurlock, Cagla Acun, et al. 2024. Evaluation of refined conversational recommendation based on. In *Proceedings of the Workshop on Generative AI for Recommender Systems and Personalization (GenAIRecP'24)*.
[169] Makesh Narsimhan Sreedhar, Kun Ni, et al. 2020. Learning improvised chatbots from adversarial modifications of natural language feedback. In *Proceedings of the Association for Computational Linguistics*. 2445–2453.
[170] Katherine Stasaski and Vikram Ramanarayanan. 2020. Automatic feedback generation for dialog-based language tutors using transformer models and active learning. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 1–12.
[171] Nisan Stiennon, Long Ouyang, et al. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* 33 (2020), 3008–3021.
[172] Hao Sun, Zhexin Zhang, et al. 2023. MoralDial: A framework to train and evaluate moral dialogue systems via moral discussions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 2213–2230.
[173] Bakhtiyar Syed, Gaurav Verma, et al. 2020. Adapting language models for non-parallel author-stylized rewriting. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, Vol. 34. 9008–9015.
[174] Gemini Team, Rohan Anil, et al. 2023. Gemini: A family of highly capable multimodal models. *Google DeepMind* (2023), 1–90.

[175] Silvia Terragni, Bruna Guedes, et al. 2022. BETOLD: A task-oriented dialog dataset for breakdown detection. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*. 23–34.

[176] Jincy Susan Thomas and Seena Thomas. 2018. Chatbot using gated end-to-end memory networks. *International Research Journal of Engineering and Technology (IRJET)* 5, 03 (2018), 3730–3735.

[177] James Thorne, Andreas Vlachos, et al. 2018. FEVER: A large-scale dataset for fact extraction and VERification. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (June 2018), 809–819.

[178] Bernadette Tix and Kim Binsted. 2024. Better results through ambiguity resolution: Large language models that ask clarifying questions. In *International Conference on Human-Computer Interaction*. 72–87.

[179] Hugo Touvron, Louis Martin, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Meta* (2023), 1–77.

[180] Usman Ahmad Usmani, Ari Happonen, and Junzo Watada. 2023. Human-centered artificial intelligence: Designing for user empowerment and ethical considerations. In *Proceedings of the 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications*. IEEE, 1–5.

[181] Lindsey Vanderlyn, Gianna Weber, et al. 2021. "It seemed like an annoying woman": On the perception and ethical considerations of affective language in text-based conversational agents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*. 44–57.

[182] Ashish Vaswani, Noam Shazeer, et al. 2017. Attention is all you need. *Proceedings of the 31st Conference on Neural Information Processing Systems* 30, 6000–6010.

[183] Anu Venkatesh, Chandra Khatri, et al. 2018. On evaluating and comparing open domain dialog systems. arXiv:1801.03625. Retrieved from https://arxiv.org/abs/1801.03625

[184] Siddharth Verma, Justin Fu, et al. 2022. CHAI: A CHatbot AI for task-oriented dialogue with offline reinforcement learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4471–4491.

[185] Michael Völske, Martin Potthast, et al. 2017. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*. 59–63.

[186] Mattias Wahde and Marco Virgolin. 2021. The five Is: Key principles for interpretable and safe conversational AI. In *Proceedings of the 4th International Conference on Computational Intelligence and Intelligent Systems*. 50–54.

[187] Eric Wallace, Pedro Rodriguez, et al. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics* 7 (2019), 387–401.

[188] Thiemo Wambsganss, Anne Höch, et al. 2021. Ethical design of conversational agents: Towards principles for a value-sensitive design. In *Proceedings of the International Conference on Wirtschaftsinformatik*. 539–557.

[189] Alex Wang, Amanpreet Singh, et al. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (Nov. 2018), 353–355.

[190] Huimin Wang, Baolin Peng, et al. 2020. Learning efficient dialogue policy from demonstrations through shaping. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6355–6365.

[191] Jiayin Wang, Fengran Mo, et al. 2024. A user-centric multi-intent benchmark for evaluating large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 3588–3612.

[192] Qianli Wang, Tatiana Anikina, et al. 2024. LLMCheckup: Conversational examination of large language models via interpretability tools and self-explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. 89–104.

[193] Xuezhi Wang, Jason Wei, et al. 2022. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations*. 1–24.

[194] Yunli Wang, Yu Wu, et al. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 3573–3578.

[195] Zijie J. Wang, Dongjin Choi, et al. 2021. Putting humans in the natural language processing loop: A survey. *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing* (April 2021), 47–52.

[196] Jason Wei, Xuezhi Wang, et al. 2022. Chain of thought prompting elicits reasoning in large language models. In *Proceedings of the 36th Conference on Neural Information Processing Systems*. Article 1800, 24824 - 24837 pages.

[197] Joseph Weizenbaum. 1966. ELIZA–A computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.

[198] Anuradha Welivita, Yubo Xie, et al. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 1251–1264.

[199] Jason E. Weston. 2016. Dialog-based language learning. *Advances in Neural Information Processing Systems* 29 (2016), 829–837.

[200] Tongshuang Wu, Michael Terry, et al. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–22.

[201] Xingjiao Wu, Luwei Xiao, et al. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* 135 (2022), 364–381.

[202] Yu Wu, Wei Wu, et al. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. (July 2017), 496–505.

[203] Zeqiu Wu, Yushi Hu, et al. 2023. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems* (2023), 1–26.

[204] Jing Xu, Megan Ung, et al. 2023. Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 13557–13572.

[205] Kerui Xu, Jingxuan Yang, et al. 2021. Adapting user preference to online feedback in multi-round conversational recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 364–372.

[206] Lin Xu, Qixian Zhou, et al. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, Vol. 33. 7346–7353.

[207] Qiongkai Xu, Lizhen Qu, et al. 2020. Personal information leakage detection in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 6567–6580.

[208] Wei Xu. 2019. Toward human-centered AI: A perspective from human-computer interaction. *Interactions* 26, 4 (2019), 42–46.

[209] Zihao Xu, Yi Liu, et al. 2024. A comprehensive study of jailbreak attack versus defense for large language models. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024*. 7432–7449.

[210] Özge Nilay Yalçın. 2020. Empathy framework for embodied conversational agents. *Cognitive Systems Research* 59 (2020), 123–132.

[211] Diyi Yang and Lucie Flek. 2021. Towards user-centric text-to-text generation: A survey. In *Proccedings of the 24th International Conference on Text, Speech, and Dialogue*. 3–22.

[212] Shunyu Yao, Jeffrey Zhao, et al. 2022. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the 11th International Conference on Learning Representations*. 1–33.

[213] Sanghyun Yi, Rahul Goel, et al. 2019. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. In *Proceedings of the 12th International Conference on Natural Language Generation*. 65–75.

[214] Saizheng Zhang, Emily Dinan, et al. 2018. Personalizing dialogue agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2204–2213.

[215] Xiang Zhang, Junbo Jake Zhao, et al. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*. 649–657.

[216] Yizhe Zhang, Siqi Sun, et al. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (July 2019), 270–278.

[217] Haiyan Zhao, Hanjie Chen, et al. 2024. Explainability for large language models: A survey. *ACM Transactions Intelligent Systems Technology* 15, 2 (2024), 1–38.

[218] Tiancheng Zhao and Maxine Eskenazi. 2016. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1–10.

[219] Denny Zhou, Nathanael Schärli, et al. 2023. Least-to-most prompting enables complex reasoning in large language models. In *Proceedings of the 11th International Conference on Learning Representations*. 1–61.

[220] Jingyan Zhou, Jiawen Deng, et al. 2022. Towards identifying social bias in dialog systems: Framework, dataset, and benchmark. In *Proceedings of the Association for Computational Linguistics*. 3576–3591.

[221] Wangchunshu Zhou, Jinyi Hu, et al. 2020. Towards interpretable natural language understanding with explanations as latent variables. *Advances in Neural Information Processing Systems* 33 (2020), 6803–6814.

[222] Yukun Zhu, Ryan Kiros, et al. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*. 19–27.

[223] Daniel M. Ziegler, Nisan Stiennon, et al. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019), 1–26.

[224] Caleb Ziems, Jane Yu, et al. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 3755–3773.

Received 26 April 2024; revised 13 March 2025; accepted 1 April 2025