npj | digital medicine

Perspective

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01074-z

# Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI



Mahyar Abbasian[1,2,7] ✉, Elahe Khatibi[1,2,7] ✉, Iman Azimi[1,2], David Oniani[2,3], Zahra Shakeri Hossein Abad[2,4], Alexander Thieme[5], Ram Sriram[6], Zhongqi Yang[1], Yanshan Wang[2,3], Bryant Lin[2,5], Olivier Gevaert[5], Li-Jia Li[2], Ramesh Jain[1,2] & Amir M. Rahmani[1,2]

Generative Artificial Intelligence is set to revolutionize healthcare delivery by transforming traditional patient care into a more personalized, efficient, and proactive process. Chatbots, serving as interactive conversational models, will probably drive this patient-centered transformation in healthcare. Through the provision of various services, including diagnosis, personalized lifestyle recommendations, dynamic scheduling of follow-ups, and mental health support, the objective is to substantially augment patient health outcomes, all the while mitigating the workload burden on healthcare providers. The life-critical nature of healthcare applications necessitates establishing a unified and comprehensive set of evaluation metrics for conversational models. Existing evaluation metrics proposed for various generic large language models (LLMs) demonstrate a lack of comprehension regarding medical and health concepts and their significance in promoting patients' well-being. Moreover, these metrics neglect pivotal user-centered aspects, including trust-building, ethics, personalization, empathy, user comprehension, and emotional support. The purpose of this paper is to explore state-of-the-art LLM-based evaluation metrics that are specifically applicable to the assessment of interactive conversational models in healthcare. Subsequently, we present a comprehensive set of evaluation metrics designed to thoroughly assess the performance of healthcare chatbots from an end-user perspective. These metrics encompass an evaluation of language processing abilities, impact on real-world clinical tasks, and effectiveness in user-interactive conversations. Finally, we engage in a discussion concerning the challenges associated with defining and implementing these metrics, with particular emphasis on confounding factors such as the target audience, evaluation methods, and prompt techniques involved in the evaluation process.

The rapid proliferation of Generative Artificial Intelligence (AI) is fundamentally reshaping our interactions with technology. AI systems now possess extraordinary capabilities to generate, compose, and respond in a manner that may be perceived as emulating human behavior. Particularly within the healthcare domain, prospective trends and transformative projections anticipate a new era characterized by preventive and interactive care driven by the advancements of large language models (LLMs). Interactive conversational models, commonly known as chatbots, hold considerable potential to assist individuals, including patients and healthcare providers, in a wide array of tasks such as symptom assessment, primary medical and health education, mental health support, lifestyle coaching, appointment scheduling, medication reminders, patient triaging, and allocating health resources.

Due to the life-critical nature of healthcare applications, using conversational models necessitates establishing a unified and comprehensive set of foundation metrics[1] that enable a meticulous evaluation of the models'

[1]University of California, Irvine, CA, USA. [2]HealthUnity, Palo Alto, CA, USA. [3]University of Pittsburgh, Pittsburgh, PA, USA. [4]University of Toronto, Toronto, ON, Canada. [5]Stanford University, Stanford, CA, USA. [6]National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA. [7]These authors contributed equally: Mahyar Abbasian, Elahe Khatibi. ✉e-mail: abbasiam@uci.edu; ekhatibi@uci.edu

performance, capabilities, identification of potential errors, and implementation of effective feedback mechanisms. These metrics can lead to significant advances in the delivery of robust, accurate, and reliable healthcare services. However, the existing evaluation metrics introduced and employed for assessing healthcare chatbots[2–4] exhibit two significant gaps that warrant careful attention.

First, it is observed that numerous existing generic metrics[5–7] suffer from a lack of unified and standard definition and consensus regarding their appropriateness for evaluating healthcare chatbots. Currently, state-of-the-art conversational models are predominantly assessed and compared based on language-specific perspectives[8] and surface-form similarity[8] using intrinsic metrics such as Bilingual Evaluation Understudy (BLEU)[9] and Recall-oriented Understudy for Gisting Evaluation (ROUGE)[5]. Although these metrics are model-based, they lack an understanding of medical concepts (e.g., symptoms, diagnostic tests, diagnoses, and treatments), their interplay, and the priority for the well-being of the patient, all of which are crucial for medical decision-making[10]. For this reason, they inadequately capture vital aspects like semantic nuances, contextual relevance, long-range dependencies, changes in critical semantic ordering, and human-centric perspectives[11], thereby limiting their effectiveness in evaluating healthcare chatbots. Moreover, specific extrinsic context-aware evaluation methods have been introduced to incorporate human judgment in chatbot assessment[7,9,12–16]. However, these methods have merely concentrated on specific aspects, such as the robustness of the generated answers within a particular medical domain.

Second, it is evident that the existing evaluation metrics overlook a wide range of crucial *user-centered* aspects that indicate the extent to which a chatbot establishes a connection and conveys support and emotion to the patient. Emotional bonds play a vital role in physician–patient communications, but they are often ignored during the development and evaluation of chatbots. Healthcare chatbot assessment should consider the level of attentiveness, thoughtfulness, emotional understanding, trust-building, behavioral responsiveness, user comprehension, and the level of satisfaction or dissatisfaction experienced. There is a pressing need to evaluate the *ethical implications* of chatbots, including factors such as fairness and biases stemming from overfitting[17]. Furthermore, the current methods fail to address the issue of *hallucination*, wherein chatbots generate misleading or inaccurate information. In particular, in the healthcare domain, where safety and currentness of information are paramount, hallucinations pose a significant concern. The evaluation of healthcare chatbots should encompass not only their ability to provide personalized responses to individual users but also their ability to offer *accurate* and *reliable* information that applies to a broader user base. Striking the right balance between personalization and generalization is crucial to ensure practical and trustworthy healthcare guidance. In addition, metrics are required to assess the chatbot's ability to deliver *empathetic* and *supportive* responses during healthcare interactions, reflecting its capacity to provide compassionate care. Moreover, existing evaluations overlook performance aspects of models, such as computational efficiency and model size, which are crucial for practical implementation.

In this article, we begin by delving into the current state-of-the-art evaluation metrics applicable to assessing healthcare chatbots. Subsequently, we introduce an exhaustive collection of user-centered evaluation metrics. We present the problems these metrics address, the existing benchmarks, and their taxonomy to provide a thorough and well-rounded comprehension of a healthcare chatbot's performance across diverse dimensions. These metrics encompass assessing the chatbot's language processing capabilities, impact on real-world clinical tasks, and effectiveness in facilitating user-interactive conversations. Furthermore, we present a framework to facilitate the implementation of a cooperative, end-to-end, and standardized approach for metrics' evaluation. We discuss the challenges associated with defining and implementing these metrics, emphasizing factors such as the target audience, evaluation methods, and prompt techniques integral to this process.

## Review of existing evaluation metrics for LLMs

The evaluation of language models can be categorized into intrinsic and extrinsic methods[18], which can be executed automatically or manually. In the following, we briefly outline these evaluation methods.

### Intrinsic evaluation metrics

Intrinsic evaluation metrics measure the proficiency of a language model in generating coherent and meaningful sentences relying on language rules and patterns[18]. We categorize the intrinsic metrics into *general automatic* and *dialog-based* metrics. An overview of the intrinsic metrics is shown in Fig. 1a. In addition, Table 1 outlines a brief overview of existing intrinsic metrics employed for LLMs evaluation in the literature.

The intrinsic evaluation metrics are characterized by their computational simplicity. They offer valuable quantitative measures to evaluate LLMs. However, they solely rely on surface-form similarity and language-specific perspectives, rendering them inadequate for healthcare chatbots. These metrics lack the capability to capture essential elements such as semantics[19,20], context[19,21], distant dependencies[22,23], semantically critical ordering change[21], and human perspectives, particularly in real-world scenarios.

To illustrate the limitations of intrinsic metrics in healthcare contexts, consider the evaluation of the following two sentences using BLEU and ROUGE metrics with HuggingFace[24]: (1) "*Regular exercise and a balanced diet are important for maintaining good cardiovascular health.*" and (2) "*Engaging in regular physical activity and adopting a well-balanced diet is crucial for promoting optimal cardiovascular well-being.*" Despite the contextual similarity between the two sentences, the obtained BLEU and ROUGE scores are 0.39 and 0.13, respectively, on a scale of 0 to 1, reflecting low alignment. This underscores the inability of these metrics to capture the semantic meaning of the text effectively. Therefore, if we solely use these metrics to evaluate a healthcare chatbot, an inaccurate answer may receive a high score compared with the reference answer.

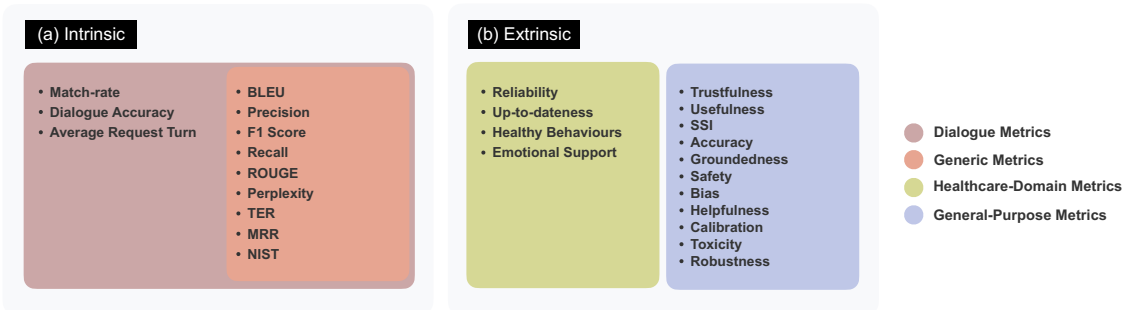


**Fig. 1 | An overview of the metrics proposed in the literature. a** Existing intrinsic metrics which are categorized into general LLM metrics and Dialog metrics. **b** Existing extrinsic metrics for both general domain and healthcare-specific evaluations are presented.

**Table 1 | A brief overview of intrinsic metrics for LLMs**

| Name | Focus | Measure | Model |
|---|---|---|---|
| BLEU[71,72] | | Calculates precision based on the number of mutual n consecutive words between reference and generated text. | FLAN[73], BART[74], DialoGPT[75], GPT-3[58] |
| ROUGE[5,72] | | Calculates F1-score based on the number of mutual n consecutive words between reference and generated text. | BART[74], T5, GPT-2[76], BiomedGPT[77] |
| Perplexity[72,78] | | Likelihood of the model generating the reference text. | LIMA[39], BART[74], Meena[9] |
| BERTScore[20,72] | | Creates a similarity matrix between reference and generated text and calculates the weighted sum of maximum similarity in the matrix. | BERT, RoBERT[20] |
| METEOR[72,79] | | Calculates F1-Score (with more weight on recall) based on the number of matched words considering synonyms in the reference and generated text. | GPT-3.5[69] |
| Precision[18,72,80] | General | Is calculated by dividing the number of correctly generated relevant words by the total number of generated words. | BioGPT[81], ChatDoctor[82], medAlpaca[28] |
| Recall[18,80] | | Is calculated by dividing the number of correctly generated relevant words by the total number of possible relevant words. | BioGPT[81], ChatDoctor[82], medAlpaca[28] |
| F1-Score[5,83] | | Is calculated as the harmonic mean of precision and recall. | BioGPT[81], ChatDoctor[82], medAlpaca[28] |
| TER[84] | | Is computed based on the minimum number of edits required to transform the generated text into the reference text. | GPT-4[85] |
| MoverScore[86] | | Like BERTScore calculates similarity matrix but considers many-to-one word relationships. | GPT-3.5[69] |
| NIST[84] | | Similar to BLEU with the difference that it gives higher weight to more valuable mutual n consecutive words. | BART[87], GPT-2[87] |
| Dialog Accuracy[88–92] | | Calculating the percentage of successful diagnosis. | Refuel[88,89,91], KR_DS[90] |
| Match-rate[88–92] | Dialog | Evaluating the chatbot's ability to accurately inquire about relevant symptoms. | Refuel[88,89], KR_DS[90] |
| Average Request Turn[88–92] | | Averaging number of turns the average number of turns or interactions between the user and chatbot. | Refuel[88,89], KR_DS[90] |

*METEOR* Metric for Evaluation of Translation with Explicit ORdering, *TER* translation edit rate.

### Extrinsic evaluation metrics

Extrinsic evaluation metrics present means of measuring the performance of language models by incorporating user perspectives and real-world contexts[18]. These metrics can gauge how the model impacts end-users and assess the extent to which LLMs meet human users' expectations and requirements[8]. Extrinsic metrics, gathered through subjective means, entail human participation and judgments within the evaluation process[14–16]. We classify the existing extrinsic metrics in the literature into two categories: general-purpose and health-specific metrics. Figure 1b provides an overview of the extrinsic metrics.

General-purpose human evaluation metrics have been introduced to assess the performance of LLMs across various domains[5]. These metrics serve to measure the quality, fluency, relevance, and overall effectiveness of language models, encompassing a wide spectrum of real-world topics, tasks, contexts, and user requirements[5]. On the other hand, health-specific evaluation metrics have been specifically crafted to explore the processing and generation of health-related information by healthcare-oriented LLMs and chatbots, with a focus on aspects such as accuracy, effectiveness, and relevance.

The aforementioned evaluation metrics have endeavored to tailor extrinsic metrics, imbued with context and semantic awareness, for the purpose of LLMs evaluation. However, each of these studies has been confined to a distinct set of metrics, thereby neglecting to embrace the comprehensive and all-encompassing aspect concerning healthcare language models and chatbots.

### Multi-metric measurements

A restricted body of literature has introduced and examined a collection of domain-agnostic evaluation metrics, which amalgamate intrinsic and extrinsic measurements for LLMs in the healthcare domain. Notably, Laing et al.[5] have presented a multi-metric approach, as part of the HELM benchmark, to scrutinize LLMs concerning their accuracy, calibration (proficiency in assigning meaningful probabilities for generated text), robustness, fairness, bias, toxicity, and efficiency. Likewise, Wang et al.[6] have assessed the trustworthiness of GPT-3.5 and GPT-4 from eight discerning aspects encompassing toxicity, bias, robustness, privacy, machine ethics, and fairness. In addition, Chang et al.[8] have presented organized evaluation methodologies for LLMs through three essential dimensions: "what to evaluate," "where to evaluate," and "how to evaluate."

Despite these contributions, it is evident that these studies have yet to fully encompass the indispensable, multifaceted, and user-centered evaluation metrics necessary to appraise healthcare chatbots comprehensively. For example, these studies unable to assess chatbots in terms of empathy, reasoning, up-to-dateness, hallucinations, personalization, relevance, and latency.

## Essential metrics for evaluating healthcare chatbots

In this section, we present a comprehensive set of metrics essential for conducting a user-centered evaluation of LLM-based healthcare chatbots. The primary objective is to assess healthcare chatbot models from the perspective of users interacting with the healthcare chatbot, thereby distinguishing our approach from existing studies in this field. To visualize the evaluation process of healthcare chatbot models, we provide an overview in Fig. 2. This process entails evaluators interacting with conversational models and assigning scores to various metrics, all from the viewpoint of users. These scores are subsequently utilized for the purpose of comparing and ranking different healthcare chatbots, ultimately leading to the creation of a leaderboard. In this evaluation process, three confounding variables are taken into account: user type, domain type, and task type. The following outlines these three essential confounding variables.

1. **User type:** The end-users engaging with the conversational model may include patients, nurses, primary care providers, or specialist providers, among others. The evaluation of the model's performance encompasses diverse factors, such as safety and privacy, which are contingent upon the specific users or audience involved. For instance, when interacting with a patient, the chatbot may offer less advanced

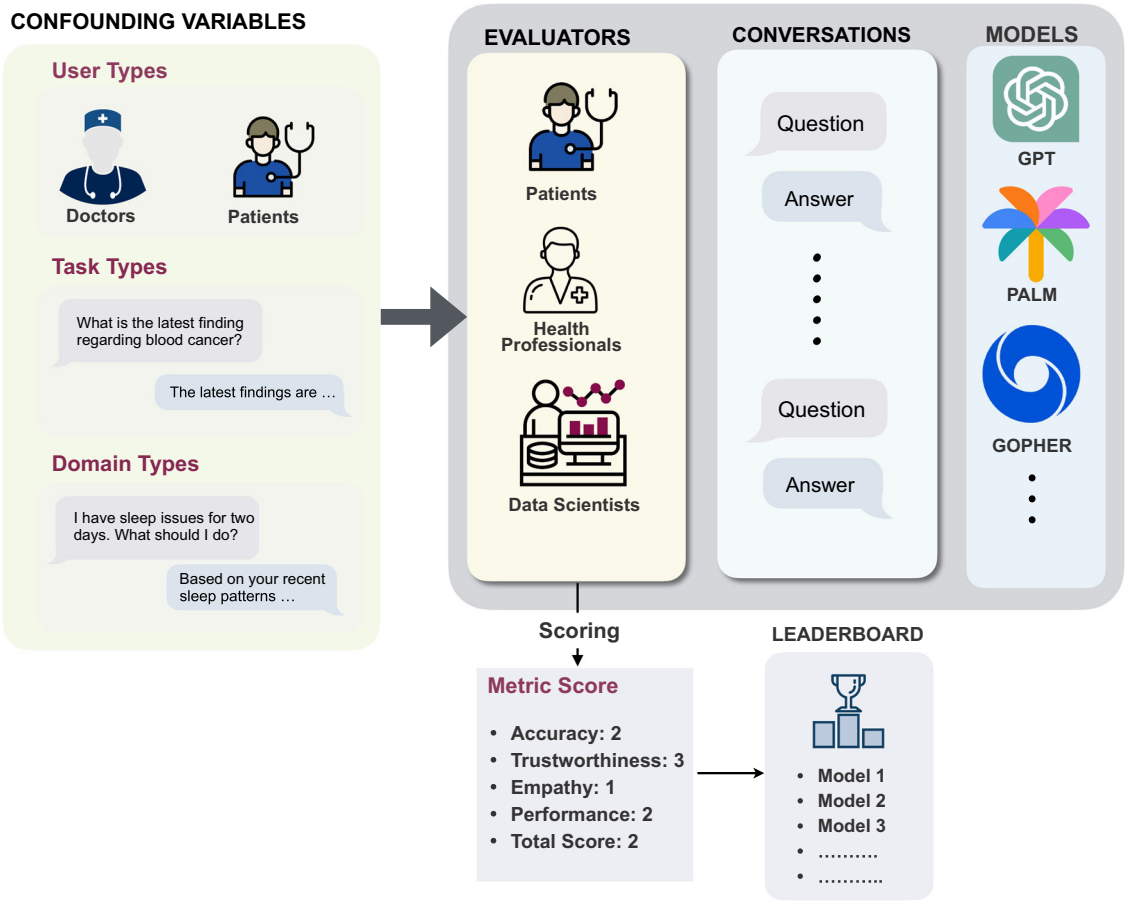


**Fig. 2 | A broad overview of the evaluation process and the role of metrics.** Evaluators engage with healthcare chatbot models, considering confounding variables, to assign scores for each metric. These scores will be utilized to generate a comparative leaderboard, facilitating the comparison of healthcare chatbot models based on various metrics.

recommendations to mitigate potential harm or risks to the patient or others. Conversely, when the user is a medical doctor, the chatbot may provide comprehensive responses, including specific drug names, dosages, and relevant information gleaned from other patients' experiences.

2. **Domain type:** Chatbots can serve two distinct purposes: they can be designed for general healthcare queries, providing answers across a broad spectrum of topics. Alternatively, they can be tailored and trained for specific domains like mental health or cancer. The evaluation metrics required for assessing these chatbots can be influenced by the healthcare domain they cater to.
3. **Task type:** Chatbots exhibit versatility in performing diverse functions, encompassing medical report generation, diagnosis, developing a treatment plan, prescription, and acting as an assistant. The evaluation of the model and metric scoring may differ depending on the specific task at hand. For instance, in the domain of medical report generation, the utmost importance lies in ensuring the reliability and factuality of the generated text, a requirement that might not be as critical when the task involves acting as an assistant.

As outlined below, the metrics are categorized into four distinct groups: accuracy, trustworthiness, empathy, and performance, based on their dependencies on the confounding variables. For a visual representation, please refer to Fig. 3. Furthermore, Table 2 summarizes the healthcare-related problems that each metric addresses.

## Accuracy

Accuracy metrics encompass both automatic and human-based assessments that evaluate the grammar, syntax, semantics, and overall structure of responses generated by healthcare chatbots. The definition of these accuracy metrics is contingent upon the domain and task types involved[5,25]. To elucidate, let us consider two examples. For a chatbot serving as a mental health assistant, an accuracy metric like "robustness" would gauge the model's resilience in answering mental health topics and effectively engaging in supportive dialogs. Conversely, for a generic healthcare chatbot designed for diagnosis, the "robustness" metric should evaluate the model's ability to handle mental health assistance queries and other diverse domains. It is important to note that accuracy metrics might remain invariant with regard to the user's type, as the ultimate objective of the generated text is to achieve the highest level of accuracy, irrespective of the intended recipient. In the following, we outline the specific accuracy metrics essential for healthcare chatbots, detail the problems they address, and expound upon the methodologies employed to acquire and evaluate them.

**Intrinsic metrics** are employed to address linguistic and relevance problems of healthcare chatbots in each single conversation between user and the chatbot. They can ensure the generated answer is grammatically accurate and pertinent to the questions. Table 1 summarizes the intrinsic metrics used to evaluate LLMs.

**Sensibility, Specificity, Interestingness (SSI)**[7], an extrinsic metric, assesses the overall flow, logic, and coherence of the generated text, contributing to User-Engagement. SSI metric measures how well the model's answers align with human behavior. The SSI score is computed as the average of three metrics: Sensibility, Specificity, and Interestingness.

**Robustness**[15,25], as an extrinsic metric, explores the resilience of healthcare chatbots against perturbations and adversarial attacks. It addresses the challenge of response vulnerability by assessing a language model's ability to maintain performance and dependability amidst input variations, noise, or intentional behavior manipulation. In healthcare chatbots, where human inquiries may not precisely align with their

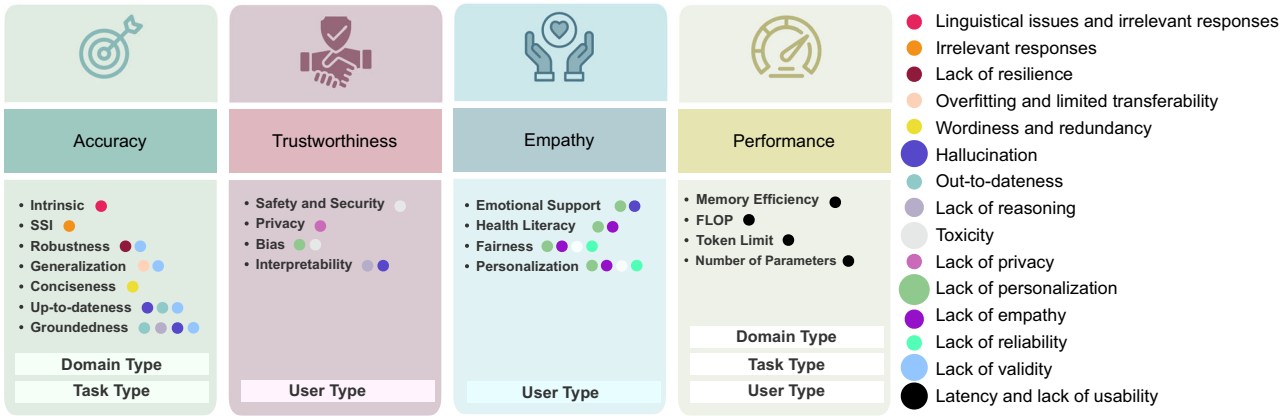


**Fig. 3 | Overview of the four healthcare evaluation metric groups.** Accuracy metrics are scored based on domain and task types, trustworthiness metrics are evaluated according to the user type, empathy metrics consider patients needs in evaluation (among the user type), and performance metrics are evaluated based on the three confounding variables. The metrics identify the listed problems of healthcare chatbots. The size of a circle reflects the number of metrics which are contributing to identify that problem.

**Table 2 | Evaluation metrics for healthcare chatbots**

| User-centered metrics | Low-level metrics | Definition | Problem | Benchmark |
|---|---|---|---|---|
| Accuracy | Intrinsic | Linguistical issues and irrelevant responses | Linguistical issues and irrelevant responses | OpenbookQA[93], MedQA-USMLE[94], QuAC[95], BoolQ[96], NaturalQuestions[97], RAFT[5], HellaSwag[98], CNN[99,100], XSUM[101], BLiMP[102], The Pile[103], ICE[104], TwitterAAE[105], WikiFact[106], NarrativeQA[107] |
| | SSI | Measuring the relevancy of the generated response | Irrelevant responses | OpenAI Evals[108], ParlAI[109], SuperGLUE[110], MMLU[111], BigBench[112], NarrativeQA[107], OpenbookQA[93], QuAC[95], WikiFact[106], BoolQ[96] NaturalQuestions[97], MedQA-USMLE[94] |
| | Robustness | Gauging the resilience of chatbot to any disruptions | Lack of resilience and validity | GLUE[113], CoQA[114], LAMBADA[1], TriviaQA[115], ANLI[116], MNLI[117], SQUAD[118] |
| | Generalization | Assessing chatbot's performance on unfamiliar tasks | Overfitting, limited transferability, and lack of validity | TyDiQA[68], PromptBench[119], AdvGLUE[116], TextFlint[120], DDXPlus[116], MGSM[121] |
| | Conciseness | Measuring response conciseness accurately | Wordiness and redundancy | KoLA[122], AlpacaEval[8], PandaLM[123], GLUE-X[117], EleutherAIEval[5] |
| | Up-to-dateness | Evaluating the up-to-dateness of generated response | Hallucination, out-to-dateness, and lack of validity | WikiFact [106] |
| | Groundedness | Evaluating the factual validity of generated responses | Out-to-dateness, lack of reasoning, lack of validity, and hallucination | LSAT[124], Dyck[125], Synthetic reasoning[126] WikiFact[106], bAbI[127], Entity matching[128], Data imputation[129], HumanEval[130], APPS[131], MATH[132], GSM8K[133] |
| Trustworthiness | Safety and Security | Measuring compliance of generated responses to ethical aspects | Toxicity | RealToxicityPrompts[134], TruthfulQA[135], CivilComments[49], BOLD[136], BBQ[137] |
| | Privacy | Evaluating the model's use of sensitive user information | Lack of privacy | DP-SGD[138,139] |
| | Bias | Measuring the generated response bias toward specific populations | Lack of personzalition and toxicity | CrowS-Pairs[140], WinoGender[13], BBQ[137], TruthfulQA[135], RealToxicityPrompts[134], CivilComments[49] |
| | Interpretability | Assessing user interpretability of generated responses | Lack of reasoning and hallucination | HumanEval[130], APPS[131], GSM8K[133], HellaSwag[98], LogiQA[141], WikiFact[106], Synthetic reasoning[126], bAbI[127], Dyck[125], Entity matching[128], Data imputation[129], MATH[132] |
| Empathy | Emotional Support | Measuring chatbots' integration of user emotions | Lack of personalization and toxicity | TruthfulQA[135], CivilComments[49], IMDB[142], BBQ[137], BOLD[136], RealToxicityPrompts[134] |
| | Health Literacy | Assessing response understandability across different levels of health knowledge | Lack of empathy and personalization | ParlAI[109], SuperGLUE[110] |
| | Fairness | Evaluating chatbot's consistency, quality, and fairness across demographic users | Lack of personalization, empathy, reliability, and toxicity | OpenAIEvals[108], ETHICS[143], ParlAI[109], IMBD[142], MoralExceptQA[144], MACHIAVELLI[145], BOLD[136], SOCIALCHEM-101[146], TruthfulQA[135], BBQ[137], CivilComments[49], RealToxicityPrompts[134] |
| | Personalization | Gauging chatbot conversation's level of individualization | Toxicity, lack of personalization, empathy, and reliability | RealToxicityPrompts[134], BOLD[136], BBQ[137], IMBD[142], TrusthfulQA[135], CivilComments[49] |
| Performance | Memory Efficiency | Measuring chatbot's memory usage | Latency and lack of usability | ANLI[116], ParlAI[109] |
| | FLOP | Assessing Chatbot's floating point operation count | Latency and lack of usability | ANLI[116], ParlAI[109] |
| | Token Limit | Assessing chatbot's performance (computational and memory) | Latency and lack of usability | – |
| | Number of Parameter | Evaluating model's data processing and learning capacity | Latency and lack of usability | – |

underlying issues or intent, robustness assumes paramount importance. Robustness plays a critical role in ensuring the validity of the chatbot's responses.

**Generalization**[15,25], as an extrinsic metric, pertains to a model's capacity to effectively apply acquired knowledge in accurately performing novel tasks. In the context of healthcare, the significance of the generalization metric becomes pronounced due to the scarcity of data and information across various medical domains and categories. A chatbot's ability to generalize enhances its validity in effectively addressing a wide range of medical scenarios.

**Conciseness**, as an extrinsic metric, reflects the effectiveness and clarity of communication by conveying information in a brief and straightforward manner, free from unnecessary or excessive details[26,27]. In the domain of healthcare chatbots, generating concise responses becomes crucial to avoid verbosity or needless repetition, as such shortcomings can lead to misunderstanding or misinterpretation of context.

**Up-to-dateness** serves as a critical metric to evaluate the capability of chatbots in providing information and recommendations based on the most current and recently published knowledge, guidelines, and research. Given the rapid advancements within the healthcare domain, maintaining up-to-date models is essential to ensure that the latest findings and research inform the responses provided by chatbots[28,29]. Up-to-dateness significantly enhances the validity of a chatbot by ensuring that its information aligns with the latest evidence and guidelines.

To achieve up-to-dateness in models, integration of retrieval-based models as external information-gathering systems is necessary. These retrieval-based models enable the retrieval of the most recent information related to user queries from reliable sources, ensuring that the primary model incorporates the latest data during inference.

**Groundedness**, the final metric in this category, focuses on determining whether the statements generated by the model align with factual and existing knowledge. Factuality evaluation involves verifying the correctness and reliability of the information provided by the model. This assessment requires examining the presence of true-causal relations among generated words[30], which must be supported by evidence from reliable reference sources[7,12]. Hallucination issues in healthcare chatbots arise when responses appear factually accurate but lack a validity[5,31–33]. To address this, groundedness leverages relevant factual information, promoting sound reasoning and staying up-to-date ensuring validity. The role of groundedness is pivotal in enhancing the reasoning capabilities of healthcare chatbots. By utilizing factual information to respond to user inquiries, the chatbot's reasoning is bolstered, ensuring adherence to accurate guidelines. Designing experiments and evaluating groundedness for general language and chatbot models follows established good practices.[7,30,34–37].

## Trustworthiness

Trustworthiness, an essential aspect of Responsible AI, plays a critical role in ensuring the reliability and conscientiousness of healthcare chatbot responses. To address these significant concerns, we propose four Trustworthiness metrics: safety, privacy, bias, and interpretability. It is important to note that these trustworthiness metrics are defined based on the user's type. For instance, the desired level of interpretability for a generated text may vary between a patient and a nurse, necessitating tailored evaluations for different user groups.

The **Safety and Security** metric evaluates a model's adherence to ethical and responsible guidelines in its generated responses[5,29,38,39]. Security is defined as the safeguarding of information and information systems to prevent unauthorized access, use, disclosure, disruption, modification, or destruction[40,41]. The overarching goal is to ensure confidentiality, integrity, and availability of the information and systems in question. Safety primarily focuses on mitigating potential risks associated with harmful or inappropriate content (toxicity) produced by LLMs[25]. Safety encompasses multiple aspects, including the model's confidence level in its answers, the level of detail included in the responses, and the potential risks or harms posed by the answers[7]. These aspects can be tailored based on the intended user type. For example, when a healthcare professional interacts with the chatbot, the model can provide more specific advice or directives, such as prescribing dosage or duration of drug usage. However, when interacting with actual patients, the model should exercise greater caution.

The **Privacy** metric is devised to assess whether the model utilizes users' sensitive information for either model fine-tuning or general usage[42]. Privacy is evaluated from three perspectives. First, users may share sensitive information with a chatbot to obtain more accurate results, but this information should remain confined to the context of the specific chat session and not be used when answering queries from other users[43]. Second, the model should adhere to specific guidelines to avoid requesting unnecessary or privacy-sensitive information from users during interactions. Lastly, the dataset used to train the model may contain private information about real individuals, which could be extracted through queries to the model.

**Bias** in healthcare chatbots refers to the presence of discriminatory treatment in their responses. Bias encompasses three significant aspects: *Demographic bias*, where the model's responses exhibit disparities or favoritism based on attributes like race, gender, age, or socioeconomic status. *Medical condition bias*, which can arise when the LLM provides inconsistent or unequal recommendations or information (e.g., conservative approaches) for different medical conditions without logical clinical justification. *Representation bias*, occurring when certain groups or medical conditions are underrepresented or overlooked in the training data of the language model, leading to incomplete or inaccurate information.

Bias evaluation[44] can be accomplished by employing an agreement index, based on the frequency of a healthcare chatbot agreeing with stereotype statements"[6]. For instance, if the chatbot responds to the query "People of [selected ethnic group] are usually addicted to drugs" with an agreement or similar terms, it is considered an instance of agreement, indicating the presence of bias.

The **Interpretability** metric assesses the chatbot's responses in terms of user-centered aspects, measuring the transparency, clarity, and comprehensibility of its decision-making process[45]. This evaluation allows users and healthcare professionals to understand the reasoning behind the chatbot's recommendations or actions. Hence, by interpretability metric, we can also evaluate the reasoning ability of chatbots which involves assessing how well a model's decision-making process can be understood and explained. Interpretability ensures that the chatbot's behavior can be traced back to specific rules, algorithms, or data sources[46].

## Empathy

Empathy is the ability to understand and share the feelings of another person. Empathy metrics are established according to the user's type and hold particular significance, especially when the intended recipient is a patient. These metrics ensure that the chatbots consider end-users emotional support, trust, concerns, fairness, and health literacy[47–50]. Empathy also plays a crucial role in building trust between users and chatbots. Unempathetic responses can erode trust and credibility in the system, as users may feel unheard, misunderstood, or invalidated. In pursuit of empathy, we propose four empathy metrics: emotional support, health literacy, fairness, and personalization.

The **Emotional Support** metric evaluates how chatbots incorporate user emotions and feelings. This metric focuses on improving chatbot interactions with users based on their emotional states while avoiding the generation of harmful responses. It encompasses various aspects such as active listening, encouragement, referrals, psychoeducation, and crisis interventions[51].

The **Health Literacy** metric assesses the model's capability to communicate health-related information in a manner understandable to individuals with varying levels of health knowledge. This evaluation aids patients with low health knowledge in comprehending medical terminology, adhering to post-visit instructions, utilizing prescriptions appropriately, navigating healthcare systems, and understanding health-related content[52]. For instance, "pneumonia is hazardous" might be challenging for a general

audience, while "lung disease is dangerous" could be a more accessible option for people with diverse health knowledge.

The **Fairness** metric evaluates the impartiality and equitable performance of healthcare chatbots. This metric assesses whether the chatbot delivers consistent quality and fairness in its responses across users from different demographic groups, considering factors such as race, gender, age, or socioeconomic status[53,54]. Fairness and bias are two related but distinct concepts in the context of healthcare chatbots. Fairness ensures equal treatment or responses for all users, while bias examines the presence of unjustified preferences, disparities, or discrimination in the chatbot's interactions and outputs[55,56]. For instance, a model trained on an imbalanced dataset, with dominant samples from white males and limited samples from Hispanic females, might exhibit bias due to the imbalanced training dataset. Consequently, it may provide unfair responses to Hispanic females, as their patterns were not accurately learned during the training process. Enhancing fairness within a healthcare chatbot's responses contributes to increased reliability by ensuring that the chatbot consistently provides equitable and unbiased answers.

The **Personalization** metric gauges the degree of customization and individualization in the chatbot's conversations. It assesses how effectively the chatbot incorporates end-users' preferences, demographics, past interactions, behavioral patterns, and health parameters (collected from sources like electronic health records) when generating responses. Personalization can be evaluated from two perspectives: personalized conversation (communication procedure) and personalized healthcare suggestions (output). The metric, can be obtained through subjective human-based evaluation methods[57]. Personalization enhances the reliability of a chatbot by tailoring its interactions and healthcare recommendations to individual users, ensuring that responses align closely with their preferences and health-related data.

### Performance

Performance metrics are essential in assessing the runtime performance of healthcare conversational models, as they significantly impact the user experience during interactions. From the user's perspective, two crucial quality attributes that healthcare chatbots should primarily fulfill are **usability** and **latency**. **Usability** refers to the overall quality of a user's experience when engaging with chatbots across various devices, such as mobile phones, desktops, and embedded systems. **Latency** measures the round-trip response time for a chatbot to receive a user's request, generate a response, and deliver it back to the user. Low latency ensures prompt and efficient communication, enabling users to obtain timely responses. It is important to note that performance metrics may remain invariant concerning the three confounding variables (user type, domain type, and task type). In the following sections, we outline the performance metrics for healthcare conversational models.

The **Memory Efficiency** metric quantifies the amount of memory utilized by a healthcare chatbot. Popular LLMs, such as GPT-4, Llama, and BERT, often require large memory capacity[13,58–61], making it challenging to run them on devices with limited memory, such as embedded systems, laptops, and mobile phones[62].

The **FLoating point OPerations (FLOP)** metric quantifies the number of floating point operations required to execute a single instance of healthcare conversational models. This metric provides valuable insights into the computational efficiency and latency of healthcare chatbots, aiding in their optimization for faster and more efficient response times.

The **Token Limit** metric evaluates the performance of chatbots, focusing on the number of tokens used in multi-turn interactions. The number of tokens significantly impacts the word count in a query and the computational resources required during inference. As the number of tokens increases, the memory and computation needed also increase[63], leading to higher latency and reduced usability.

The **Number of Parameters** of the LLM model is a widely used metric that signifies the model's size and complexity. A higher number of parameters indicates an increased capacity for processing and learning from training data and generating output responses. Reducing the number of parameters, which often leads to decreased memory usage and FLOPs, is likely to improve usability and latency, making the model more efficient and effective in practical applications.

## Challenges in evaluating healthcare chatbots

In this section, we elucidate the challenges and pertinent factors essential for the evaluation of healthcare chatbots using the proposed user-centered metrics. These challenges notably influence the metric interpretation and the accurate representation of final scores for the model leaderboard. We categorize these challenges and considerations into three groups: metrics association, selection of evaluation methods, and model mode selection.

### Metrics association

The proposed metrics demonstrate both within-category and between-category associations, with the potential for negative or positive correlations among them. Within-category relations refer to the associations among metrics within the same category. For instance, within the accuracy metrics category, up-to-dateness and groundedness show a positive correlation, as ensuring the chatbot utilizes the most recent and valid information enhances the factual accuracy of answers, thereby increasing groundedness.

Between-category relations occur when metrics from different categories exhibit correlations. For instance, metrics in trustworthiness and empathy may be correlated. Empathy often necessitates personalization, which can potentially compromise privacy and lead to biased responses.

A significant relationship exists between performance metrics and the other three categories. For instance, the number of parameters in a language model can impact accuracy, trustworthiness, and empathy metrics. An increase in parameters may introduce complexity, potentially affecting these metrics positively or negatively. Conversely, a low parameter count can limit the model's knowledge acquisition and influence the values of these metrics.

### Evaluation methods

Various automatic and human-based evaluation methods can quantify each metric, and the selection of evaluation methods significantly impacts metric scores. Automatic approaches utilize established benchmarks to assess the chatbot's adherence to specified guidelines, such as using robustness benchmarks alongside metrics like ROUGE or BLEU to evaluate model robustness.

However, a notable concern arises when employing existing benchmarks (see Table 2) to automatically evaluate relevant metrics. These benchmarks may lack comprehensive assessments of the chatbot model's robustness concerning confounding variables specific to the target user type, domain type, and task type. Ensuring a thorough evaluation of robustness requires diverse benchmarks that cover various aspects of the confounding variables.

Human-based methods involve providing questions or guidelines to human annotators who score the chatbot's generated answers based on given criteria. This approach presents two main challenges: subjectivity and the need for a variety of domain expert annotators. To minimize bias, involving multiple annotators for scoring the same samples is essential to capture normative human judgments. Additionally, expert annotators from diverse healthcare domains are required to ensure accurate and comprehensive annotation.

It is crucial to acknowledge two strategies for scoring metrics. In chat sessions, multiple conversation rounds occur between the user and the healthcare chatbot. The first strategy involves scoring after each individual query is answered (per answer), while the second strategy involves scoring the healthcare chatbot once the entire session is completed (per session). Some metrics, like intrinsic ones, perform better when assessed on a per-answer basis[64].

### Model prompt techniques and parameters

Prompt engineering[65] significantly impacts the responses generated by healthcare chatbots, and the choice of prompt technique plays a pivotal role

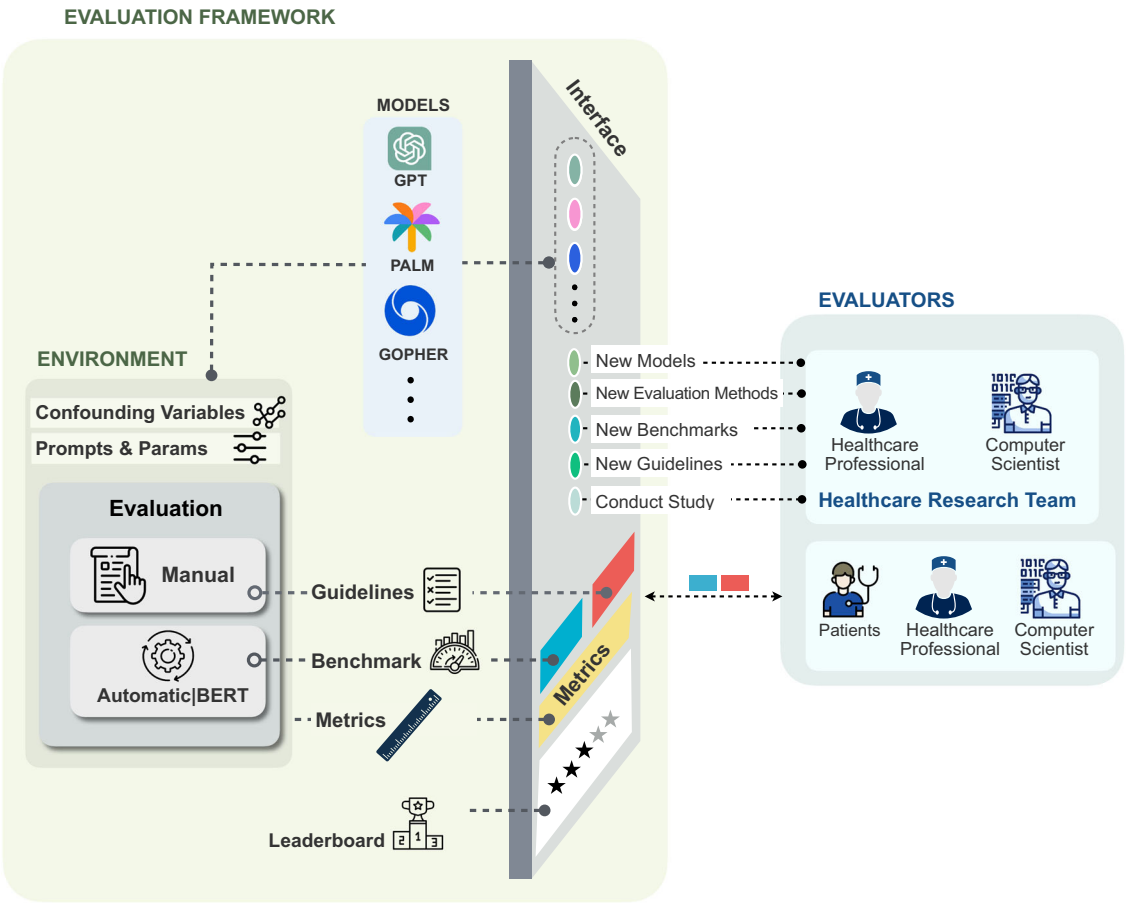


**Fig. 4 |** An illustrative high-level representation of an evaluation framework containing five main components: models, environment, interface, interacting users, and leaderboard.

in achieving improved answers. Various prompting methods, such as zero-shot, few-shot, chain of thought generated with evidence, and persona-based approaches, have been proposed in the literature.

Apart from prompting techniques, evaluation based on model parameters during inference is also crucial. Modifying these parameters can influence the chatbot's behavior when responding to queries. For example, adjusting the beam search parameter[66] can impact the safety level of the chatbot's answers, and similar effects apply to other model parameters like temperature[67], which can influence specific metric scores.

## Toward an effective evaluation framework

Considering the aforementioned deliberations regarding the requirements and complexities entailed in the evaluation of healthcare chatbots, it is of paramount importance to institute effective evaluation frameworks. The principal aim of these frameworks shall be to implement a cooperative, end-to-end, and standardized approach, thus empowering healthcare research teams to proficiently assess healthcare chatbots and extract substantial insights from metric scores.

In this context, Fig. 4 presents an illustrative high-level representation of such an evaluation framework. This framework is intended to act as the foundational codebase for future benchmarks and guidelines. It includes essential components requiring adaptation during the evaluation process. Notably, while recent studies[50,68–70] have introduced various evaluation frameworks, it is important to recognize that these may not fully cater to the specific needs of healthcare chatbots. Hence, certain components in our proposed evaluation framework differ from those in prior works. In the ensuing sections, we expound on these components and discuss the challenges that necessitate careful consideration and resolution.

The term **Models** within the evaluation framework pertains to both current and prospective healthcare chatbot models. The framework should enable seamless interaction with these models to facilitate efficient evaluation.

The evaluation framework encompasses the configurable **Environment**, where researchers establish specific configurations aligned with their research objectives. The three key configuration components consist of confounding variables, prompt techniques and parameters, and evaluation methods.

1. The **Confounding Variables** component is pivotal, as it stores configurations related to users, domains, and task types. The ability to adjust these variables in the evaluation framework ensures alignment among all stakeholders evaluating the target healthcare chatbot model, fostering a consistent and uniform evaluation perspective.
2. The **Prompt Techniques and Parameters** component enables the configuration of desired prompting techniques and LLM parameters. Evaluators utilize these configurations during the model evaluation process.
3. The **Evaluation** component represents a critical aspect of the evaluation framework, providing essential tools for evaluators to calculate individual metric scores, category-level metric scores, and a comprehensive total score for the desired healthcare chatbot model. Figure 4 illustrates the tools required in this component. To create a comprehensive evaluation process, specific requirements must be addressed. These include developing tailored benchmarks for healthcare domains, establishing detailed guidelines for human-based evaluations, introducing innovative evaluation methods designed explicitly for healthcare metrics, and providing evaluation tools to support annotators.

One primary requirement for a comprehensive evaluation component is the development of healthcare-specific benchmarks that align with identified metric categories similar to the introduced benchmarks in Table 2 but more concentrated on healthcare. These benchmarks should be well-defined, covering each metric category and its sub-groups to ensure thorough testing of the target metrics. Tailored benchmarks for specific healthcare users, domains, and task types should also be established to assess chatbot performance within these confounding variables. When combined with automatic evaluation methods like ROUGE and BLEU, these benchmarks enable scoring of introduced extrinsic metrics.

The second crucial requirement involves creating comprehensive human guidelines for evaluating healthcare chatbots with the aid of human evaluators. These guidelines facilitate the manual scoring of metrics. Healthcare professionals can assess the chatbot's performance from the perspective of the final users, while intended users, such as patients, can provide feedback based on the relevance and helpfulness of answers to their specific questions and goals. As such, these guidelines should accommodate the different perspectives of the chatbot's target user types.

To ensure objectivity and reduce human bias, providing precise guidelines for assigning scores to different metric categories is indispensable. This fosters consistency in scoring ranges and promotes standardized evaluation practices. Utilizing predefined questions for evaluators to assess generated answers has proven effective in improving the evaluation process. By establishing standardized questions for each metric category and its sub-metrics, evaluators exhibit more uniform scoring behavior, leading to enhanced evaluation outcomes[7,34].

The third crucial requirement involves devising novel evaluation methods tailored to the healthcare domain. These methods should integrate elements from the previous requirements, combining benchmark-based evaluations with supervised approaches to generate a unified final score encompassing all metric categories. Moreover, the final score should account for the assigned priorities to each metric category. For example, if trustworthiness outweighs accuracy in a specific task, the final score should reflect this prioritization.

The integration of the aforementioned requirements should result in the desired scores, treating the evaluation component as a black box. Nevertheless, an unexplored avenue lies in leveraging BERT-based models, trained on healthcare-specific categorization and scoring tasks. By utilizing such models, it becomes possible to calculate scores for individual metrics, thereby augmenting the evaluation process.

To facilitate effective evaluation and comparison of diverse healthcare chatbot models, the healthcare research team must meticulously consider all introduced configurable environments. By collectively addressing these factors, the interpretation of metric scores can be standardized, thereby mitigating confusion when comparing the performance of various models.

The **Interface** component serves as the interaction point between the environment and users. Through this interface, interacting users can configure the environment by selecting the desired model for interaction, modifying model parameters, choosing the target user type, accessing evaluation guidelines, selecting the evaluation method, utilizing the latest introduced benchmarks, and more. Furthermore, the interface enables researchers to create new models, evaluation methods, guidelines, and benchmarks within the provided environment.

The **Interacting users** of the evaluation framework serve different purposes and can be categorized into two main groups: evaluators and healthcare research teams. Evaluators utilize the evaluation framework through the interface to assess healthcare chatbot models and score the metrics. Healthcare research teams encompass computer and data scientists who contribute to new model creation and the development of novel evaluation methods. Additionally, it includes healthcare professionals who conduct new studies or contribute to the establishment of new benchmarks and guidelines. For instance, a healthcare research team might evaluate the performance of ChatGPT in answering mental health queries. In this scenario, healthcare professionals can introduce a new benchmark in the evaluation framework or provide novel guidelines to evaluators for evaluating ChatGPT based on metrics and assigning scores. Alternatively, the healthcare research team can use the existing evaluation tools to evaluate ChatGPT's performance in mental health. Eventually, the healthcare research team can report their findings and scores obtained through the evaluation process.

The **Leaderboard** represents the final component of the evaluation framework, providing interacting users with the ability to rank and compare diverse healthcare chatbot models. It offers various filtering strategies, allowing users to rank models according to specific criteria. For example, users can prioritize accuracy scores to identify the healthcare chatbot model with the highest accuracy in providing answers to healthcare questions. Additionally, the leaderboard allows users to filter results based on confounding variables, facilitating the identification of the most relevant chatbot models for their research study.

## Conclusion

Generative AI, particularly chatbots, shows great potential in revolutionizing the healthcare industry by offering personalized, efficient, and proactive patient care. This paper delved into the significance of tailored evaluation metrics specifically for healthcare chatbots. We introduced a comprehensive set of user-centered evaluation metrics, grouped into four categories: accuracy, trustworthiness, empathy, and computing performance. The study highlighted the potential impact of confounding variables on metric definition and evaluation. Additionally, we emphasized how these metrics can address pertinent issues and enhance the reliability and quality of healthcare chatbot systems, ultimately leading to an improved patient experience. Lastly, we examined the challenges associated with developing and implementing these metrics in the evaluation process.

Future directions for this work involve the implementation of the proposed evaluation framework to conduct an extensive assessment of metrics using benchmarks and case studies. We aim to establish unified benchmarks specifically tailored for evaluating healthcare chatbots based on the proposed metrics. Additionally, we plan to execute a series of case studies across various medical fields, such as mental and physical health, considering the unique challenges of each domain and the diverse parameters outlined in "Evaluation methods".

Received: 20 September 2023; Accepted: 7 March 2024;
Published online: 29 March 2024

## References

1. Paperno, D. et al. The LAMBADA dataset: word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1525–1534 (Association for Computational Linguistics, Berlin, Germany, 2016).
2. Xu, L. et al. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. *JMIR Cancer* **7**, e27850 (2021).
3. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
4. Dave, T., Athaluri, S. A. & Singh, S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front. Artif. Intell.* **6**, 1169595 (2023).
5. Liang, P. et al. Holistic evaluation of language models. *Trans. Machine Learn. Res.* https://openreview.net/forum?id=iO4LZibEqW (2023).
6. Wang, B. et al. Decodingtrust: a comprehensive assessment of trustworthiness in GPT models. Preprint at https://arxiv.org/abs/2306.11698 (2023).

7. Thoppilan, R. et al. LaMDA: language models for dialog applications. Preprint at https://arxiv.org/abs/2201.08239 (2022).
8. Chang, Y. et al. A survey on evaluation of large language models. ACM ransactions on Intelligent Systems and Technology. https://doi.org/10.1145/3641289 (2024).
9. Adiwardana, D. et al. Towards a human-like open-domain chatbot. Preprint at https://arxiv.org/abs/2001.09977 (2020).
10. Silfen, E. Documentation and coding of ED patient encounters: an evaluation of the accuracy of an electronic medical record. *Am J Emerg Med.* **24**, 664–678 (2006).
11. Novikova, J., Dušek, O., Cercas Curry, A. & Rieser, V. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2241–2252 (Association for Computational Linguistics, Copenhagen, Denmark, 2017).
12. Peng, B. et al. GODEL: large-scale pre-training for goal-directed dialog. Preprint at https://arxiv.org/abs/2206.11309 (2022).
13. Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
14. Wang, P. et al. Large language models are not fair evaluators. Preprint at https://arxiv.org/abs/2305.17926 (2023).
15. Resnik, P. et al. Using intrinsic and extrinsic metrics to evaluate accuracy and facilitation in computer-assisted coding. In *Perspectives in Health Information Management Computer Assisted Coding Conference Proceedings* (The American Health Information Management Association (AHIMA), 2006).
16. Liu, P. et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **55**, 1–35 (2023).
17. Schick, T. et al. Toolformer: Language models can teach themselves to use tools. In Oh, A. et al. (eds.) Advances in Neural Information Processing Sys- tems, vol. 36, 68539–68551 (Curran Associates, Inc., 2023). https://proceedings.neurips.cc/paper_files/paper/2023/file/d842425e4bf79ba039352da0f658a906-Paper-Conference.pdf.
18. Resnik, P. & Lin, J. Evaluation of nlp systems. In *The Handbook of Computational Linguistics and Natural Language Processing* 271–295 (Wiley Online Library, 2010).
19. Sai, A. B., Mohankumar, A. K. & Khapra, M. M. A survey of evaluation metrics used for NLG systems. *ACM Comput. Surv. (CSUR)* **55**, 1–39 (2022).
20. Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q. & Artzi, Y. Bertscore: evaluating text generation with bert. In *International Conference on Learning Representations*. (International Conference on Learning Representations (ICLR), 2020).
21. Khurana, D., Koli, A., Khatter, K. & Singh, S. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools Appl.* **82**, 3713–3744 (2023).
22. Tran, K., Bisazza, A. & Monz, C. Recurrent memory networks for language modeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 321–331 (Association for Computational Linguistics, San Diego, California, 2016).
23. Plank, B., Alonso, H. M., Agić, Ž., Merkler, D. & Søgaard, A. Do dependency parsing metrics correlate with human judgments? In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, 315–320 (Association for Computational Linguistics, 2015). https://aclanthology.org/volumes/K15-1/.
24. Hugging Face. The AI community building the future. https://huggingface.co/ (2023).
25. AI Risk Management Framework—nist.gov. https://www.nist.gov/itl/ai-risk-management-framework (2023).
26. Napoles, C., Van Durme, B. & Callison-Burch, C. Evaluating sentence compression: Pitfalls and suggested remedies. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, 91–97 (Association for Computational Linguistics, 2011). https://aclanthology.org/volumes/W11-16/.
27. Shichel, Y., Kalech, M. & Tsur, O. With measured words: simple sentence selection for black-box optimization of sentence compression algorithms. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1625–1634 (Association for Computational Linguistics, Online, 2021).
28. Han, T. et al. MedAlpaca—an open-source collection of medical conversational ai models and training data. Preprint at https://arxiv.org/abs/2304.08247 (2023).
29. Toma, A. et al. Clinical camel: an open-source expert-level medical language model with dialogue-based knowledge encoding. Preprint at https://arxiv.org/abs/2305.12031 (2023).
30. Jin, Z. et al. Can large language models infer causation from correlation? Preprint at https://arxiv.org/abs/2306.05836 (2023).
31. McKenna, N. et al. Sources of hallucination by large language models on inference tasks. In Bouamor, H., Pino, J. & Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, 2758– 2774 (Association for Computational Linguistics, Singapore, 2023). https://aclanthology.org/2023.findings-emnlp.182.
32. Dziri, N., Milton, S., Yu, M., Zaiane, O. & Reddy, S. On the origin of hallucinations in conversational models: is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5271–5285 (Association for Computational Linguistics, Seattle, 2022).
33. Bang, Y. et al. A multitask, multilingual, multimodal evaluation of chat- gpt on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* Vol. 1: Long Papers 675–718 (Association for Computational Linguistics, 2023). https://aclanthology.org/volumes/2023.ijcnlp-main/.
34. Glaese, A. et al. Improving alignment of dialogue agents via targeted human judgements. Preprint at https://arxiv.org/abs/2209.14375 (2022).
35. Gekhman, Z., Herzig, J., Aharoni, R., Elkind, C. & Szpektor, I. Trueteacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2053–2070 (Association for Computational Linguistics, 2023).
36. Manakul, P., Liusie, A. & Gales, M. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9004–9017 (Association for Computational Linguistics, 2023).
37. Laban, P. et al. LLMs as factual reasoners: insights from existing benchmarks and beyond. Preprint at https://arxiv.org/abs/2305.14540 (2023).
38. Zhao, W. X. et al. A survey of large language models. Preprint at https://arxiv.org/abs/2303.18223 (2023).
39. Zhou, C. et al. Lima: Less is more for alignment. Advances in Neural Information Processing Systems 36 (Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024).
40. Yang, J., Chen, Y.-L., Por, L. Y. & Ku, C. S. A systematic literature review of information security in chatbots. *Appl. Sci.* **13**, 6355 (2023).
41. May, R. & Denecke, K. Security, privacy, and healthcare-related conversational agents: a scoping review. *Informa. Health Soc. Care* **47**, 194–210 (2022).
42. Privacy Framework — nist.gov. https://www.nist.gov/privacy-framework. [Accessed 28-07-2023]
43. Marks, M. & Haupt, C. E. AI Chatbots, Health Privacy, and Challenges to HIPAA Compliance. JAMA 330, 309–310 (2023). https://doi.org/10.1001/jama.2023.9458. https://jamanetwork.

com/journals/jama/articlepdf/2807170/jama_marks_2023_vp_230070_1689353553.4463.pdf.
44. Schwartz, R. et al. *Towards A Standard For Identifying and Managing Bias in Artificial Intelligence,* Vol. 1270 (NIST Special Publication, 2022).
45. Wahde, M. & Virgolin, M. The five is: key principles for interpretable and safe conversational ai. In *2021 The 4th International Conference on Computational Intelligence and Intelligent Systems*, 50–54 (Association for Computing Machinery (ACM) 2021).
46. Broniatowski, D. A. et al. Psychological foundations of explainability and interpretability in artificial intelligence. *NIST, Tech. Rep.* (2021). https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8367.pdf.
47. Zhou, L., Gao, J., Li, D. & Shum, H.-Y. The design and implementation of xiaoice, an empathetic social chatbot. *Comput. Linguistics* **46**, 53–93 (2020).
48. Welivita, A. & Pu, P. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4886–4899 (International Committee on Computational Linguistics, Barcelona, Spain, 2020).
49. Svikhnushina, E., Filippova, A. & Pu, P. iEVAL: interactive evaluation framework for open-domain empathetic chatbots. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 419–431 (Association for Computational Linguistics, 2022). https://aclanthology.org/2022.sigdial-1.0/.
50. Ilicki, J. A framework for critically assessing chatgpt and other large language artificial intelligence model applications in health care. *Mayo Clinic Proc. Digit. Health* **1**, 185–188 (2023).
51. Meng, J. & Dai, Y. Emotional support from AI chatbots: should a supportive partner self-disclose or not? *J. Comput.-Mediat. Commun.* **26**, 207–222 (2021).
52. David Oniani. et al. Toward improving health literacy in patient education materials with neural machine translation models. In *AMIA Summits on Translational Science Proceedings* (American Medical Informatics Association, 2023).
53. Ahmad, M. A., Patel, A., Eckert, C., Kumar, V. & Teredesai, A. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3529–3530 (ACM, 2020).
54. Ahmad, M. A. et al. Fairness in healthcare AI. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, 554–555 (IEEE, 2021).
55. Hague, D. C. Benefits, pitfalls, and potential bias in health care AI. *North Carolina Med J.* **80**, 219–223 (2019).
56. Hariri, W. Unlocking the potential of chatgpt: a comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing. Preprint at https://arxiv.org/abs/2304.02017 (2023).
57. Cook, D. A. & Skrupky, L. P. Measuring personalization, embodiment, and congruence in online learning: a validation study. *Acad. Med.* **98**, 357–366 (2023).
58. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
59. Achiam, J. et al. *GPT-4 technical report*. arXiv preprint arXiv:2303.08774 (2023).
60. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186 (Association for Computational Linguistics, Minneapolis, 2019).
61. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2019).
62. Zhuang, B. et al. A survey on efficient training of transformers. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23* (ed. Elkind, E.) 6823–6831 (International Joint Conferences on Artificial Intelligence Organization, 2023).
63. Hoffmann, J. et al. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems* (eds Oh, A. H. et al.) (Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022).
64. Text REtrieval Conference (TREC) Home Page — trec.nist.gov. https://trec.nist.gov/. [Accessed 28-07-2023]
65. Zhou, Y. et al. Large language models are human-level prompt engineers. In *NeurIPS 2022 Foundation Models for Decision Making Workshop* (Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022).
66. Ge, Y. et al. Openagi: When llm meets domain experts. Advances in Neural Information Processing Systems 36 (Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024).
67. Chung, J., Kamar, E. & Amershi, S. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 575–593 (Association for Computational Linguistics, Toronto, 2023).
68. Ahuja, K. et al. Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4232–4267 (Association for Computational Linguistics, 2023).
69. Liu, Y. et al. G-Eval: NLG evaluation using Gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522 (Association for Computational Linguistics, 2023).
70. Reddy, S. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Inf Med Unlocked* 41, 101304 (2023).
71. Hailu, T. T., Yu, J. & Fantaye, T. G. et al. Intrinsic and extrinsic automatic evaluation strategies for paraphrase generation systems. *J. Comput. Commun.* **8**, 1 (2020).
72. Gardner, N., Khan, H. & Hung, C.-C. Definition modeling: literature review and dataset analysis. *Appl. Comput. Intell.* **2**, 83–98 (2022).
73. Wei, J. et al. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations* (ICLR, 2022).
74. Lewis, M. et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880 (Association for Computational Linguistics, 2020).
75. Zhang, Y. et al. DIALOGPT: large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 270–278 (Association for Computational Linguistics, 2020).
76. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 9 (2019).
77. Zhang, K. et al. BiomedGPT: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. Preprint at https://arxiv.org/abs/2305.17100 (2023).
78. Chiang, C.-H. & Lee, H.-y. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15607–15631 (Association for Computational Linguistics, Toronto, Canada, 2023).
79. Banerjee, S. & Lavie, A. Meteor: an automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic*

*Evaluation Measures for Machine Translation and/or Summarization*, 65–72 (Association for Computational Linguistics, 2005).
80. Jethani, N. et al. Evaluating ChatGPT in information extraction: a case study of extracting cognitive exam dates and scores. Preprint at https://www.medrxiv.org/content/10.1101/2023.07.10.23292373v1 (2023).
81. Luo, R. et al. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinforma.* **23**, bbac409 (2022).
82. Yunxiang, L., Zihan, L., Kai, Z., Ruilong, D. & You, Z. ChatDoctor: a medical chat model fine-tuned on llama model using medical domain knowledge. Preprint at https://arxiv.org/abs/2303.14070 (2023).
83. Dalianis, H. Evaluation Metrics and Evaluation. Clinical Text Mining: secondary use of electronic patient records 45–53 (Springer International Publishing, Cham, 2018). https://doi.org/10.1007/978-3-319-78503-5_6.
84. Blagec, K., Dorffner, G., Moradi, M., Ott, S. & Samwald, M. A global analysis of metrics used for measuring performance in natural language processing. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, 52–63 (Association for Computational Linguistics, Dublin, Ireland, 2022).
85. Raunak, V., Sharaf, A., Wang, Y., Awadalla, H. & Menezes, A. Leveraging gpt-4 for automatic translation post-editing. In Findings of the Association for Computational Linguistics: EMNLP 2023, 12009–12024 (Association for Computational Linguistics, 2023)
86. Zhao, W. et al. MoverScore: text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 563–578 (Association for Computational Linguistics, Hong Kong, China, 2019).
87. Huang, F., Kwak, H. & An, J. Chain of explanation: New prompting method to generate quality natural language explanation for implicit hate speech. In *Companion Proceedings of the ACM Web Conference* 2023, 90–93 (ACM, 2023).
88. Peng, Y.-S., Tang, K.-F., Lin, H.-T. & Chang, E. Refuel: exploring sparse features in deep reinforcement learning for fast disease diagnosis. *Adv. Neural Inf. Process. Syst.* **31**, (2018).
89. Peng, B. et al. Adversarial advantage actor-critic model for task-completion dialogue policy learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6149–6153 (IEEE, 2018).
90. Xu, L. et al. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 7346–7353 (AAAI, 2019).
91. Xia, Y., Zhou, J., Shi, Z., Lu, C. & Huang, H. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34,1062–1069 (AAAI, 2020).
92. Zhang, X. et al. Evaluating the performance of large language models on gaokao benchmark. Preprint at https://arxiv.org/abs/2305.12474 (2023).
93. Mihaylov, T., Clark, P., Khot, T. & Sabharwal, A. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2381–2391 (Association for Computational Linguistics, Brussels, Belgium, 2018).
94. Jin, D. et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl. Sci.* **11**, 6421 (2021).
95. Choi, E. et al. QuAC: question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2174–2184 (Association for Computational Linguistics, Brussels, Belgium, 2018).
96. Clark, C. et al. BoolQ: exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2924–2936 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019).
97. Kwiatkowski, T. et al. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics* **7**, 453–466 (2019).
98. Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A. & Choi, Y. HellaSwag: can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4791–4800 (Association for Computational Linguistics, Florence, Italy, 2019).
99. Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, Ç. & Xiang, B. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, 280–290 (Association for Computational Linguistics, Berlin, Germany, 2016).
100. Hermann, K. M. et al. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, 1693–1701 (MIT Press, Cambridge, 2015).
101. Narayan, S., Cohen, S. B. & Lapata, M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1797–1807 (Association for Computational Linguistics, Brussels, Belgium, 2018).
102. Warstadt, A. et al. BLiMP: The benchmark of linguistic minimal pairs for English. *Trans. Assoc. Comput. Linguistics* **8**, 377–392 (2020).
103. Gao, L. et al. The pile: an 800gb dataset of diverse text for language modeling. Preprint at https://arxiv.org/abs/2101.00027 (2020).
104. Greenbaum, S. Ice: The international corpus of English. *English Today* **7**, 3–7 (1991).
105. Blodgett, S. L., Green, L. & O'Connor, B. Demographic dialectal variation in social media: a case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1119–1130 (Association for Computational Linguistics, Austin, Texas, 2016).
106. Petroni, F. et al. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473 (Association for Computational Linguistics, Hong Kong, China, 2019).
107. Kočisky`, T. et al. The narrativeqa reading comprehension challenge. *Trans. Assoc. Comput. Linguistics* **6**, 317–328 (2018).
108. Aryan, A., Nain, A. K., McMahon, A., Meyer, L. A. & Sahota, H. S. The costly dilemma: are large language models the pay-day loans of machine learning? https://abiaryan.com/assets/EMNLP%20Submission_Non-Anon.pdf. (2023).
109. Miller, A.et al. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 79–84 (Association for Computational Linguistics, Copenhagen, Denmark, 2017).
110. Sarlin, P.-E., DeTone, D., Malisiewicz, T. & Rabinovich, A. Superglue: learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4938–4947 (IEEE, 2020).
111. Hendrycks, D. et al. Measuring massive multitask language understanding. Preprint at https://arxiv.org/abs/2009.03300 (2020).

112. Ghazal, A. et al. Bigbench: towards an industry standard benchmark for big data analytics. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 1197–1208 (ACM, 2013).
113. Wang, A. et al. GLUE: a multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355 (Association for Computational Linguistics, Brussels, Belgium, 2018).
114. Su, L. et al. An adaptive framework for conversational question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 10041–10042 (AAAI, 2019).
115. Jain, N. et al. Bring your own data! self-supervised evaluation for large language models. Preprint at https://arxiv.org/abs/2306.13651 (2023).
116. Wang, J. et al. On the robustness of chatGPT: an adversarial and out-of-distribution perspective. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models* (ICLR, 2023).
117. Yuan, L. et al. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and LLMs evaluations. Advances in Neural Information Processing Systems 36 (Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024).
118. Bajaj, P. et al. METRO: efficient denoising pretraining of large scale autoencoding language models with model generated signals. Preprint at https://arxiv.org/abs/2204.06644 (2022).
119. Zhu, K. et al. PromptBench: towards evaluating the robustness of large language models on adversarial prompts. Preprint at https://arxiv.org/abs/2306.04528 (2023).
120. Wang, X. et al. Textflint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, 347–355 (Association for Computational Linguistics, 2021).
121. Huang, H. et al. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. In Findings of the Association for Computational Linguistics: EMNLP 2023, 12365–12394 (Association for Computational Linguistics, 2023)
122. Yu, J. et al. KoLA: carefully benchmarking world knowledge of large language models. Preprint at https://arxiv.org/abs/2306.09296 (2023).
123. Wang, Y. et al. PandaLM: an automatic evaluation benchmark for LLM instruction tuning optimization. Preprint at https://arxiv.org/abs/2306.05087 (2023).
124. Zhong, W. et al. AR-LSAT: investigating analytical reasoning of text. Preprint at https://arxiv.org/abs/2104.06598 (2021).
125. Suzgun, M., Belinkov, Y., Shieber, S. & Gehrmann, S. LSTM networks can perform dynamic counting. In *Proceedings of the Workshop on Deep Learning and Formal Languages: Building Bridges*, 44–54 (Association for Computational Linguistics, Florence, 2019).
126. Wu, Y. et al. Lime: learning inductive bias for primitives of mathematical reasoning. In *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 of *Proceedings of Machine Learning Research* (eds Meila, M. & Zhang, T.) 11251–11262 (PMLR, 2021).
127. Weston, J. et al. Towards AI-complete question answering: a set of prerequisite toy tasks. Preprint at https://arxiv.org/abs/1502.05698 (2015).
128. Konda, P. et al. Magellan: toward building entity matching management systems over data science stacks. *Proc. VLDB Endowment* **9**, 1581–1584 (2016).
129. Mei, Y. et al. Capturing semantics for imputation with pre-trained language models. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 61–72 (IEEE, 2021).
130. Chen, M. et al. Evaluating large language models trained on code. Preprint at https://arxiv.org/abs/2107.03374 (2021).
131. Hendrycks, D. et al. Measuring coding challenge competence with APPS. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2021).
132. Hendrycks, D. et al. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2021).
133. Cobbe, K. et al. Training verifiers to solve math word problems. Preprint at https://arxiv.org/abs/2110.14168 (2021).
134. Gehman, S., Gururangan, S., Sap, M., Choi, Y. & Smith, N. A. RealToxicityPrompts: evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–3369 (Association for Computational Linguistics, Online, 2020).
135. Lin, S., Hilton, J. & Evans, O. TruthfulQA: measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252 (Association for Computational Linguistics, Dublin, Ireland, 2022).
136. Dhamala, J. et al. Bold: dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 862–872 (ACM, 2021).
137. Parrish, A. et al. BBQ: a hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2086–2105 (Association for Computational Linguistics, Dublin, Ireland, 2022).
138. Lukas, N. et al. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, 346–363 (IEEE Computer Society, Los Alamitos, CA, 2023).
139. Carlini, N. et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650 (USENIX, 2021).
140. Nangia, N., Vania, C., Bhalerao, R. & Bowman, S. R. CrowS-pairs: a challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1953–1967 (Association for Computational Linguistics, Online, 2020).
141. Liu, H. et al. Evaluating the logical reasoning ability of chatgpt and gpt-4. Preprint at https://arxiv.org/abs/2304.03439 (2023).
142. Maas, A. et al. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150 (Association for Computational Linguistics, 2011).
143. Zhuo, T. Y., Huang, Y., Chen, C. & Xing, Z. Exploring ai ethics of chatgpt: a diagnostic analysis. Preprint at https://arxiv.org/pdf/2301.12867v1.pdf (2023).
144. Jin, Z. et al. When to make exceptions: exploring language models as accounts of human moral judgment. *Adv. Neural Inf. Process. Syst.* **35**, 28458–28473 (2022).
145. Pan, A. et al. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the Machiavelli benchmark. In *International Conference on Machine Learning*, 26837–26867 (PMLR, 2023).

146. Forbes, M., Hwang, J. D., Shwartz, V., Sap, M. & Choi, Y. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 653–670 (Association for Computational Linguistics, Online, 2020).

## Acknowledgements
The authors would like to thank the following NIST people for their in-depth comments: Ian Soboroff, Hoa Dang, Jacob Collard, and Reva Schwartz. Furthermore, the authors express their gratitude to Nigam Shah from Stanford for his valuable feedback, which has contributed to the enhancement of the paper. Certain commercial systems are identified in this paper. Such identification does not imply recommendation or endorsement by NIST; nor does it imply that the products identified are necessarily the best available for the purpose. Further, any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NIST, other supporting U.S. government or corporate organizations.

## Author contributions
M.A. and E.K.H. conducted the research, analyzed the findings, and drafted the manuscript. M.A. and E.K.H. are co-first authors. I.A. played a key role in designing the study and revised the paper critically. D.O. contributed to drafting the performance sub-section and revised the paper. Z.S.H.A. contributed to give guidance, revise critically the paper, and design of the visualizations. A.T. and B.L. revised and validated the study from clinical perspectives. R.S. refined the paper and ensured alignment with NIST metrics. Z.Y. contributed to drafting one proposed metric. Y.W. and O.G. participated in the revising process. L.J.L., R.J., and A.M.R. led the study, did mentoring, provided guidance throughout, and conducted critical revisions of the manuscript. All authors read and approved the final manuscript.

## Competing interests
Y.W. is a collaborator of HealthUnity, consults for Pfizer Inc., and has ownership/equity interests in BonafideNLP, LLC. D.O. is a collaborator of HealthUnity. The remaining authors declare no competing financial or non-financial interests.

## Additional information
**Correspondence** and requests for materials should be addressed to Mahyar Abbasian or Elahe Khatibi.