BenchCouncil Transactions on Benchmarks, Standards and Evaluations 3 (2023) 100136

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/

Full length article

Benchmarking, ethical alignment, and evaluation framework for conversational AI: Advancing responsible development of ChatGPT

Partha Pratim Ray

Department of Computer Applications, Sikkim University, Gangtok, Sikkim 737102, India

ARTICLE INFO

Keywords:
Conversational AI
ChatGPT
Evaluation framework
Benchmarks
Adaptive standards
Intelligent evaluation

ABSTRACT

Conversational AI systems like ChatGPT have seen remarkable advancements in recent years, revolutionizing human–computer interactions. However, evaluating the performance and ethical implications of these systems remains a challenge. This paper delves into the creation of rigorous benchmarks, adaptable standards, and an intelligent evaluation methodology tailored specifically for ChatGPT. We meticulously analyze several prominent benchmarks, including GLUE, SuperGLUE, SQuAD, CoQA, Persona-Chat, DSTC, BIG-Bench, HELM and MMLU illuminating their strengths and limitations. This paper also scrutinizes the existing standards set by OpenAI, IEEE's Ethically Aligned Design, the Montreal Declaration, and Partnership on AI's Tenets, investigating their relevance to ChatGPT. Further, we propose adaptive standards that encapsulate ethical considerations, context adaptability, and community involvement. In terms of evaluation, we explore traditional methods like BLEU, ROUGE, METEOR, precision–recall, F1 score, perplexity, and user feedback, while also proposing a novel evaluation approach that harnesses the power of reinforcement learning. Our proposed evaluation framework is multidimensional, incorporating task-specific, real-world application, and multi-turn dialogue benchmarks. We perform feasibility analysis, SWOT analysis and adaptability analysis of the proposed framework. The framework highlights the significance of user feedback, integrating it as a core component of evaluation alongside subjective assessments and interactive evaluation sessions. By amalgamating these elements, this paper contributes to the development of a comprehensive evaluation framework that fosters responsible and impactful advancement in the field of conversational AI.

1. Introduction

In recent years, the rapid rise of conversational AI systems has reshaped human–computer interactions, propelling us towards a future where natural language conversations with machines become commonplace. Among the myriad of AI systems, ChatGPT, a product of OpenAI, has emerged as a paragon, showcasing remarkable language generation capabilities [1,2]. As this field gains momentum, the necessity to create stringent benchmarks, adaptable standards, and intelligent evaluation criteria becomes paramount to drive responsible development and constant refinement of systems like ChatGPT [3–5].

ChatGPT has garnered significant attention for its impressive language generation capabilities and ability to engage in contextually relevant conversations. However, the evaluation of such systems presents unique challenges that need to be addressed to ensure their continuous improvement and responsible development.

The need for robust benchmarks, adaptive standards, and intelligent evaluation criteria arises from the increasing demand for conversational AI systems that can understand and respond to human queries, provide meaningful interactions, and maintain ethical considerations [6–9]. The evaluation of these systems requires a comprehensive and multidimensional approach that goes beyond traditional metrics and embraces the complexities of language understanding, context awareness, and ethical alignment.

Motivated by these challenges, this paper proposes a comprehensive evaluation framework for ChatGPT that encompasses prominent benchmarks, adaptive standards, and intelligent evaluation methods [10–12]. The framework aims to enhance the performance assessment, ethical alignment, and user satisfaction of ChatGPT. By providing a clear roadmap for evaluation, the proposed framework ensures the responsible and impactful development of ChatGPT and future conversational AI systems [13,14].

Our research objectives are multi-pronged:

- **Performance Assessment Enhancement:** We endeavor to design task-specific benchmarks and evaluation metrics to assess ChatGPT's prowess across an array of conversational tasks, emphasizing its comprehension of context, maintenance of coherence, and delivery of precise and relevant responses.

E-mail address: ppray@cus.ac.in.

https://doi.org/10.1016/j.tbench.2023.100136
Received 20 June 2023; Received in revised form 19 July 2023; Accepted 27 July 2023
Available online 9 August 2023

- **Ethical Alignment:** Given the profound influence of AI in our lives, the development of adaptive standards is crucial for ensuring that ChatGPT complies with ethical guidelines. We leverage the principles outlined in recognized frameworks such as IEEE's Ethically Aligned Design and the Montreal Declaration, to mitigate potential biases, safeguard user privacy, and promote responsible data handling.
- **Innovating Evaluation Techniques:** We place significant emphasis on refining evaluation methodologies that gauge the quality and effectiveness of ChatGPT. By examining metrics beyond traditional measures, harnessing user feedback, and utilizing reinforcement learning techniques, we aspire to provide a comprehensive and nuanced evaluation.

Our work makes substantial contributions to the field. Firstly, we offer an in-depth analysis of leading benchmarks in conversational AI, providing insights into their strengths and limitations. Secondly, we investigate the applicability of existing ethical standards to ChatGPT and propose adaptive standards that ensure ethical and responsible conversational AI practices. Thirdly, we examine prevalent evaluation methods and propose an innovative, multi-dimensional approach to benchmarking ChatGPT. We also underscore the value of user-centered evaluation, and advocate for the integration of user feedback, subjective assessments, and interactive evaluation sessions into the overall evaluation framework.

Our ultimate goal is to develop an integrated evaluation framework that facilitates the development of conversational AI systems that are not only proficient linguistically, but also ethically aligned, user-centric, and adaptable to evolving challenges and expectations. The ensuing sections will unpack the specifics of our evaluation framework for ChatGPT, offering a comprehensive analysis that serves to drive the responsible and impactful development of conversational AI systems.

2. State-of-the-art of benchmarks, standards, and evaluation criteria

Benchmarking is required for ChatGPT to ensure the model's performance meets the objectives and standards set by its developers and users. A few key reasons for this necessity are:

- **Quality Assurance:** Benchmarking helps verify that the model's responses are accurate, contextually appropriate, and free from factual errors or misconceptions. It checks whether the model can understand and generate text in a manner that meets the expectations for human-like conversation.
- **Improvement Over Time:** By benchmarking, developers can identify the model's strengths and weaknesses. This information guides the future improvement of the model, enhancing its performance over time.
- **User Experience:** Benchmarking is crucial to ensure a positive user experience. The model should respond to users in a way that is engaging, helpful, and respectful. The ability to manage various conversational scenarios is key to meeting user expectations.
- **Ethical Compliance:** With benchmarking, developers can ensure that the model handles sensitive topics appropriately, respects user privacy, and adheres to the guidelines for responsible AI usage.
- **Comparison with Other Models:** Finally, benchmarking allows for an objective comparison of ChatGPT with other AI models. This can aid in choosing the best tool for specific applications and helps in communicating the model's capabilities to potential users or stakeholders.

Benchmarking the efficacy of ChatGPT demands meticulous planning, along with the strategic implementation of multiple key measures.

- Firstly, human-like conversation simulation should be checked: can it maintain relevant and engaging dialogue, mirroring the coherence, empathy, humor, and complexity a human might offer?
- Next, factual accuracy is critical — the AI should provide up-to-date, reliable information consistent with its knowledge cut-off. Natural language understanding and generation are essential too, evidenced by the model's ability to parse complex input and create grammatically sound, clear and concise output. Furthermore, the model's capacity for context-awareness is crucial, keeping track of ongoing conversations and adapting responses to situational nuances.
- Lastly, but not least, ethical considerations must be evaluated, observing how well the model respects privacy, avoids inappropriate content, and handles sensitive topics. Therefore, a comprehensive benchmark for ChatGPT necessitates a holistic assessment, scrutinizing not only its intellectual prowess but also its ability to maintain meaningful, responsible, and human-like interactions.

This subsection critically evaluates the existing benchmarks, standards, and evaluation methods utilized in the field of conversational AI, focusing on their strengths, weaknesses, and limitations. It provides a comprehensive review of prominent benchmarks such as GLUE, SuperGLUE, SQuAD, CoQA, Persona-Chat, DSTC, BIG-Bench, HELM and MMLU along with an analysis of the standards set by OpenAI, IEEE's Ethically Aligned Design, the Montreal Declaration, and Partnership on AI's Tenets. Additionally, it discusses common evaluation methods like BLEU, ROUGE, METEOR, precision–recall, F1 score, perplexity, and user feedback.

2.1. Benchmarks

2.1.1. GLUE and SuperGLUE

General Language Understanding Evaluation (GLUE) [15] and its successor, SuperGLUE [16], are benchmarks designed to evaluate the performance of models across a wide range of NLP tasks. GLUE consists of nine tasks, including question-answering, sentiment analysis, and textual entailment. SuperGLUE builds upon GLUE and includes more challenging tasks, pushing the boundaries of NLP models.

- Strengths
  - Comprehensive: GLUE and SuperGLUE cover diverse tasks, enabling evaluation of models' generalization capabilities.
  - Research Focus: These benchmarks encourage researchers to develop models that perform well across multiple NLP tasks.
- Weakness
  - Task-Specific Limitations: GLUE and SuperGLUE may not capture all nuances and complexities of specific tasks.
  - Limited Scope: The benchmarks may not cover all possible types of NLP tasks.

2.1.2. SQuAD

The Stanford Question Answering Dataset (SQuAD) is a popular benchmark for question answering models [17]. It provides passages and corresponding questions that require understanding of the passage to answer.

- Strengths
  - Contextual Understanding: SQuAD assesses a model's ability to comprehend and extract information from passages.
  - High-Quality Dataset: The dataset is carefully curated, providing reliable evaluation data for question-answering tasks.

- Weakness
  - Task-Specific: SQuAD primarily focuses on question answering and may not generalize well to other conversational tasks.
  - Complexity Representation: The complexity of questions and passages may not fully represent the diversity of real-world applications.

2.1.3. Conversational question answering (CoQA)

The CoQA benchmark is designed to evaluate models on their ability to handle conversational question answering, which requires understanding of the conversation history [18].

- Strengths
  - Conversation Context: CoQA evaluates models' ability to maintain context and generate coherent responses in a conversational setting.
  - Realistic Interactions: The dataset captures the dynamic nature of conversations, adding a layer of complexity to the evaluation. Research Focus: These benchmarks encourage researchers to develop models that perform well across multiple NLP tasks.
- Weakness
  - Limited Conversational Data: Availability of conversational datasets like CoQA may be restricted, hindering broader evaluation.
  - Complexity of Context: Modeling conversational context accurately can be challenging, and the benchmark may not fully capture all contextual nuances.

2.1.4. Persona-chat

Persona-Chat focuses on maintaining consistent personas during conversations [19]. It consists of a dataset of over 131,000 utterances where models are trained to engage in dialogue while adhering to predefined personas.

- Strengths
  - Persona Consistency: Persona-Chat evaluates models on their ability to sustain and embody specific personas during conversations.
  - Human-like Interaction: The benchmark encourages the development of more engaging and natural conversational AI models.
- Weakness
  - Persona Requirement: The necessity to maintain personas may not be applicable or relevant to all conversational AI applications.
  - Overlooking Other Aspects: Focusing solely on persona consistency may divert attention from other essential factors such as accuracy and relevance of responses.

2.1.5. Dialogue system technology challenges (DSTC)

DSTC consists of annual competitions that offer a benchmark for various dialogue-related tasks [20]. The challenges encompass a wide range of dialogue system facets, including dialogue state tracking, sentiment analysis, and natural language understanding.

- Strengths
  - Task Variety: DSTC covers diverse dialogue-related tasks, allowing evaluation across multiple dimensions of dialogue systems.
  - Research Advancement: The competitions encourage the development of innovative techniques and foster collaboration in the field.
- Weakness
  - Contextual Diversity: Given the complexity and variability of human conversations, it can be challenging for a benchmark like DSTC to sufficiently cover the diversity of conversational contexts that a dialogue system might encounter in real-world applications.
  - Competition Limitations: The competition format may restrict flexibility for researchers to explore different approaches.

2.1.6. BIG-Bench

BIG-Bench is a benchmark for evaluating large language models, specifically focusing on assessing their performance across various language tasks [21]. It covers a wide range of tasks such as text classification, summarization, translation, question answering, and more. BIG-Bench utilizes a large-scale dataset to provide a comprehensive evaluation of the models. It is an open-source benchmark that promotes collaboration and reproducibility in the research community.

- Strengths
  - Comprehensive Evaluation: BIG-Bench aims to provide a comprehensive evaluation framework for large language models by covering various language tasks, including text classification, summarization, translation, question answering, and more.
  - Diverse Benchmark Tasks: It includes a wide range of benchmark tasks, allowing researchers to assess the model's performance across different domains and linguistic capabilities.
  - Large-Scale Dataset: BIG-Bench utilizes a large-scale dataset, enabling robust evaluation and providing a more realistic assessment of the model's capabilities.
  - Open-Source and Reproducible: The benchmark is open-source, facilitating collaboration among researchers, and providing a reproducible evaluation platform.
- Weakness
  - Limited Task-Specific Evaluation: While BIG-Bench covers a wide range of language tasks, it may lack task-specific evaluations that focus on the nuances and requirements of individual tasks. This term refers to the tendency in some previous works to evaluate AI systems based on a narrow set of tasks, often those for which the system was specifically trained or designed. While this approach can provide valuable insights into the system's performance on those specific tasks, it can also be somewhat limiting as it might not reflect the system's adaptability to other tasks or contexts. For instance, a chatbot trained for customer service might perform well in that specific context but struggle to carry on a casual, open-ended conversation. It is important to include a range of tasks in the evaluation to get a better sense of the system's versatility and adaptability.
  - Potential Bias in Dataset: Depending on the data sources used for training, there might be biases present in the benchmark dataset, which could impact the fairness and generalizability of the evaluation results.
  - Resource-Intensive: The large-scale dataset and comprehensive evaluation framework of BIG-Bench require significant computational resources, which may limit its accessibility for certain researchers or organizations.

Table 1
Comparison of various benchmarks.

Benchmark	Key features	Number of tasks/datasets	Task diversity	Context awareness	Persona consistency
GLUE/SuperGLUE	Comprehensive, diverse tasks	9 for GLUE, 8 for SuperGLUE	Yes	No	No
SQuAD	Contextual understanding	2	No	No	No
CoQA	Conversation history awareness	1	No	Yes	No
Persona-Chat	Persona consistency	1	No	No	Yes
DSTC	Variety of dialogue tasks	Varies annually	Yes	Yes	No
BIG-Bench	Comprehensive evaluation across language tasks open-source	3	No	No	No
HEML	Assessment of contextual understanding and reasoning challenging tasks	2	Yes	Yes	Yes
MMLU	Evaluation across multiple languages and domains standardized metrics	1	No	No	No

2.1.7. Holistic evaluation of language models (HELM)

HELM aims to evaluate language models by assessing their contextual understanding and reasoning abilities. It focuses on designing challenging tasks that require deep comprehension, including linguistic nuances, common sense, and logical reasoning [22]. HELM incorporates evaluations in multiple languages to ensure linguistic diversity and cross-lingual evaluation. It provides an open evaluation platform for researchers to compare their models against state-of-the-art models.

- Strengths
  - Emphasis on Contextual Understanding: HELM focuses on assessing the contextual understanding and reasoning abilities of language models by designing challenging tasks that require deep comprehension.
  - Linguistic and Commonsense Knowledge Evaluation: It incorporates evaluation metrics that measure the model's understanding of linguistic nuances, common sense, and logical reasoning, providing a holistic assessment.
  - Linguistic Diversity: HELM includes diverse evaluation tasks that cover multiple languages, ensuring that the benchmark is not limited to English-centric evaluations.
  - Open Evaluation Platform: HELM provides an open evaluation platform, enabling researchers to submit their models and compare their performance against state-of-the-art models.
- Weakness
  - Limited Coverage of Language Tasks: HELM may not cover the full spectrum of language tasks, focusing more on the contextual understanding aspect. This may restrict its applicability to specific evaluation scenarios.
  - Evaluation Complexity: The evaluation tasks designed in HELM can be complex, requiring advanced linguistic and reasoning capabilities, which may pose challenges for models that are not specifically trained for such tasks.
  - Reliance on Human Annotations: Some HELM tasks may require human annotations or human evaluations, which could introduce subjectivity and potential biases in the evaluation process.

2.1.8. Multilingual multi-domain language understanding (MMLU)

MMLU focuses on evaluating language models across multiple languages and domains [23]. It covers evaluations in various domains, including news, e-commerce, and conversational data. MMLU incorporates languages from different language families to promote cross-lingual evaluation and linguistic diversity. The benchmark employs standardized evaluation metrics for fair comparisons between different language models. Table 1 compares various benchmarks.

- Strengths
  - Multilingual Evaluation: MMLU focuses on evaluating language models across multiple languages, providing insights into the models' performance on a global scale.
  - Cross-Domain Evaluation: It covers evaluations in various domains, including news, e-commerce, and conversational data, ensuring a diverse assessment of models' performance in different contexts.
  - Linguistic Diversity: MMLU incorporates languages from different language families, increasing the coverage of languages and promoting cross-lingual evaluation.
  - Standardized Evaluation Metrics: MMLU employs standardized evaluation metrics, allowing for fair comparisons between different language models.
- Weakness
  - Limited Task Coverage: MMLU may not cover all possible language tasks, potentially missing some specialized tasks or domains that require specific evaluation criteria.
  - Dependency on Available Multilingual Data: The evaluation in MMLU heavily relies on the availability of multilingual data, which may limit the scope of evaluation for certain language pairs or low-resource languages.
  - Potential Dataset Bias: The dataset used in MMLU may exhibit biases based on the sources and collection methods, which can impact the fairness and generalizability of the evaluation results.

2.1.9. Openai's guidelines

OpenAI's Guidelines encompass principles related to AI behavior, safety, broad access, and long-term robustness [24]. These guidelines provide a framework for responsible AI development and deployment.

- Strengths
  - Comprehensive Framework: OpenAI's Guidelines offer a holistic approach to AI development, considering ethical implications and long-term impact.
  - Societal Considerations: The guidelines emphasize the importance of fairness, safety, and avoiding undue concentration of power.
- Weakness
  - OpenAI-specific: The guidelines are specific to OpenAI's approach and may not directly apply to other organizations or models.
  - Balancing Trade-offs: Implementing all the principles may require difficult trade-offs between competing priorities.

2.1.10. IEEE's Ethically Aligned Design

IEEE's Ethically Aligned Design provides a set of principles, recommendations, and guidelines for ethically aligned AI development [25]. It emphasizes the importance of ensuring AI systems align with ethical values and human rights.

Table 2
Comparison of various standards.

Standard	Key principles	Developed by	Applicable to	Adoption	Trade-off considerations	Ethical considerations
OpenAI's Guidelines	AI behavior, safety	OpenAI	AI and AGI	Used by OpenAI	Yes	Yes
IEEE's Ethically Aligned Design	Ethical AI implementation	IEEE	AI and AIS	Used worldwide	Yes	Yes
Montreal Declaration	Well-being, Autonomy, Justice	University of Montreal	AI and AIS	Endorsed by organizations	Yes	Yes
Partnership on AI's Tenets	Cooperation, Safety, Fair Access	Partnership on AI	AI and AIS	Endorsed by partners	Yes	Yes

- Strengths
  - Ethical Framework: IEEE's document offers a comprehensive framework for designing AI systems that prioritize ethical considerations.
  - Wide Adoption: The standards have gained recognition and are widely adopted across industries and research communities.
- Weakness
  - High-level Principles: The principles provided by IEEE are general, requiring interpretation and adaptation to specific AI systems and applications.
  - Balancing Ethical Considerations: Incorporating all ethical principles may involve challenging trade-offs and complex decision-making.

2.1.11. The Montreal Declaration for responsible AI

The Montreal Declaration presents a comprehensive ethical framework for AI development [26]. It outlines principles such as respect for autonomy, protection of privacy, and promotion of well-being.

- Strengths
  - Holistic Ethical Approach: The Montreal Declaration covers a broad range of ethical considerations, promoting responsible AI development.
  - Broad Endorsement: The declaration has received endorsements from various organizations, fostering awareness and acceptance of responsible AI practices.
- Weakness
  - High-level Guidance: The principles may require further elaboration and contextualization to ensure practical application in different AI domains.
  - Potential Conflicts: Balancing multiple ethical principles may lead to conflicts when implementing AI systems.

2.1.12. Partnership on AI's tenets

The Partnership on AI's Tenets outlines principles for cooperation, safety, fairness, and broad access to AI technology [27]. It highlights the importance of addressing societal challenges and promoting responsible AI practices. Table 2 compares various standards.

- Strengths
  - Cooperative Approach: The tenets encourage collaboration among stakeholders to ensure responsible and beneficial AI development.
  - Focus on Safety and Fairness: The principles emphasize safety measures, unbiased research, and inclusive deployment of AI technologies.
- Weakness
  - Trade-off Considerations: Implementing all tenets may involve complex trade-offs, as some principles might conflict with each other in specific scenarios.
  - Broad Interpretation: The tenets' high-level nature requires further clarification and guidance for practical implementation.

2.2. Evaluation criteria

This subsection discusses common evaluation methods used in conversational AI, including BLEU [28], ROUGE [29], METEOR [30], precision–recall, F1 score, perplexity, and user feedback.

2.2.1. BLEU, ROUGE, and METEOR

Bilingual Evaluation Understudy (BLEU) is commonly used for evaluating the quality of machine translation or text generation. It compares the n-gram overlap between the generated text and one or more reference texts as shown in Eq. (1).

$$BLEU = BP * exp(sum(wi * \log(\pi))) \tag{1}$$

- BP (Brevity Penalty) is a penalty term that accounts for the difference in length between the generated and reference texts.
- π is the modified n-gram precision, which measures the ratio of n-grams in the generated text that appear in the reference text.
- wi is the weight assigned to each n-gram precision.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is used for evaluating the quality of text summarization or document similarity. It measures the overlap of n-grams, word sequences, and other features between the generated summary and the reference summary. Eqs. (2) and (3) present the formulations of ROUGE-N and ROUGE-L.

$$ROUGE - N = \frac{CN}{TN} \tag{2}$$

$$ROUGE - L = \frac{LCS}{TNW} \tag{3}$$

- CN: Count of overlapping N-grams, TN: Count of N-grams in the reference summary, LCS: Longest Common Subsequence, TNW: Total Number of Words in the reference summary
- ROUGE-N calculates the precision of n-gram matches between the generated and reference summaries.
- ROUGE-L measures the longest common subsequence between the generated and reference summaries.

Metric for Evaluation of Translation with Explicit Ordering (METEOR) is another metric commonly used for evaluating machine translation or text generation as shown in Eq. (4). It incorporates measures of precision, recall, and alignment errors, along with stemming and synonymy matching.

$$METEOR = (1 - \alpha) * P + \alpha * R * (1 - P) \tag{4}$$

- P: Precision measures the ratio of matching unigrams between the generated and reference texts.
- R: Recall measures the ratio of matching unigrams in the generated text against the reference text.
- Penalty penalizes the generated text for incorrect word order or alignment errors.
- α is a parameter that controls the trade-off between precision and recall.

BLEU, ROUGE, and METEOR are widely used metrics for evaluating machine translation and text summarization. They compare the model-generated output to human-generated reference texts.

- Strengths
  - Quantitative Assessment: These metrics provide quantitative measures of model performance, facilitating objective evaluation.
  - Scalability: BLEU, ROUGE, and METEOR can be automatically computed, enabling efficient evaluation across large datasets.
- Weakness
  - Limitations in Capturing Quality: While useful for assessing certain aspects of text generation, these metrics may not capture all aspects of text quality, such as coherence or relevance.
  - Reference Dependency: The choice of reference texts may not always represent the only correct or best possible output.

2.2.2. Precision, recall, F1 score

Precision, recall, and the F1 score are evaluation metrics commonly used in information extraction, question answering, and other tasks. Precision measures the proportion of true positives among all identified entities as shown in Eq. (5), while recall measures the proportion of true positives among all actual positives as shown in Eq. (6). The F1 score is the harmonic mean of precision and recall as shown in Eq. (7).

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{5}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{6}$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{7}$$

- True Positives represent the number of correctly identified instances.
- False Positives represent the number of incorrect instances identified as positive.
- False Negatives represent the number of missed instances.
- Precision measures the proportion of true positives among all identified instances.
- Recall measures the proportion of true positives among all actual positive instances.
- F1 Score is the harmonic mean of precision and recall, providing a single metric that balances the two.

- Strengths
  - Quantitative Assessment: Precision, recall, and the F1 score provide quantitative measures of model performance, enabling comparison and benchmarking.
  - Balance between False Positives and Negatives: The F1 score considers both precision and recall, allowing trade-offs between false positives and false negatives.
- Weakness
  - Task-Specific Limitations: These metrics may not fully represent model performance for tasks that require more nuanced evaluation criteria.
  - Optimal Balance Variation: The optimal balance between precision and recall may vary depending on the specific application or task.

2.2.3. Perplexity

Perplexity is a measure of how well a probability model predicts a sample. In the context of language models, a lower perplexity score indicates better performance. Perplexity is a metric commonly used to evaluate the performance of language models, including conversational AI systems. It measures how well a language model predicts a given sequence of words. Perplexity is calculated based on the probability distribution of the language model. A lower perplexity score indicates that the language model can better predict the next word in a sequence and, therefore, has better performance. The formula for perplexity is as follows and shown in Eq. (8):

$$Perplexity = 2^{-\frac{\log P(w)}{N}} \tag{8}$$

- $P(w)$ represents the probability of the word sequence given by the language model.
- N represents the number of words in the sequence.

In essence, perplexity measures how surprised the language model would be to see the actual word sequence. A lower perplexity score suggests that the language model is more certain and accurate in its predictions. To use perplexity in evaluation, the language model is typically trained on a large dataset and then tested on a separate evaluation dataset. The perplexity score is calculated by applying the formula to the evaluation dataset. Lower perplexity scores indicate better performance and a better ability of the language model to predict the next word accurately. It is important to note that perplexity is often used as an internal evaluation metric during the training and fine-tuning of language models. While it provides a quantitative measure of how well the model fits the training data, it may not always directly correlate with the overall quality or coherence of generated text. Therefore, perplexity is usually used in conjunction with other evaluation methods, such as human evaluation or task-specific metrics, to get a more comprehensive understanding of the language model's performance.

1. Strengths
   (a) Interpretable Measure: Perplexity provides a single, interpretable measure of language model performance.
   (b) Automatic Computation: Perplexity can be calculated automatically, enabling scalable evaluation across large datasets.
2. Weakness
   (a) Limited Text Quality Representation: Perplexity may not always correlate with qualitative measures of text quality, such as coherence or relevance.
   (b) Assumption of Data Distribution: It assumes that the test data follows the same distribution as the training data, which may not always hold true in real-world scenarios.

2.2.4. User feedback

User feedback provides a qualitative evaluation of a conversational AI system. It involves collecting feedback from users regarding their satisfaction, engagement, and overall experience with the system. Table 3 presents the comparison of various evaluation criteria. Here are some steps to effectively utilize user feedback for evaluation.

- Design Feedback Collection Mechanisms: Implement mechanisms that allow users to provide feedback easily. This can include in-app rating systems, feedback forms, surveys, or even direct interaction with users through interviews or focus groups.
- Define Evaluation Goals: Clearly define the evaluation goals and the specific aspects of the system that you want to assess with user feedback. This could include factors like system responsiveness, accuracy of responses, naturalness of conversation, or overall user satisfaction.

Table 3
Comparison of various evaluation criteria.

Evaluation method	Used in tasks	Key feature	Nature	User involvement	Scalability	Subjectivity	Quantitative
BLEU, ROUGE, METEOR	Translation, Summarization	Text overlap, precision, recall	Quantitative	No	Yes	Low	Yes
Precision, Recall, F1 Score	Information Extraction, QA	True positive, false positive, false negative	Quantitative	No	Yes	No	Yes
Perplexity	Language modeling	Language model perplexity	Quantitative	No	Yes	No	Yes
User Feedback	Conversational AI systems, user satisfaction	Subjective user satisfaction scores	Qualitative	Yes	Yes	Yes	No

- Gather Structured and Unstructured Feedback: Collect both structured and unstructured feedback from users. Structured feedback can be in the form of ratings, rankings, or Likert scale responses, while unstructured feedback can include open-ended comments or suggestions. Structured feedback provides quantifiable metrics, while unstructured feedback captures nuanced insights.
- Analyze Quantitative Metrics: Analyze structured feedback to gather quantitative metrics. This can involve calculating averages, aggregating ratings, or analyzing trends over time. These metrics can provide a quantifiable understanding of user satisfaction or specific aspects of the system's performance.
- Analyze Qualitative Insights: Analyze unstructured feedback to extract qualitative insights. This involves categorizing and summarizing user comments, identifying recurring themes or issues, and extracting actionable insights. Qualitative feedback provides rich context and helps identify areas for improvement.
- Triangulate Feedback with Other Evaluation Measures: Combine user feedback with other evaluation measures, such as performance metrics or task-specific assessments. This helps gain a comprehensive understanding of the system's performance and identifies correlations between user feedback and objective measures.
- Iterative Improvement: Use user feedback as a basis for iterative improvement. Identify areas where the system falls short or where user satisfaction can be enhanced, and prioritize enhancements accordingly. Regularly incorporate user feedback into the system's development cycle to drive continuous improvement.
- Address User Concerns: Actively address user concerns and issues raised through feedback. Communicate updates, improvements, or resolutions to users to demonstrate responsiveness and maintain user trust.
- Engage Users in Co-creation: Engage users in the co-creation process by involving them in feedback-driven feature prioritization, design decisions, or beta testing. This fosters a sense of ownership, enhances user satisfaction, and ensures the system aligns with user expectations.

- Strengths
  - User-Centric Assessment: User feedback captures the subjective experience and satisfaction, providing valuable insights into system performance.
  - Comprehensive Evaluation: User feedback encompasses aspects that may not be fully captured by quantitative metrics, such as the system's naturalness and overall user experience.

- Weakness
  - Resource-Intensive: Collecting and analyzing user feedback can be time-consuming and resource-intensive, requiring dedicated efforts.
  - Subjectivity and Variability: User feedback can be subjective, and opinions may vary among users, making it challenging to generalize the evaluation results.

2.2.5. New why metrics extend beyond traditional measures?

Traditional evaluation metrics for conversational AI systems, such as BLEU, ROUGE, and F1 scores, are extremely valuable for assessing the accuracy of generated responses based on reference responses. However, these measures do not fully capture some crucial aspects of conversation quality, such as context-sensitivity, dialogue coherence, user satisfaction, and the relevance of responses.

For instance, context-sensitivity refers to the AI's ability to adapt its responses based on the conversational context. This aspect cannot be properly captured by traditional metrics, which evaluate responses independently of the conversational context. Therefore, we propose the Contextual Sensitivity Index (CSI) to quantitatively assess the AI's ability to adjust its responses based on the conversation context.

Dialogue coherence is another important aspect often overlooked by traditional metrics. A conversation should maintain a logical and meaningful flow. To evaluate this, we propose a Dialogue Coherence Measure, which can quantify the degree of coherence in the conversation flow.

User satisfaction is one of the ultimate goals of any conversational AI system. Traditional metrics often fall short in capturing the subjective experience of users. By incorporating user feedback and human evaluation into our framework, we can gather insights into user satisfaction and the perceived quality of conversations.

Lastly, the relevance of responses is another crucial aspect. A response may be grammatically correct and similar to reference responses (resulting in high scores in traditional metrics) but may still be irrelevant or inappropriate in a given context. To capture this, we propose a Relevance Measure, which assesses the pertinence of generated responses.

While we recognize that some of these measures have been used in other contexts or for specific tasks, our proposed framework integrates them into a comprehensive evaluation system for conversational AI. The combination of these measures provides a more nuanced and holistic evaluation of the AI's performance, filling gaps left by traditional metrics. We hope this clarifies the need for these "beyond traditional" measures in our proposed framework.

Let us delve deeper into why the proposed metrics extend beyond traditional measures in the context of evaluating conversational AI systems.

- **Contextual Sensitivity Index (CSI):** Traditional metrics are inherently context-agnostic. They measure the linguistic closeness of the generated response to a pre-determined "gold standard" response. However, this fails to capture an essential attribute of natural conversations — the context-dependency. Conversations are not merely exchanges of information but are deeply influenced by the context they are embedded in. Therefore, it is crucial to assess a model's capability of being sensitive to the context, a factor traditional measures do not address. CSI, as we propose, quantifies this context sensitivity. It can detect if the model appropriately adjusts its responses to changes in the context, such as alterations in topic, sentiment, or nuances introduced by the user. For instance, in a support chat scenario, if the user goes from asking about a product's feature to expressing frustration about it,

the model should adjust its responses accordingly, demonstrating empathy and providing assistance.
The CSI might be a normalized score that compares a model's responses in different contexts.

$$CSI = \frac{f(\text{Contextual Response Variation})}{g(\text{Contextual Stimuli Variation})} \quad (9)$$

Here, f() could be a function that measures the degree of variation in the model's responses given a change in the contextual stimuli. g() could be a function quantifying the variation in the contextual stimuli.
Strengths

- CSI can capture a model's ability to adapt its responses to the changes in context.
- It focuses on a crucial aspect of conversational AI that traditional metrics overlook: context sensitivity.

Weakness

- Determining an appropriate measure of "contextual variation" might be challenging.
- Some elements of context might be subtle or hard to quantify.

- **Dialogue Coherence Measure:** Conversations are not random sequences of exchanges but follow a certain logic or flow. They are coherent narratives. While traditional metrics might capture fluency or grammatical correctness, they are ill-equipped to assess the conversational coherence over extended dialogues. We propose a dialogue coherence measure that goes beyond sentence-level assessment and looks at the conversation as a whole, from the start to the current utterance, capturing both local and global coherence.
This measure could assess both local (adjacent turn-to-turn) and global (entire conversation) coherence.

$$\text{Coherence Score} = \alpha * \text{LC} + \beta * \text{GC} \quad (10)$$

LC: Local coherence could be quantified as the semantic similarity between adjacent utterances.
GC: Global coherence could be quantified by considering the semantic drift over the course of the conversation.
Strengths

- It takes into account the entire conversational flow, not just individual utterances.
- It can capture the logical consistency and progression of a conversation.

Weakness

- Deciding on suitable weights (α and β) for local and global coherence might be tricky.
- Semantic drift computation could be computationally heavy for long conversations.

- **User Feedback and Human Evaluation:** Traditional metrics are quantitative, automated, and lack the human touch. They do not factor in the user's perception of the conversation or subjective experience, which is crucial as the ultimate aim of conversational AI is to engage and assist humans effectively. This is where user feedback and human evaluation play a key role. Users can provide insights into factors traditional metrics cannot perceive: Did they find the conversation engaging? Did the AI understand and satisfy their intent? Did they find the response natural, empathetic, or creative, even if it deviated from standard responses?
This metric could be an aggregate score of different facets of user feedback.

$$\text{User Score} = \Sigma[w_i * X_i] \quad (11)$$

Here, Xi could represent different aspects of user feedback (like satisfaction, understanding, helpfulness), and wi could be their corresponding weights.
Strengths

- It is a direct measure of user satisfaction, which is the ultimate goal of a conversational AI.
- It can capture aspects like naturalness, empathy, and creativity that automated metrics may miss.

Weakness

- User feedback might be subjective and could vary widely between individuals.
- Collecting and analyzing user feedback can be resource-intensive.

- **Relevance Measure:** Linguistic closeness to a reference response does not necessarily equate to relevance. A response could be grammatically correct and align well with a reference response yet be entirely irrelevant to the conversation at hand. Therefore, a relevance measure is crucial to assess how well a model's responses align with the current context and the user's needs and expectations. It goes beyond the myopic view of traditional metrics and looks at the bigger picture — the conversation's goal. The Relevance Measure can assess how closely the AI's responses align with the content and purpose of the preceding conversational turns. It is crucial to ensure that the AI does not deviate significantly from the topic, which would make the conversation feel disjointed and reduce user satisfaction.
This measure assesses how closely the AI's responses align with the content and intent of the preceding conversational turns.

$$RM = \frac{\text{Number of Relevant Responses}}{\text{Total Number of Responses}} \quad (12)$$

Strengths

- It directly evaluates how well the AI maintains the context and stays on topic.
- It can prevent the AI from generating off-topic responses, which are a common problem in chatbot conversations.

Weakness

- The definition of "relevance" can be subjective and may differ across various conversational contexts.
- Some conversations might require the AI to shift topics appropriately, which this metric might penalize.

- **Task Success Rate (TSR):** The Task Success Rate is a vital metric when conversational AI systems are designed to perform specific tasks, such as answering customer inquiries, booking appointments, or making reservations. This metric provides a direct measure of the system's ability to complete these tasks correctly and is a clear indicator of the system's practical value to users.
TSR is a crucial measure for task-oriented conversational AI systems. It provides a direct measure of the AI's ability to correctly complete the assigned tasks.

$$TSR = \frac{\text{Number of Successful Tasks}}{\text{Total Number of Tasks}} \quad (13)$$

Strengths

- It directly measures the system's ability to perform its intended function.
- It is straightforward to calculate and understand.

Weakness

- The definition of "task success" can vary widely across different types of tasks and may be hard to standardize.

Table 4
Comparison of benchmarks, standards, and evaluation Criteria between ChatGPT and other AI models.

Parameters	ChatGPT	Other GPT/Deep learning models
Benchmark Purpose	Assess conversational performance and interactivity	Measure task-specific performance and capabilities
Focus Areas	Coherence, context maintenance, multi-turn dialogue	Task-specific metrics, accuracy, precision, recall
Task Diversity	Multiple conversational benchmarks	Task-specific benchmarks (e.g., translation, QA, etc.)
Persona Consistency	Assessing persona adherence and consistency	Not applicable to most models
Ethical Considerations	Evaluating bias mitigation, responsible behavior	General ethical guidelines and data handling practices
Standards Purpose	Define guidelines and principles for conversational AI	General technical and ethical standards
Development Organizations	OpenAI, research community, industry stakeholders	Research community, organizations, standards bodies
Applicability	Conversational AI systems	Broad range of deep learning models and applications
Trade-off Considerations	Balancing user experience, performance, and ethics	Model complexity, training data requirements, fairness
Evaluation Criteria Flexibility	Customized for conversational characteristics	Task-specific evaluation metrics and benchmarks
User-Centric Evaluation	User satisfaction, engagement, interaction quality	Task-specific performance, accuracy, user feedback
Adaptability to New Challenges	Dynamic evaluation criteria for emerging needs	May require updates for new tasks or problem domains

– It does not capture the quality of the system's interactions with users outside the context of task completion.

In summary, these metrics and methods stretch beyond the traditional metrics' capacity to evaluate a conversation's quality, offering a more comprehensive understanding of the model's conversational competence. While some of these measures might have been used in some contexts, their use in evaluating conversational AI is relatively new. Their integration into our proposed framework represents a major step forward in the development of more holistic, nuanced, and user-centric evaluation methodologies.

3. Insight of benchmarks, standards and evaluations of ChatGPT

3.1. Differences between ChatGPT and other AI models

Benchmarks, standards, and evaluation criteria for ChatGPT may differ from those used for other GPT or deep learning models due to the specific nature of conversational AI systems. Here's how they differ [31,32].

- **Benchmarks:** ChatGPT benchmarks focus on assessing the performance and capabilities of the model in conversational scenarios, emphasizing factors like interactivity, coherence, and context maintenance. Traditional benchmarks for other GPT or deep learning models may focus on specific tasks like machine translation, question answering, or sentiment analysis, with less emphasis on the dynamic and interactive nature of conversations.
- **Standards:** Standards for ChatGPT encompass guidelines and principles specific to conversational AI, addressing ethical considerations, user experience, and responsible behavior in interactive dialogue systems. Standards for other GPT or deep learning models may focus on general ethical guidelines, technical performance metrics, or data handling practices, but may not explicitly address the complexities and challenges unique to conversational AI.
- **Evaluation Criteria:** Evaluation criteria for ChatGPT emphasize aspects such as context awareness, persona consistency, user satisfaction, and relevance in multi-turn conversations. Evaluation criteria for other GPT or deep learning models may focus on metrics like accuracy, precision, recall, or F1 score, typically measured on specific tasks or datasets.

The key distinction lies in the specific requirements and characteristics of conversational AI systems like ChatGPT, which necessitate tailored benchmarks, standards, and evaluation criteria. Conversational AI places emphasis on interactivity, contextual understanding, user experience, and ethical considerations that differ from the task-specific evaluation used for other deep learning models. Here's an expanded comparison table that provides more details and parameters for comparing benchmarks, standards, and evaluation criteria between ChatGPT and other GPT or deep learning models. Table 4 shows the comparison of benchmarks, standards, and evaluation Criteria between ChatGPT and other AI models.

3.2. Key issues in evaluation criteria

We present some existing challenges in the evaluation of conversational AI systems like ChatGPT, along with specific points that highlight the complexities and considerations involved [33].

- **Contextual Understanding:** Capturing and maintaining context across multiple turns of conversation, resolving coreferences and handling ambiguous or implicit references, and understanding and addressing user intent and nuanced queries.
- **Coherence and Relevance:** Ensuring that the generated responses remain coherent and relevant throughout a conversation, aligning with the user's expectations and intent, and avoiding generic or nonsensical responses that do not address the user's query.
- **Bias and Fairness:** Detecting and mitigating biases in the generated responses, ensuring fairness and equitable treatment across different user demographics, and avoiding the propagation of harmful stereotypes or discriminatory views.
- **Ethical Considerations:** Protecting user privacy and responsibly handling sensitive information, avoiding the generation of offensive, harmful, or misleading content, and ensuring transparency in communicating the system's capabilities and limitations.
- **Evaluation Metrics:** Developing comprehensive metrics that capture the nuances of conversational AI, incorporating both quantitative and qualitative measures to assess system performance, and striking a balance between different evaluation criteria to provide a holistic assessment.
- **User Engagement and Satisfaction:** Maintaining user engagement throughout a conversation, providing responses that are not only accurate but also engaging and natural, and ensuring user satisfaction by meeting their expectations and preferences.
- **Robustness and Error Handling:** Handling out-of-scope queries and gracefully responding to unsupported requests, detecting and addressing cases where the model generates incorrect or nonsensical responses, and effectively managing errors or misinterpretations during the conversation.
- **Scalability and Generalization:** Ensuring that the system's performance generalizes well to unseen scenarios, scaling the system to handle high volumes of concurrent conversations, and evaluating its performance on diverse datasets and real-world use cases.
- **User-Centered Design:** Incorporating user feedback and involving users in the evaluation process, designing systems that adapt to individual user preferences and needs, and striking a balance between system capabilities and user expectations to optimize the user experience.
- **Real-Time Interaction:** Enabling fast and seamless responses in real-time conversations, minimizing latency to ensure timely interaction for a smooth user experience, and managing the computational requirements for real-time response generation.

3.3. Adaptability aspects for ChatGPT

Achieving adaptability in the creation of new benchmarks, standards, and evaluation criteria for ChatGPT involves considering the dynamic nature of the field, evolving requirements, and emerging challenges. Here are some key aspects to consider for achieving adaptability [34].

- **Flexibility in Design:** To promote adaptability, benchmarks, standards, and evaluation criteria should be designed with flexibility in mind. This includes accommodating future changes and advancements by allowing for iterative updates and revisions based on emerging research, user feedback, and evolving needs. Incorporating modularity in the design enables easy addition or modification of evaluation components as the field progresses.
- **Community Collaboration:** Collaboration among researchers, developers, and stakeholders is essential for adaptability. Foster an environment of open discussions, knowledge sharing, and participation to collectively define benchmarks, standards, and evaluation criteria. Establish community-driven processes to gather input, incorporate diverse perspectives, and validate the proposed criteria, ensuring that they reflect the needs and requirements of the wider community.
- **Engagement with User Feedback:** User feedback plays a crucial role in creating adaptable benchmarks, standards, and evaluation criteria. Actively seek and incorporate user perspectives to ensure that the criteria align with their needs, expectations, and desired outcomes. Regularly assess and integrate user feedback to refine and update the benchmarks and evaluation protocols, making them more relevant and effective.
- **Consideration of Emerging Challenges:** Staying informed about emerging challenges and novel use cases in conversational AI is essential for adaptability. Continuously evaluate the relevance of existing benchmarks and standards, identifying gaps and new requirements. Proactively address ethical, fairness, and privacy concerns that arise as conversational AI systems evolve, ensuring that the criteria are responsive to the changing landscape.
- **Iterative Improvement:** Approach the creation of benchmarks, standards, and evaluation criteria as an iterative process. Gather feedback from researchers, developers, and the wider community to refine and enhance the criteria over time. Embrace a growth mindset that welcomes continuous improvement as new insights and techniques emerge, keeping the benchmarks and standards up-to-date and reflective of the latest advancements.
- **Regular Updates and Versioning:** Establish mechanisms for regular updates and versioning of benchmarks, standards, and evaluation criteria. Release new versions that incorporate feedback, address limitations, and adapt to the evolving landscape. Transparently communicate updates and changes to the wider community, ensuring that stakeholders are aware of the latest developments and can align their practices accordingly.
- **Balance Consistency and Flexibility:** Strive for consistency in evaluation methodologies to enable benchmarking and comparison across different systems. However, strike a balance between consistency and flexibility to accommodate diverse use cases, domains, and emerging challenges. Allow for customization and adaptation of evaluation criteria based on specific application requirements, enabling the benchmarks and standards to cater to a wide range of needs and contexts.

3.4. Proposed framework

Our proposed evaluation framework for ChatGPT is a six-layered and comprehensive model that accounts for a wide range of evaluation criteria from task-specific to human-based assessments. This robust evaluation framework is necessary to ensure that the AI system is capable of understanding and responding to human queries, maintaining coherence, providing relevant and accurate responses, and upholding ethical considerations. Here is a deeper dive into each layer of the evaluation framework.

3.4.1. Background

The following section presents a theoretical proposal for a benchmarking and evaluation framework specifically developed for ChatGPT. At present, this is a conceptual proposition and not a hands-on implementation. We have formulated the proposed framework with the intention of laying the groundwork for future practical applications and developments in the field of Conversational AI. Our framework comprises of diverse evaluation tasks and standards, which are representative of a wide array of potential use cases. We acknowledge that the scope of Conversational AI is vast and continuously evolving; hence, our proposal is not exhaustive but focuses on the most critical aspects of this field. We believe that our proposed framework, with its structured evaluation tasks and progressive standards, could offer valuable insights to guide the responsible and effective development and deployment of conversational models like ChatGPT. We also emphasize that our framework is inherently adaptable to incorporate future advancements and emerging trends in AI. With this flexible design, it can continue to serve as a reliable guide, reflecting and addressing the ever-changing landscape of AI technologies. Fig. 1 presents the architecture of the 6-layered proposed framework.

3.4.2. A multi-dimensional approach of proposed framework

- **Task-Oriented Benchmarks:** These are tasks specifically designed to test various capabilities of ChatGPT. This category is broken down into further subsections. Task-specific benchmarks focus on evaluating the performance of a conversational AI system on specific predefined tasks or domains. These benchmarks are designed to assess the system's ability to understand and generate responses relevant to the given task. Examples of task-specific benchmarks include question-answering datasets like SQuAD or translation datasets like Workshop on Machine Translation (WMT).
  - **Factual Understanding:** Tasks testing the ability to understand and generate factual information. For instance, questions about historical events, scientific concepts, or general knowledge.
  - **Contextual Understanding:** Tasks to evaluate the model's ability to grasp and maintain context over a conversation. An example could be a sequence of questions where each question relies on information from the previous one.
  - **Coherence:** Tasks to assess whether the model maintains coherence in long responses or over a long conversation. Ambiguity Resolution: Tasks that test the system's ability to handle ambiguous queries and requests.

  Measuring task-oriented benchmarks can involve following techniques.

  - **Accuracy:** Measure the percentage of correctly answered questions or tasks.
  - **F1 Score:** Compute the harmonic mean of precision and recall, particularly used in question answering benchmarks.
  - **BLEU:** Measure the quality of machine-generated translations by comparing them to reference translations.
  - **ROUGE:** Assess the quality of machine-generated summaries by comparing them to reference summaries.
  - **METEOR:** Evaluate the quality of machine-generated translations by considering precision, recall, and alignment.

  Strength

  - Task-oriented benchmarks focus on evaluating the performance of a conversational AI system on specific predefined tasks or domains.

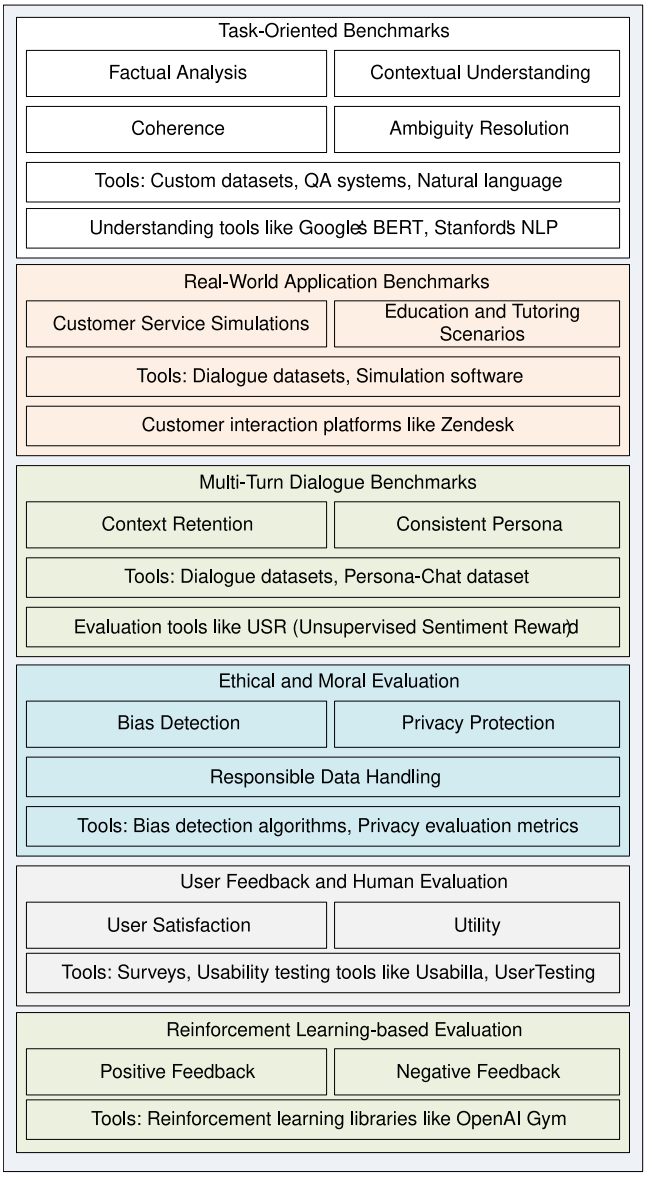

Fig. 1. Proposed framework for ChatGPT benchmark evaluation.

- They provide a clear and measurable evaluation criterion based on the successful completion of the task.
- These benchmarks help assess the system's ability to understand and generate responses relevant to the given task.
- They enable direct comparison and evaluation of different systems on the same task.

Weakness

- Selecting the right set of task-oriented benchmarks can be challenging due to the diverse range of tasks and domains.
- It may be difficult to capture the full complexity of real-world tasks within a benchmark, leading to potential limitations in generalizability.
- Designing high-quality task-specific datasets for benchmarking can be time-consuming and resource-intensive.

• **Real-World Application Benchmarks:** These involve creating real-world scenarios for evaluation. Such benchmarks aim to evaluate the system's performance in real-world scenarios, where the conversations are more diverse and dynamic. These benchmarks simulate realistic conversational settings and evaluate the system's ability to handle complex interactions, maintain context, and provide appropriate responses. Real-world application benchmarks often involve more open-ended conversations, such as customer support dialogues or interactive chat sessions. A few examples include.

- **Customer Service Simulations:** ChatGPT should be able to manage a customer's request, provide accurate information, and offer a satisfactory resolution. Evaluation can be based on success rate, resolution time, and customer satisfaction.
- **Education and Tutoring Scenarios:** Tasks could involve explaining complex concepts, answering educational queries, and interacting in a pedagogically effective manner.

Measuring real-world application benchmarks can involve following techniques.

- **Human Evaluation:** Conduct manual assessments where human judges rate the quality of the system's responses based on criteria such as relevance, fluency, and appropriateness.
- **User Feedback:** Collect feedback from real users who interact with the system in real-world scenarios, such as customer support or chat applications.
- **Contextual Coherence:** Measure the system's ability to maintain context and coherence throughout a conversation by evaluating the flow and continuity of dialogue exchanges.
- **Relevance:** Assess the relevance of the system's responses to the specific user queries or prompts in a real-world context.
- **User Satisfaction:** Gather user ratings or feedback to gauge their overall satisfaction and experience with the system.

In determining representative real-world tasks or applications to serve as benchmarks, we propose a three-pronged approach:

- **User-Driven Selection:** The first step involves identifying the primary ways users interact with ChatGPT. This involves extensive user studies, surveys, and data analysis to discern the most common and critical use-cases. For instance, if most users engage with ChatGPT for drafting emails or generating text, then benchmarks should include tasks that directly assess these functions. On the other hand, if users frequently engage in conversational dialogues with ChatGPT, the benchmarks should reflect tasks that measure its performance in conversational understanding and generation.
  - **Survey and User Studies:** To identify the main interactions users have with ChatGPT, we can conduct extensive surveys and user studies. These could involve a combination of quantitative and qualitative methods. For instance, users could be asked to rank various use-cases of ChatGPT according to their frequency of usage. They could also be invited to participate in focus groups or interviews to share their experiences and expectations in more detail.
  - **Data Analysis:** ChatGPT interaction data (while respecting user privacy and data protection norms) could be analyzed to identify common patterns and tasks. Techniques such as data mining, clustering, and sequence analysis could help identify frequent user interactions. Natural language processing techniques can also be applied to the data to extract patterns and insights, providing a rich source of user-driven tasks.

- **Advanced Analytical Techniques:** Employing methods such as sentiment analysis, latent semantic analysis, or topic modeling to the user interaction data can reveal not only common tasks but also users' attitudes towards those tasks. Additionally, machine learning techniques such as association rule learning can help discover relationships between different types of interactions, revealing more complex tasks or task sequences.
- **User Persona Creation:** By developing user personas, or representations of different types of users based on behavior patterns, needs, motivations, and goals, we can derive an understanding of the needs and wants of different user groups. This will guide us towards representative tasks that cater to a wide range of users.

Diversity of Benchmarks: The evaluation should not be limited to one or two tasks that represent the most common uses. It should include a wide variety of tasks that measure different aspects of the system's capabilities. This could include tasks like question answering (to test understanding), text generation (to test creative abilities), summarization (to test conciseness), and translation (to test language capabilities). By selecting diverse tasks, we can create a more holistic view of ChatGPT's performance and versatility.

- **Cognitive Task Diversity:** The selected tasks should not only reflect user interactions but also encompass a wide spectrum of cognitive capabilities. For instance, question answering tasks measure the system's understanding and reasoning abilities. In contrast, text generation tasks test the system's creativity and coherence. Having a diverse set of tasks would ensure a comprehensive evaluation.
- **Domain-Specific Benchmarks:** Given the wide-ranging applications of ChatGPT, it is also crucial to include domain-specific tasks. For instance, if ChatGPT is being used for drafting legal documents or medical prescriptions, including benchmarks relevant to those fields would provide a more accurate performance measure.
- **Multimodal Tasks:** With the advancement of AI, many chatbots have evolved to process multiple types of input (like text, voice, image, etc.). Including multimodal tasks in the benchmark set can help evaluate the AI's capabilities across different modalities.
- **Inter-task dependencies:** In real-world applications, a conversation often involves a series of interdependent tasks. Therefore, considering tasks in isolation may not fully represent the AI's capabilities. Including compound tasks, which require the completion of one task to start the next, can provide more comprehensive insights.

Edge Case Inclusion: Real-world use often involves scenarios that were not explicitly catered to during system design. These "edge cases" are critical for evaluating a system's robustness and generalization ability. For instance, the benchmarks could include dialogues that involve ambiguous references or require extensive world knowledge. It could also involve multilingual conversations, or conversations that require the system to handle sensitive topics tactfully. By including these tasks in the benchmarks, we can assess how well ChatGPT adapts to less-than-ideal or unexpected scenarios.

- **Ethical and Sensitive Scenarios:** Including tasks involving sensitive topics is crucial in assessing how the AI handles such situations. This could involve creating hypothetical scenarios where the user brings up a potentially distressing topic, and assessing how well the AI responds.
- **Handling Ambiguity:** Tasks should also be designed to measure the AI's ability to handle ambiguity. This could involve dialogues that contain ambiguous references, require inference from context, or involve languages other than English. Assessing these abilities would provide valuable insights into the AI's robustness and ability to generalize from its training.
- **Stress Testing:** This involves testing the AI system under extreme conditions, such as rapid-fire questioning, nonsensical input, or challenging factual questions. These tests can reveal the system's resilience and ability to handle unexpected situations.
- **Long Conversations:** Including tasks that involve long conversations can test the AI's ability to maintain context and coherence over an extended interaction. This is crucial in real-world applications, where conversations often go beyond simple question-answering.

Therefore, the process of selecting representative tasks as benchmarks goes beyond merely picking the most common use-cases. It involves an in-depth understanding of the system's intended use, its capabilities, and potential real-world scenarios it might encounter. By employing such a comprehensive approach, we ensure that the benchmarks chosen provide a detailed, holistic, and robust evaluation of ChatGPT's performance.

Strength

- Real-world application benchmarks aim to evaluate the system's performance in more diverse and dynamic conversational scenarios.
- They simulate realistic conversational settings and assess the system's ability to handle complex interactions, maintain context, and provide appropriate responses.
- These benchmarks provide a more comprehensive evaluation of the system's practical usability and performance.
- They help identify challenges and limitations that arise in real-world applications.

Weakness

- Designing and curating real-world application benchmarks can be challenging due to the need for diverse and representative datasets.
- Evaluating performance in real-world scenarios may introduce subjectivity, as user expectations and preferences can vary.
- It may be difficult to ensure consistent evaluation criteria across different real-world applications, potentially limiting direct comparison between systems.

- **Multi-Turn Dialogue Benchmarks:** These benchmarks assess the model's performance in extended conversations. This type of benchmarks evaluate the system's ability to engage in extended conversations involving multiple turns or exchanges. These benchmarks assess the system's contextual understanding, coherence, and ability to maintain a consistent dialogue flow over multiple interactions. They often involve complex dialogue datasets that capture the nuances of natural conversations. Specific evaluations may include.

- **Context Retention:** Evaluating the model's ability to remember previous turns of the conversation and use them to inform responses.
- **Consistent Persona:** Assessing whether the AI can maintain a consistent persona throughout a conversation.

Measuring multi-turn dialogue benchmarks may involve following techniques.

- **Dialogue Coherence:** Evaluate the overall coherence and continuity of the dialogue by assessing how well the system understands and responds to multiple turns of conversation.
- **Context Retention:** Measure the system's ability to remember and refer back to previous parts of the conversation accurately.
- **Consistency:** Assess the consistency of the system's responses across multiple turns, ensuring that the system maintains a coherent personality or persona throughout the dialogue.
- **Fluency:** Evaluate the system's ability to generate fluent and natural-sounding responses within the context of a multi-turn dialogue.
- **Engagement:** Measure the level of user engagement and interaction throughout the multi-turn dialogue, considering factors such as response length, prompt-following, and overall dialogue flow.

Strength

- Multi-turn dialogue benchmarks assess the system's performance in extended conversations involving multiple turns or exchanges.
- They evaluate the system's contextual understanding, coherence, and ability to maintain a consistent dialogue flow over multiple interactions.
- These benchmarks capture the complexities of natural conversations and test the system's ability to handle long-term dependencies.
- They provide insights into the system's ability to remember previous turns, maintain a consistent persona, and engage in coherent dialogues.

Weakness

- Designing high-quality multi-turn dialogue datasets that capture the intricacies of natural conversations can be challenging.
- Evaluating multi-turn dialogues requires more complex evaluation metrics beyond traditional measures, which can be subjective.
- Assessing system performance in multi-turn dialogues may require significant computational resources and time.

- **Ethical and Moral Evaluation:** To assess the ethical and moral aspects of a conversational AI system, various techniques can be employed. Bias analysis involves analyzing the system's training data and generated responses to identify potential biases related to gender, race, religion, or other protected attributes. Fairness metrics like disparate impact analysis, demographic parity, or equalized odds can be used to evaluate the system's responses across different demographic groups and identify any disparities or biases. Privacy assessment involves analyzing how the system handles user data, ensuring compliance with privacy regulations such as GDPR or HIPAA. Ethical alignment frameworks like OpenAI's ethical principles or IEEE's Ethically Aligned Design can be used to evaluate the system's adherence to ethical guidelines and principles such as fairness, transparency, accountability, and avoiding harm. This could include.
  - **Bias Detection** Analyzing the system's outputs to identify any potential biased behavior.
  - **Privacy Protection:** Evaluating the system's ability to avoid sensitive topics, not to store or misuse private user data.
  - **Responsible Data Handling:** Ensuring the AI does not manipulate or misuse data.

  Various evaluation techniques can be imposed to evaluate ethical and moral aspects as follows.

  - **Bias Analysis:** Conduct an in-depth examination of the system's training data and generated responses to identify potential biases in terms of gender, race, religion, or other protected attributes. This can involve statistical analysis, correlation studies, and fairness metrics to quantify the presence and impact of biases.
  - **Fairness Metrics:** Utilize fairness metrics, such as disparate impact analysis, demographic parity, or equalized odds, to evaluate the fairness of the system's responses across different demographic groups. These metrics can help identify and address any disparities or biases in the system's behavior.
  - **Privacy Assessment:** Perform a privacy impact assessment to analyze how the system handles user data, including data collection, storage, and sharing practices. Evaluate whether the system adheres to privacy regulations and guidelines, such as GDPR or HIPAA, and ensure that user privacy is protected.
  - **Ethical Alignment Frameworks:** Assess the system's adherence to ethical guidelines and frameworks, such as OpenAI's ethical principles, IEEE's Ethically Aligned Design, or the Montreal Declaration. This involves evaluating the system's behavior against specific ethical principles, such as fairness, transparency, accountability, and avoiding harm.

  Strength

  - Ethical and moral evaluation focuses on assessing the system's behavior in alignment with ethical guidelines and principles.
  - It helps identify potential biases, privacy concerns, and responsible data handling practices.
  - These evaluations promote fairness, transparency, accountability, and avoidance of harm in conversational AI systems.
  - They address societal concerns and contribute to the responsible development and deployment of AI technologies.

  Weakness

  - Ethical evaluation may involve subjective judgments, making it challenging to define and enforce standardized criteria.
  - Assessing ethical aspects often requires domain-specific expertise and understanding of societal norms and values.
  - Evaluating the long-term societal impact of conversational AI systems can be difficult, as ethical considerations evolve over time.

- **User Feedback and Human Evaluation:** Collecting user feedback is crucial in evaluating conversational AI systems. Surveys and questionnaires can be designed and distributed to gather feedback on aspects like user satisfaction, usefulness, naturalness, and perceived biases. User ratings can be obtained by allowing users to rate individual responses based on relevance, fluency, coherence, and appropriateness. Conducting preference tests enables users to compare and rank different system responses, revealing their preferences and identifying the most

desirable outputs. Human judgment can be employed by employing human judges to evaluate the system's responses against predefined criteria, assessing coherence, relevance, naturalness, and adherence to ethical and moral standards. Various metrics can be used.

- **User Satisfaction:** Measuring whether the user is satisfied with the interaction. Utility: Checking if the system provides useful and accurate information.
- **Understandability:** Assessing whether the user understands the system's responses.

Evaluation of user feedback and human involvement can me done in following ways.

- **Surveys and Questionnaires:** Design and distribute surveys or questionnaires to collect user feedback on various aspects of the conversational AI system, including satisfaction, usefulness, naturalness, and perceived biases. Use Likert scales, rating scales, or open-ended questions to gather quantitative and qualitative feedback.
- **User Ratings:** Allow users to rate individual responses generated by the system based on criteria such as relevance, fluency, coherence, and appropriateness. Aggregate these ratings to measure the overall quality of the system's output.
- **Comparative Evaluation:** Conduct preference tests where users are presented with different system responses and asked to compare and rank them based on preferred qualities. This helps identify user preferences and determine the most desirable responses.
- **Human Judgment:** Employ human judges who evaluate the system's responses based on predefined evaluation criteria. Judges assess aspects such as coherence, relevance to the user's query, naturalness, and adherence to ethical and moral standards.

Strength

- User feedback and human evaluation provide valuable insights into user satisfaction, usability, and the overall quality of system responses.
- They capture subjective aspects such as relevance, fluency, coherence, and appropriateness from the user's perspective.
- Human evaluation allows for the assessment of nuanced qualities that are challenging to capture through automated metrics alone.
- User feedback enables continuous improvement and iteration of the conversational AI system based on real user experiences.

Weakness

- Collecting user feedback and conducting human evaluations can be time-consuming and resource-intensive.
- Subjective nature of user feedback and human judgment may introduce biases or inconsistencies in the evaluation process.
- Scaling user feedback and human evaluation across a large user base can be challenging, leading to limited sample sizes.

- **Reinforcement Learning-based Evaluation:** Reinforcement learning techniques can be utilized to evaluate and improve conversational AI systems. Defining reward models that capture the desired behavior and objectives of the system guides the reinforcement learning process. Offline evaluation involves simulating or replaying user interactions to assess the system's performance using historical dialogues or synthetic user interactions. This helps evaluate the quality of generated responses based on predefined evaluation metrics. Online evaluation involves deploying the system in a live setting and collecting real-time user feedback. Through techniques like active learning, users can provide feedback on specific responses, which is then used to update the model and improve its performance over time. It uses the following criteria.

  - **Positive Feedback:** If the model performs well on a task, it is rewarded, encouraging such behavior in the future.
  - **Negative Feedback:** If the model performs poorly or makes a mistake, it is penalized, discouraging such behavior in the future.

  Measuring the reinforcement learning-based evaluation can be done in below mentioned ways.

  - **Reward Models:** Define reward models that capture desired behavior and objectives for the conversational AI system. These reward models guide the reinforcement learning process, allowing the system to learn and improve its responses based on the feedback received.
  - **Offline Evaluation:** Simulate or replay user interactions offline to evaluate the system's performance. This involves using historical dialogues or synthetic user interactions to assess the quality of generated responses against predefined evaluation metrics.
  - **Online Evaluation:** Deploy the system in a live setting and collect real-time user feedback. This can be done through active learning techniques, where the system prompts users for feedback on specific responses. The collected feedback is then used to update the model and improve its performance over time.

  Strength

  - Reinforcement learning-based evaluation involves training the system using reward models to optimize its behavior.
  - It allows for adaptive learning and improvement of the conversational AI system over time.
  - This evaluation approach can address the limitations of static benchmarks by enabling the system to learn from user interactions.
  - Reinforcement learning-based evaluation provides a dynamic and iterative evaluation process.

  Weakness

  - Designing effective reward models that capture the desired behavior can be challenging.
  - Reinforcement learning-based evaluation requires substantial computational resources and time.
  - The trial-and-error learning process of reinforcement learning may lead to unintended consequences and potential ethical concerns.
  - It may be difficult to interpret and explain the inner workings of the system trained through reinforcement learning.

3.4.3. Workloads for benchmark

Determining the precise number of workloads or the types of workloads sufficient for comprehensive benchmarking is a complex task. In an ideal scenario, benchmarks should cover a broad spectrum of scenarios that a system could encounter. However, it is impractical and nearly impossible to include every possible workload due to the inherent diversity of real-world interactions and applications. Hence, we propose a balanced and representative selection of workloads.

- **Task-Specific Workloads:** A good starting point is to include a variety of task-specific workloads that reflect different types of

tasks a conversational AI might be expected to perform. For instance, this could include tasks such as booking a flight, setting an appointment, providing a weather update, and answering trivia questions. This can test the system's ability to understand and respond to specific intents.

- **Domain-Specific Workloads:** Additionally, benchmarks should incorporate workloads specific to various domains like healthcare, finance, and education, to name a few. Different domains have unique language patterns, terminologies, and contextual nuances, providing a rigorous test for the system's adaptability and contextual understanding.
- **General Conversation Workloads:** Finally, the workload should also include more free-form conversational interactions that are not tied to a specific task or domain. This can help evaluate the system's ability to carry on a meaningful, coherent, and engaging conversation.

The combination of these workloads would be determined by the target application of the AI model. For example, a conversational AI designed for customer service might have a higher focus on task-specific and domain-specific (i.e., customer service-related) workloads, whereas a general-purpose AI might require a more balanced mix. The key here is to ensure that the chosen workloads are representative and challenging enough to cover a wide range of scenarios the system could face, yet still feasible to be implemented in practice. The exact number and selection of workloads would vary based on these considerations. However, it is essential to continuously update and expand these workloads as new tasks, domains, and use-cases emerge.

When determining the number and types of workloads that are sufficient for benchmarking, it is important to strike a balance between comprehensiveness and feasibility. While it is indeed challenging and impractical to cover every possible workload as a benchmark, there are strategies to ensure an effective and representative evaluation.

- **Diversity of Workloads:** Instead of aiming for exhaustive coverage, focus on selecting workloads that represent a diverse range of tasks, domains, and conversational scenarios. This can include a mix of common real-world tasks, industry-specific use cases, and challenging or complex scenarios.
- **Importance and Relevance:** Prioritize workloads that are widely used or have significant practical importance. Consider tasks that are commonly encountered in real-world applications, as well as those that pose specific challenges or require sophisticated language understanding and generation capabilities.
- **Coverage of Key Domains:** Identify key domains or industries where conversational AI systems are expected to perform well. This can include healthcare, customer support, education, e-commerce, and others. Select representative workloads from these domains to evaluate the system's performance in domain-specific contexts.
- **Scalability:** Consider the scalability of workloads. While it may not be feasible to cover every possible variation, ensure that the selected workloads cover a sufficient range of complexities, including variations in conversational styles, user intents, and system responses.
- **Balancing Breadth and Depth:** Aim for a balance between breadth and depth in workload coverage. While it is important to cover a wide range of workloads, ensure that each workload is evaluated in sufficient detail to capture nuances and intricacies specific to that task.
- **User-Centric Approach:** Incorporate user feedback and preferences in workload selection. Consider the tasks that users commonly seek assistance with or find challenging. This can help identify workloads that align with user needs and expectations.
- **Continuous Evaluation:** Recognize that the landscape of workloads and user requirements is dynamic. As new tasks and domains emerge, continuously evaluate and update the benchmark suite to reflect evolving demands.

3.4.4. Integration of the metrics with proposed framework

The integration of these newly proposed metrics within your existing framework adds nuanced, context-aware dimensions to the evaluation of conversational AI models like ChatGPT. These proposed metrics—Contextual Sensitivity Index (CSI), Dialogue Coherence Measure, Relevance Measure, and Task Success Rate (TSR), along with traditional metrics like BLEU, ROUGE, METEOR, F1 score, precision, recall, and perplexity—offer an extensive spectrum of evaluation criteria.

- **Contextual Sensitivity Index (CSI):** The CSI serves as a thermometer for the AI's ability to perceive the ebb and flow of the conversational context. In customer service scenarios, the AI's responses should not only be accurate but also empathetic, particularly if the customer is showing signs of frustration. In chatbot applications for mental health support, the weightage of CSI could be significantly higher, as understanding and adjusting to the user's emotional context is vital.
- **Relevance Measure:** Ensuring relevance in AI responses is crucial for maintaining user engagement and satisfaction. For instance, in a digital assistant application where the user asks for weather updates, a response about the latest news headlines, although perfect in grammar and syntax, is irrelevant. Applications that demand direct answers to user queries, such as virtual assistants or customer support bots, should assign a higher weightage to this metric.
- **Task Success Rate (TSR):** In task-oriented applications, the system's competency is directly linked to how successfully it performs a particular task. In a restaurant reservation bot, TSR would measure how accurately the bot processes user input (date, time, venue, etc.) to complete the booking. The higher the success rate, the more reliable the bot is perceived by the users. Applications that are built to perform specific tasks should assign a higher weightage to TSR.
- **Dialogue Coherence Measure:** This metric is essential for applications involving multi-turn dialogues. For instance, a tutoring bot should follow the topic discussed, maintain the continuity of ideas, and avoid abrupt topic switches. A higher weightage could be given to this measure in scenarios involving extended dialogues, such as tutoring, therapy, or general conversation bots.
- **User Satisfaction:** This could involve various parameters such as the AI's response speed, relevance, coherence, and politeness. Depending on user feedback and the specific use case, the importance of these parameters may differ. For example, in time-sensitive applications, like customer support, users might value response speed more, while in therapy bots, users may value politeness and coherence more. Weights should be adjusted accordingly to align with user preferences.
- **Traditional Metrics (BLEU, ROUGE, METEOR, F1 score, Precision, Recall, Perplexity):** Each of these metrics offers different insights about the linguistic capabilities of the model. For example, in a language translation bot, metrics like BLEU and METEOR would have higher weightage as they measure how close the translated text is to the reference translation. However, in a question-answering bot, Precision and Recall may have higher weightage as they measure how accurately the bot retrieves the relevant information.

Imposing Weights on Metrics

In order to compute a comprehensive score, each metric could be normalized to a standard scale, perhaps between 0 and 1 or 0 to 100, to allow for comparison across different measures. Following this, the overall score could be computed as a weighted average of these normalized scores. The weights assigned to each metric could be decided based on several factors such as the specific use case of the model, user feedback, or empirical evidence from pilot studies.

For instance, if the ChatGPT model is primarily used for customer support, higher weight might be given to TSR, Relevance Measure, and CSI, since these would be critical for the successful operation in a customer service environment. Conversely, if the model is being used for creative writing or storytelling, Dialogue Coherence Measure and traditional language generation metrics (like BLEU, ROUGE, METEOR) might receive higher weighting. Furthermore, these weights could be dynamically adjusted based on user feedback. For instance, if users consistently indicate that they value relevance and coherence over perfect grammatical correctness, the weights for Relevance Measure and Dialogue Coherence Measure could be increased, and weights for traditional metrics like BLEU and ROUGE could be decreased.

Assigning different weights to metrics in your benchmarking framework requires careful consideration of the specifics of the AI model, its application area, and its user base. Here's how this process might unfold:

- **Understand the Use Case:** The primary use case of the AI model should be the first determinant of weights. For instance, if you are evaluating a customer service chatbot, you might assign more weight to the Task Success Rate (TSR) and Relevance Measure, as these aspects are crucial for solving customer issues. On the other hand, a therapeutic chatbot might require a higher emphasis on the Contextual Sensitivity Index (CSI) and Dialogue Coherence Measure.
- **Consider User Preferences and Feedback:** Feedback from users can provide insights into which aspects of the AI model are most important to them. Regular user surveys, user-testing sessions, and analyses of user reviews and ratings can help you understand what users value in the AI's performance. This understanding can guide the weight assignment. For example, if users particularly appreciate coherent and context-sensitive responses, assign higher weights to the CSI and Dialogue Coherence Measure.
- **Leverage Domain Expert Opinions:** Domain experts can provide valuable guidance on assigning weights. For instance, a linguistics expert might suggest a higher weightage for traditional NLP metrics like BLEU and ROUGE for language learning applications. Meanwhile, a psychologist might advise prioritizing ethical considerations and context sensitivity for therapeutic applications.
- **Use a Data-Driven Approach:** Machine learning techniques can be applied to automatically adjust the weights based on empirical evidence. Regression analysis, for example, could be used to find the correlation between different metrics and overall user satisfaction. The metrics most strongly correlated with satisfaction would receive higher weights.
- **Iterative Refinement:** The initial weights should not be set in stone; they should be subject to regular reassessment and refinement. Continually analyzing user feedback, monitoring changes in user behavior, and staying attuned to advancements in the AI field will provide the data necessary to adjust weights over time, ensuring the benchmarking framework remains effective and relevant.

The goal of assigning weights is to tailor the evaluation framework to provide the most meaningful assessment of an AI model's performance in its intended application. Hence, this process must be thoughtful, flexible, and continuously evolving. Ultimately, the key is to maintain a level of flexibility and adaptability in your framework. The ability to adapt the weights based on these factors will ensure that your benchmarking framework remains relevant and effective in evaluating and improving the performance of conversational AI models like ChatGPT.

3.4.5. Feasibility analysis of proposed framework

Feasibility analysis is key to understand how well the proposed framework would be taken into the consideration for realistic use.

- **Technical Feasibility:** The proposed evaluation framework involves advanced techniques such as natural language processing, reinforcement learning, and machine learning. While these techniques are well-established within the AI research community, they require a high level of technical expertise. There are a number of open-source tools and libraries available (such as NLTK, Gensim, and Scikit-learn for NLP tasks, and Tensorflow, PyTorch, and OpenAI Gym for RL tasks), which can be leveraged to implement the components of this framework. However, effectively integrating these techniques into a cohesive system is a complex task that may require considerable time and effort.
- **Operational Feasibility:** Operationally, this framework involves the collection and processing of large amounts of data, which may present challenges related to data storage, computational resources, and privacy concerns. The development of this framework would likely require significant computational power, potentially requiring the use of high-performance computing resources or cloud-based solutions.
- **Economic Feasibility:** Economically, the development and implementation of this comprehensive evaluation framework could be costly. Costs would be associated with hiring skilled personnel, acquiring computational resources, collecting and processing data, and maintaining and updating the framework over time. Therefore, a careful cost–benefit analysis should be conducted to assess the economic feasibility of this project.
- **Legal Feasibility:** Given that this framework involves the collection and processing of user data, it is essential to consider legal and regulatory requirements, such as data protection laws. The use of reinforcement learning techniques also raises ethical considerations, as these methods often involve trial-and-error learning, which could potentially result in unintended consequences.
- **Schedule Feasibility:** The development of this framework would likely be a time-consuming process. Each component of the framework, from the task-specific benchmarks to the user-centric evaluation, involves substantial research and development. It is crucial to develop a realistic project timeline that accounts for these complexities.

3.4.6. Adaptability analysis of proposed framework

Adaptability of the proposed framework faces several key challenges as mentioned below.

- **Complexity of Implementation:** The framework comprises multiple components, each requiring specialized knowledge in areas such as natural language processing, machine learning, ethics in AI, and reinforcement learning. Getting a team with such a diverse skill set can be a challenge.
- **Time and Resource Intensive:** Due to its comprehensive nature, implementing this framework could be time-consuming. Additionally, creating or obtaining the datasets for evaluation, particularly those related to real-world applications, could be costly and labor-intensive.
- **Evolving Nature of AI:** The rapid advancement in AI and NLP technologies would require the framework to be continuously updated and refined to stay relevant, which might be challenging.
- **Scalability:** If the ChatGPT model is updated frequently, or if there are many versions to evaluate, scaling the proposed framework might be difficult.
- **Interpretability and Transparency:** Even with a comprehensive evaluation, explaining the inner workings of AI models (like ChatGPT) in a comprehensible manner remains a challenge. This could make the adoption of the framework difficult for those seeking easily interpretable evaluation results.

- **Ethical Considerations:** The framework aims to address ethical issues such as bias, privacy, and data handling. However, defining and enforcing these standards can be difficult due to the subjective nature of ethics and the global variation in ethical norms and regulations.
- **Acceptance from the Scientific Community:** Given the novelty and the comprehensive nature of the proposed framework, it may take time for it to be accepted and adopted by the larger scientific and research community. Rigorous peer review and validation would be necessary to achieve widespread adoption.

3.4.7. SWOT analysis of proposed framework

SWOT (Strengths, Weaknesses, Opportunities and Threats) analysis is a useful tool to evaluate the strengths, weaknesses, opportunities, and threats associated with the proposed framework.

- **Strengths:**
  - Comprehensive Framework: The proposed framework covers various aspects of evaluation, including task-specific benchmarks, real-world application benchmarks, and user-centric evaluation. It provides a holistic approach to assess the performance and capabilities of ChatGPT.
  - Integration of Advanced Techniques: The framework incorporates advanced techniques such as natural language processing, reinforcement learning, and user feedback analysis. This integration enables a more nuanced evaluation of ChatGPT's language generation and contextual understanding abilities.
  - Alignment with Ethical Considerations: The framework emphasizes ethical and responsible AI development by proposing adaptive standards and considering issues such as bias, privacy, and transparency. It aims to ensure that ChatGPT meets the highest ethical standards.
- **Weaknesses:**
  - Implementation Complexity: Implementing the proposed framework would require a high level of technical expertise and computational resources. The integration of different components, such as data collection, benchmark creation, and user feedback analysis, can be challenging and time-consuming.
  - Lack of Concrete Artifacts: The framework provides a conceptual structure but lacks specific tools or artifacts for implementation. This may make it difficult for researchers and practitioners to adopt the framework without additional guidance.
- **Opportunities:**
  - Advancements in AI Technologies: The rapid development of AI technologies provides opportunities to leverage new techniques, algorithms, and tools in the evaluation framework. Incorporating cutting-edge approaches can enhance the accuracy, efficiency, and effectiveness of the evaluation process.
  - Collaboration and Knowledge Sharing: The proposed framework encourages collaboration among researchers, industry experts, and practitioners. This collaborative approach can lead to the sharing of best practices, datasets, and evaluation methodologies, fostering continuous improvement and standardization.
- **Threats:**
  - Data Privacy and Security Concerns: The collection and processing of user data for evaluation purposes raise privacy and security concerns. Adhering to data protection regulations and implementing robust security measures is essential to mitigate these threats.
  - Bias and Fairness Issues: As ChatGPT learns from large-scale datasets, it may inherit biases present in the training data. Ensuring fairness and mitigating bias in the evaluation process is a critical challenge. Failing to address these issues could lead to biased outcomes and ethical concerns.

3.4.8. Adaptive standards of proposed framework

Adaptive standards play a crucial role in guiding the development and deployment of the proposed framework. By evolving the standards to align with emerging challenges and ethical considerations, we can ensure responsible and effective use of the system.

- **Ethically Aligned Design**
  - Incorporate principles from frameworks such as IEEE's Ethically Aligned Design and the Montreal Declaration to guide the ethical development and deployment of ChatGPT.
  - Integrate fairness, transparency, accountability, and privacy considerations into the standards to address potential biases, ensure responsible data handling, and protect user privacy.
- **Contextual Adaptability**
  - Establish standards that promote adaptability to diverse conversational contexts and user preferences.
  - Enable ChatGPT to dynamically adjust its responses based on user feedback, adapting to individual user needs and societal changes.
- **Collaboration and Openness**
  - Foster collaboration among researchers, developers, and users to collectively define adaptive standards for ChatGPT.
  - Emphasize open-source contributions, shared knowledge, and community-driven development to ensure transparency and inclusivity in the standard-setting process.

3.4.9. Use of proposed framework for intelligent evaluation

By "Intelligent Evaluation", we refer to the process of incorporating multi-faceted, nuanced methods to capture the depth of ChatGPT's performance via the proposed framework. This involves going beyond traditional measures, leveraging user feedback, and employing reinforcement learning for evaluation.

- **Metrics Beyond Traditional Measures** While traditional metrics like BLEU, ROUGE, and F1 score provide a quantitative measure of system performance, they may not fully capture aspects such as context-sensitivity, dialogue coherence, and relevance of responses. We propose to supplement these with metrics that focus on evaluating dialogue quality and contextual understanding. For example, one could use the Contextual Sensitivity Index (CSI), a metric we propose that quantifies the degree to which a model's responses vary appropriately with changes to the conversational context.
- **User Feedback and Human Evaluation:** This involves collecting qualitative feedback from users regarding their interaction with ChatGPT, which can provide insights into user satisfaction and the perceived quality of conversations. This can be carried out through user studies or surveys post-interaction.
- **Application of Reinforcement Learning in Evaluation:** In reinforcement learning-based evaluation, an agent (in this case, ChatGPT) learns to make decisions by taking actions in an environment to maximize some notion of cumulative reward. For instance, a dialogue manager could be trained to optimize the cumulative reward of maintaining user engagement and minimizing harmful or inappropriate responses. We outline a reinforcement learning-based evaluation pipeline in Algorithm 1 and provide implementation details to aid in reproducibility.

4. Challenges and future directions for benchmarking, standards, and evaluation for ChatGPT

In this section, we discuss the key challenges and future directions in benchmarking, standards, and evaluation for ChatGPT [35,36].

4.1. Key challenges

- **Data and Representativeness:** The availability of diverse and representative datasets is crucial for benchmarking ChatGPT. However, existing datasets may exhibit biases or lack representation across various demographic and cultural groups, leading to skewed model performance. Future research should focus on creating more inclusive datasets that encompass a wide range of languages, cultures, and perspectives. Additionally, techniques such as data augmentation and debiasing methods can be explored to reduce biases in training data.
- **Scalability and Efficiency:** As ChatGPT becomes more powerful and complex, scalability and efficiency become critical concerns. Handling high volumes of concurrent conversations and ensuring real-time interactions pose challenges in benchmarking and evaluation. To address these challenges, future research should focus on developing benchmarks and evaluation methodologies that specifically measure the scalability and efficiency of ChatGPT. Techniques such as distributed computing, parallelization, and model compression can be investigated to improve scalability and reduce inference latency.
- **Explainability and Interpretability:** The black-box nature of ChatGPT limits its explainability, making it difficult to understand how decisions are made and potentially raising ethical concerns. Lack of interpretability hinders the establishment of transparent standards and the evaluation of bias and fairness. Future research should focus on developing methods to enhance the explainability and interpretability of ChatGPT. Techniques such as model introspection, attention visualization, and rule-based post-processing can be explored to shed light on the decision-making processes and ensure transparency in the system's behavior.
- **Adversarial Attacks and Security:** ChatGPT may be vulnerable to adversarial attacks, where malicious actors attempt to manipulate or deceive the system by inputting carefully crafted inputs. Ensuring the security and robustness of ChatGPT in real-world scenarios is essential. Future research should investigate adversarial attack techniques specific to ChatGPT and develop robust defenses against such attacks. Techniques such as adversarial training, input sanitization, and ensemble methods can be explored to enhance the system's security and resilience.
- **Real-Time User Feedback Integration:** Incorporating real-time user feedback into the evaluation process can be logistically challenging. Gathering and processing user feedback in a timely manner to provide actionable insights for model improvement is a complex task. Future research should focus on developing efficient mechanisms to collect and process real-time user feedback during interactive conversations. Techniques such as natural language understanding, sentiment analysis, and active learning can be leveraged to derive meaningful insights and guide model adaptation in real-time.
- **Multimodal Conversational AI:** The integration of multimodal inputs, such as text, images, and audio, presents new challenges for benchmarking and evaluation. Evaluating the performance of multimodal conversational AI systems like ChatGPT requires specialized benchmarks and evaluation criteria. Future research should focus on creating multimodal benchmarks and evaluation methodologies that assess the performance of ChatGPT in processing and generating responses from multiple modalities. Additionally, novel metrics and evaluation techniques need to be developed to capture the multimodal aspects of conversational AI accurately.

4.2. Future direction

- **Enhanced Data Collection:** Future research should prioritize the creation of more diverse and inclusive datasets for benchmarking ChatGPT. This includes capturing a wide range of languages, cultures, and perspectives to reduce biases and improve the system's performance across different demographics and contexts. Techniques such as data augmentation, crowdsourcing, and domain adaptation can be further explored to enhance dataset representativeness [37,38].
- **Scalability and Efficiency Improvements:** To address the scalability and efficiency challenges, future research should focus on developing benchmarks and evaluation methodologies specifically designed to measure ChatGPT's performance under high loads and real-time interaction scenarios. Techniques such as distributed computing, parallelization, model optimization, and hardware acceleration can be investigated to enhance the scalability and efficiency of ChatGPT in practical deployment scenarios.
- **Improved Explainability and Interpretability:** Future research should strive to improve the explainability and interpretability of ChatGPT by developing methods that shed light on its decision-making processes. This can include techniques such as rule-based post-processing, attention mechanisms, counterfactual explanations, and interactive visualization tools, which provide insights into the factors influencing ChatGPT's responses and facilitate the establishment of transparent standards and the evaluation of bias and fairness.
- **Robustness against Adversarial Attacks:** To ensure the security and robustness of ChatGPT, future research should focus on investigating adversarial attack techniques specific to ChatGPT and developing robust defenses against such attacks. Techniques such as adversarial training, input sanitization, ensemble methods, and anomaly detection can be explored to enhance the system's resilience against malicious inputs and adversarial manipulations.
- **Improved Integration in Real-Time User Feedback:** Efficient mechanisms for collecting and processing real-time user feedback during interactive conversations should be developed. This can involve leveraging natural language understanding techniques, sentiment analysis, active learning, and reinforcement learning to derive meaningful insights from user feedback in real-time. The integration of real-time user feedback will provide valuable insights for model adaptation, improvement, and personalized user experiences.
- **Advancements in Multimodal Conversational AI:** As multimodal inputs gain prominence in conversational AI, future research should focus on developing specialized benchmarks and evaluation methodologies for multimodal conversational AI systems like ChatGPT. This includes creating benchmarks that assess ChatGPT's performance in processing and generating responses from multiple modalities, such as text, images, and audio. Additionally, novel metrics and evaluation techniques need to be developed to capture the multimodal aspects of conversational AI accurately, considering factors such as modality integration, coherence, user satisfaction, and multimodal context understanding.

5. Conclusion

This paper has presented a comprehensive evaluation framework that addresses the challenges and complexities of evaluating conversational AI systems like ChatGPT. We have examined prominent benchmarks, including GLUE, SuperGLUE, SQuAD, CoQA, Persona-Chat, DSTC, BIG-Bench, HELM, and MMLU, and assessed their strengths and limitations in evaluating ChatGPT's performance. These benchmarks offer standardized tasks and evaluation metrics to measure the system's

contextual understanding, coherence in generating responses, and conversational relevance. To ensure ethical and responsible development, we have proposed adaptive standards aligned with recognized frameworks such as OpenAI's principles, IEEE's Ethically Aligned Design, the Montreal Declaration, and Partnership on AI's Tenets. These standards promote fairness, transparency, accountability, and privacy, while accommodating the evolving challenges of conversational AI. Intelligent evaluation methods play a crucial role in measuring the quality and effectiveness of ChatGPT. We have explored metrics beyond traditional measures, incorporating user feedback and reinforcement learning techniques. By leveraging these methods, we can comprehensively assess response coherence, context-awareness, fluency, relevance, and user engagement. Our evaluation framework incorporates task-specific benchmarks, real-world application benchmarks, and multi-turn dialogue benchmarks to enhance adaptability and representativeness. These benchmarks capture the nuances and complexities of conversational AI, providing a holistic evaluation of ChatGPT's performance. Through this comprehensive evaluation framework, we aim to drive the responsible and impactful development of ChatGPT and conversational AI systems. By continually refining benchmarks, adapting standards, and utilizing intelligent evaluation methods, we can foster systems that deliver natural, contextually aware, and ethically sound conversational experiences. As the field of conversational AI evolves, our evaluation framework serves as a foundation for ongoing research, collaboration, and improvement. We hope that this framework inspires further advancements, promotes user-centric design, and ensures that ChatGPT and future conversational AI systems meet the highest standards of performance, ethics, and user satisfaction.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Javaid, A. Haleem, R.P. Singh, S. Khan, I.H. Khan, Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system, in: BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2023, 100115.

[2] M.T.R. Laskar, M.S. Bari, M. Rahman, M.A.H. Bhuiyan, S. Joty, J.X. Huang, A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets, 2023, arXiv preprint arXiv:2305.18486.

[3] F. Muftić, M. Kadunić, A. Mušinbegović, A. Abd Almisreb, Exploring medical breakthroughs: A systematic review of ChatGPT applications in healthcare, South. Eur. J. Soft Comput. 12 (1) (2023) 13–41.

[4] X. Zhang, C. Li, Y. Zong, Z. Ying, L. He, X. Qiu, Evaluating the performance of large language models on GAOKAO benchmark, 2023, arXiv preprint arXiv:2305.12474.

[5] N.G. Vidhya, D. Devi, A. Nithya, T. Manju, Prognosis of exploration on Chat GPT with artificial intelligence ethics, Braz. J. Sci. 2 (9) (2023) 60–69.

[6] Y. Huang, A. Gomaa, T. Weissmann, J. Grigo, H.B. Tkhayat, B. Frey, F. Putz, Benchmarking ChatGPT-4 on ACR radiation oncology in-training exam (TXIT): Potentials and challenges for AI-Assisted medical education and decision making in radiation oncology, 2023, arXiv preprint arXiv:2304.11957.

[7] Y. Huang, A. Gomaa, S. Semrau, M. Haderlein, S. Lettmaier, T. Weissmann, F. Putz, Benchmarking ChatGPT-4 on ACR radiation oncology in-training (TXIT) exam and red journal gray zone cases: Potentials and challenges for AI-Assisted medical education and decision making in radiation oncology, 2023, Available at SSRN 4457218.

[8] X. Ohmer, E. Bruni, D. Hupkes, Evaluating task understanding through multilingual consistency: A ChatGPT case study, 2023, arXiv preprint arXiv:2305.11662.

[9] D. Sobania, M. Briesch, C. Hanna, J. Petke, An analysis of the automatic bug fixing performance of ChatGPT, 2023, arXiv preprint arXiv:2301.08653.

[10] J. Oppenlaender, J. Hämäläinen, Mapping the challenges of HCI: An application and evaluation of ChatGPT and GPT-4 for cost-efficient question answering, 2023, arXiv preprint arXiv:2306.05036.

[11] Q. Chen, H. Sun, H. Liu, Y. Jiang, T. Ran, X. Jin, et al., A comprehensive benchmark study on biomedical text generation and mining with ChatGPT, 2023, bioRxiv, 2023-04.

[12] W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, et al., Agieval: A human-centric benchmark for evaluating foundation models, 2023, arXiv preprint arXiv:2304.06364.

[13] B. Wang, X. Yue, H. Sun, Can ChatGPT defend the truth? Automatic dialectical evaluation elicits LLMs' Deficiencies in reasoning, 2023, arXiv preprint arXiv:2305.13160.

[14] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, X. He, Is ChatGPT fair for recommendation? Evaluating fairness in large language model recommendation, 2023, arXiv preprint arXiv:2305.07609.

[15] Glue, 2023, https://gluebenchmark.com/. (Accessed 17 June 2023).

[16] SuperGLUE, 2023, https://super.gluebenchmark.com/. (Accessed 17 June 2023).

[17] SQuAD, 2023, https://huggingface.co/datasets/squad. (Accessed 17 June 2023).

[18] CoQA, 2023, https://stanfordnlp.github.io/coqa/. (Accessed 17 June 2023).

[19] Persona-Chat, S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, J. Weston, Personalizing dialogue agents: I have a dog, do you have pets too? 2018, arXiv preprint arXiv:1801.07243.

[20] DSTC, 2023, https://github.com/alexa/alexa-with-dstc10-track2-dataset. (Accessed 17 June 2023).

[21] BIG-Bench, 2023, https://github.com/google/BIG-bench. (Accessed 8 July 2023).

[22] HELM, 2023, https://crfm.stanford.edu/helm/latest/. (Accessed 8 July 2023).

[23] MMLU, 2023, https://arxiv.org/abs/2212.10455. (Accessed 8 July 2023).

[24] OpenAI's policy, 2023, https://openai.com/policies/usage-policies. (Accessed 17 June 2023).

[25] IEEE Ethically aligned design, 2023, https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf. (Accessed 17 June 2023).

[26] Montreal declaration for responsible development, 2023, https://monoskop.org/images/d/d2/Montreal_Declaration_for_a_Responsible_Development_of_Artificial_Intelligence_2018.pdf. (Accessed 17 June 2023).

[27] Partnership on AI's tenet, 2023, https://partnershiponai.org/. (Accessed 17 June 2023).

[28] BLEU, 2023, https://machinelearningmastery.com/calculate-bleu-score-for-text-python/. (Accessed 17 June 2023).

[29] ROUGE, 2023, https://huggingface.co/spaces/evaluate-metric/rouge. (Accessed 17 June 2023).

[30] METEOR, 2023, https://huggingface.co/spaces/evaluate-metric/meteor. (Accessed 17 June 2023).

[31] A. Koubaa, W. Boulila, L. Ghouti, A. Alzahem, S. Latif, Exploring ChatGPT capabilities and limitations: A critical review of the nlp game changer, 2023.

[32] X. He, X. Shen, Z. Chen, M. Backes, Y. Zhang, Mgtbench: Benchmarking machine-generated text detection, 2023, arXiv preprint arXiv:2303.14822.

[33] C. Chan, J. Cheng, W. Wang, Y. Jiang, T. Fang, X. Liu, Y. Song, Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations, 2023, arXiv preprint arXiv:2304.14827.

[34] J. Li, X. Cheng, W.X. Zhao, J.Y. Nie, J.R. Wen, HELMA: A large-scale hallucination evaluation benchmark for large language models, 2023, arXiv preprint arXiv:2305.11747.

[35] I. Jahan, M.T.R. Laskar, C. Peng, J. Huang, Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers, 2023, arXiv preprint arXiv:2306.04504.

[36] P.P. Ray, ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, 2023, Internet of Things and Cyber-Physical Systems.

[37] C.K. Lo, What is the impact of ChatGPT on education? A rapid review of the literature, Educ. Sci. 13 (4) (2023) 410.

[38] M. Haman, M. Åkolník, Using ChatGPT to conduct a literature review, Account. Res. (2023) 1–3.