BenchCouncil Transactions
on Benchmarks, Standards and Evaluations

TBench
Volume 3, Issue 3
2023

Original Articles

Review Articles

ISSN: 2772-4859

BenchCouncil Transactions on Benchmarks, Standards and Evaluations (TBench) is an open-access multi-disciplinary journal dedicated to benchmarks, standards, evaluations, optimizations, and data sets. This journal is a peer-reviewed, subsidized open access journal where The International Open Benchmark Council pays the OA fee. Authors do not have to pay any open access publication fee. However, at least one of the authors must register BenchCouncil International Symposium on Benchmarking, Measuring and Optimizing (Bench) (https://www.benchcouncil.org/bench/) and present their work. It seeks a fast-track publication with an average turnaround time of one month.

CONTENTS

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.

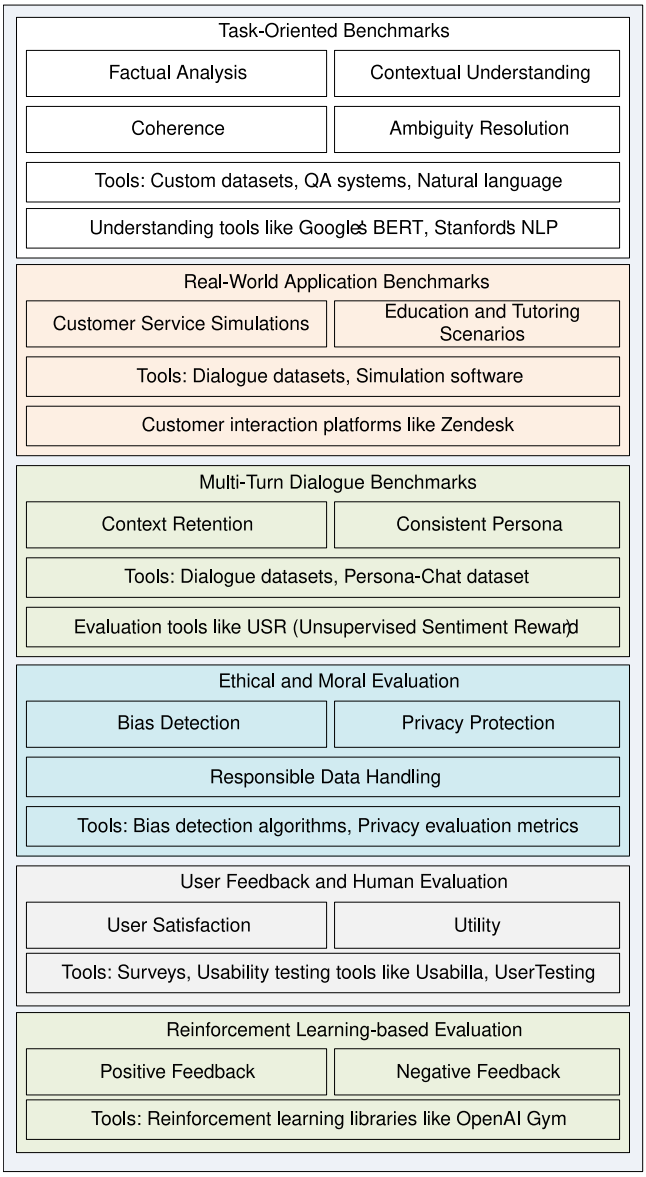

Fig. 1. Proposed framework for ChatGPT benchmark evaluation.

- 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.

BenchCouncil Transactions on Benchmarks, Standards and Evaluations 3 (2023) 100140

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

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

Full Length Article

Analyzing the potential benefits and use cases of ChatGPT as a tool for improving the efficiency and effectiveness of business operations

Rohit Raj [a], Arpit Singh [b], Vimal Kumar [a,*], Pratima Verma [c]

[a] Department of Information Management, Chaoyang University of Technology, Taichung-413310, Taiwan, China
[b] Department of Information Systems and Analytics, O. P. Jindal Global University, Sonipat, India 131029
[c] Department of Strategic Management, Indian Institute of Management Kozhikode, India, 673570

ARTICLE INFO

Keywords:
ChatGPT
benefits
business
efficiency
automation

ABSTRACT

The study addresses the potential benefits for companies of adopting ChatGPT, a popular chatbot built on a large-scale transformer-based language model known as a generative pre-trained transformer (GPT). Chatbots like ChatGPT may improve customer service, handle several client inquiries at once, and save operational costs. Moreover, ChatGPT may automate regular processes like order tracking and billing, allowing human employees to focus on more complex and strategic responsibilities. Nevertheless, before deploying ChatGPT, enterprises must carefully analyze its use cases and restrictions, as well as its strengths and disadvantages. ChatGPT, for example, requires training data that is particular to the business domain and might produce erroneous and ambiguous findings. The study identifies areas of deployment of ChatGPT's possible benefits in enterprises by drawing on the literature that is currently accessible on ChatGPT, massive language models, and artificial intelligence. Then, using the PSI (Preference Selection Index) and COPRAS (Complex Proportional Assessment) approaches, potential advantages are taken into account and prioritized. By highlighting current trends and possible advantages in the industry, this editorial seeks to provide insight into the present state of employing ChatGPT in enterprises and research. ChatGPT may also learn biases from training data and create replies that reinforce those biases. As a result, enterprises must train and fine-tune ChatGPT to specific operations, set explicit boundaries and limitations for its use, and implement appropriate security measures to avoid malicious input. The study highlights the research gap in the dearth of literature by outlining ChatGPT's potential benefits for businesses, analyzing its strengths and limits, and offering insights into how organizations might use ChatGPT's capabilities to enhance their operations.

1. Introduction

As artificial intelligence (AI) continues to grow and become advanced, more businesses are exploring ways to integrate technologies led by AI into their operations [1]. One such technology that has gained significant traction from businesses around the world is chatbots. Chatbots are automated systems that use natural language processing (NLP) algorithms capable of simulating conversations with humans, providing customers with instant support and assistance [2]. Chat-generative pre-trained (ChatGPT) is a popular chatbot that is a large language model trained by OpenAI with the potential of providing several benefits to businesses [3]. ChatGPT is based on a large-scale transformer-based language model called generative pre-trained transformer (GPT), which was first introduced by OpenAI in the year 2018 [4]. It was trained on a large corpus of text data to learn the patterns and structures of natural language using unsupervised learning methods. This enabled ChatGPT to learn from raw data without any explicit supervision [5]. After the success of GPT, OpenAI created ChatGPT, which is particularly meant to replicate human-like user dialogues. ChatGPT was trained using a vast dataset of online interactions from social media, forums, and other sources. The training data was carefully selected to ensure that the model learned the intricacies of human language and could generate high-quality responses to a variety of questions [6].

Chatbots like ChatGPT are particularly useful for businesses in

* Corresponding Author: Vimal Kumar, Assistant Professor, Department of Information Management, Chaoyang University of Technology, Taichung, Taiwan, Contact: +886-966620704, +886-4-23323000, Fax: +886-4-23742337
E-mail addresses: rohitraj2034@gmail.com (R. Raj), asingh6@jgu.edu.in (A. Singh), vimaljss91@gmail.com, vimalkr@gm.cyut.edu.tw (V. Kumar), way2pratima@gmail.com, pratima@iimk.ac.in (P. Verma).

https://doi.org/10.1016/j.tbench.2023.100140
Received 24 June 2023; Received in revised form 28 August 2023; Accepted 29 August 2023
Available online 1 September 2023

enhancing their customer service by providing instant support and assistance to customers [7]. Since ChatGPT is an automated platform, it can stay operational 24/7 eliminating the need for human customer service representatives to be available around the clock [8]. Customers can be provided with instant services with nearly perfect accuracy without having to wait for human representatives to respond [9]. ChatGPT can also handle multiple customer inquiries simultaneously, providing a scalable solution for businesses with large customer bases [10]. In addition to improving customer service, ChatGPT can prove instrumental in assisting organizations in streamlining their operations thereby reducing costs [11]. ChatGPT can allow human employees to focus on more complicated and strategic duties by automating routine tasks such as order monitoring and billing. Additionally, by reducing the need for extra people to handle customer concerns and requests, ChatGPT can assist cut operational expenses.

To leverage ChatGPT's capabilities effectively, businesses must carefully consider its use cases and limitations. It is paramount that businesses clearly understand how ChatGPT works, its strengths, and its limitations before implementing it [12]. ChatGPT is a language-trained model that is completely based on the raw data supplied to train it. Hence, it requires training data that is specific to the business domain. There are instances where ChatGPT outputs erroneous and vague results particularly involving domain-specific knowledge or context [13]. The possibility of misunderstanding the context thereby leading to erroneous responses is quite high. ChatGPT can learn biases from the training data that it is exposed to, which can lead to erroneous or discriminatory responses. For example, if the training data includes biased language or perspectives, ChatGPT may generate responses that perpetuate those biases [3]. Natural language is often ambiguous, and ChatGPT might encounter issues with disambiguating phrases or sentences, leading to erroneous responses [14]. The ability of ChatGPT to generate human-like responses can be exploited to generate fake news or phishing attacks [15]. Businesses must train and fine-tune ChatGPT to do specific activities, check its responses regularly, and fix any biases or inconsistencies in the training data to reduce these mistakes. It is also critical to establish clear boundaries and constraints for ChatGPT's use, as well as to put proper security measures in place to prevent malicious input [16].

The prospective benefits are considered and prioritized when using the PSI and COPRAS techniques. We employ PSI (Preference Selection Index) and COPRAS (Complex Proportional Assessment) approaches to prioritize the identified areas of deployment for ChatGPT's benefits in enterprises. They also consider potential benefits such as cost savings, Enhanced Customer Experience (ECE), and Greater Human-Computer Collaboration (GHC). However, there seems to be a need for more clarity and detail regarding the methodology used, especially in comparison to related works. The paper proposes the use of PSI and COPRAS approaches to prioritize the potential benefits and use cases of ChatGPT in business operations. The paper would benefit from providing a more detailed explanation of these approaches. This would help readers understand how these methods work, how they are applied to the context of ChatGPT, and how they contribute to the analysis.in this revised manuscript, we have explained the PSI and COPRAS, which were chosen as the analytical tools for prioritization. What specific advantages do these approaches offer in evaluating the potential benefits and use cases of ChatGPT compared to other methods? Highlighting the strengths of these approaches would enhance the rationale behind their selection. The paper mentions that ChatGPT's usage in businesses is presented in related work, but it does not elaborate on how these related works evaluate the benefits. To address this gap, the authors should explicitly compare their chosen methodology with those used in related works. What are the differences and similarities? How does the paper's approach contribute to a better understanding of the benefits compared to existing research? It's important for the paper to clearly outline its contributions compared to existing research. What novel insights or advancements does the paper bring to the field of using ChatGPT in business operations? How does the combination of PSI, COPRAS, and the specific context of ChatGPT distinguish this study from others? We have explained the methodology section, which provides a detailed step-by-step illustration of how PSI and COPRAS are applied to the context of ChatGPT's benefits and use cases. This includes the data collection process, criteria selection, analysis, and interpretation of results. The more detailed the explanation, the better readers can grasp the study's analytical framework. In a nutshell, the study seems to have a clear focus on analyzing the benefits and use cases of ChatGPT in business operations, it needs to provide more comprehensive explanations of the chosen methodology, highlight the differences from related work, and clearly articulate its contributions to the research field. This will enhance the paper's overall clarity and impact.

Overall, ChatGPT has found immense utility across various fields including language translation, chatbots, and content creation. ChatGPT is continually learning and enhancing its capabilities as a result of user interactions. Its uses include everything from customer service to language translation to creative writing. ChatGPT is an outstanding display of machine learning and artificial intelligence's capabilities in language processing. However, little research has been done to explore ChatGPT's potential benefits for boosting business operations, which represents a significant gap in the literature. This paper aims to fill this gap by identifying ChatGPT's possible benefits for businesses, discussing its strengths and limitations, and providing insights on how businesses might use ChatGPT's capabilities to improve their operations. We hope that this research will help enterprises make informed decisions about using ChatGPT by providing a better knowledge of its possible impact on business operations.

The remainder of the paper presents a detailed literature review and research methodology in Section 2 and Section 3 respectively. Section 4 presents a data analysis and results followed by the discussion and findings in Section 5 with the implications of this study. Finally, Section 6 outlines the conclusions with limitations and the future scope of this study.

2. Literature review

The recent limited literature on the subject of ChatGPT's usage in businesses is summarized in the following section.

In a study based on text classification, a corpus of 233,914 English tweets was analyzed using ChatGPT to identify the dominant themes which were collected within the first month of the launch of ChatGPT. Three dominant themes emerged namely news, technology, and reactions. The authors pointed out that the AI chatbot, ChatGPT can be effectively used in five functional domains including critical writing, essay writing, prompt writing, code writing, and answering questions. This research revealed that ChatGPT can have both positive and negative consequences on technology and humans. The major issues generated by the use of ChatGPT include job evolution, changes in the technological landscape, the pursuit of general artificial intelligence, and ethical considerations [17]. ChatGPT can provide efficient services including customer service applications and the creation of virtual assistants for voice and text conversations. It also offers topic detection, emotion detection, and sentiment analysis capabilities to enhance user understanding. It has a positive impact on digital marketing, e-commerce, healthcare, education, and finance [8]. In an interesting research in the domain of finance, it was found that ChatGPT can efficiently generate research studies that are plausible and useful, even in its basic state. The output can be further refined to better quality by adding private data and researcher expertise. Using the peer-review process, the evaluations of the created research give empirical verification of their potential contribution. The ethical implications of employing ChatGPT as a research instrument are unknown, and two points must be considered. On the one hand, ChatGPT might be viewed as a democratizing instrument capable of leveling the research production gaps between the global south and wealthy nations. Yet, it raises problems regarding the

correct credit and ownership of research conducted with its support [18]. The focus of the research article authored by [19] is on the application of ChatGPT for resolving programming bugs. This study explored the utility of ChatGPT in providing debugging aid, bug prediction, and bug explanation to address programming issues. Moreover, the paper emphasizes the incorporation of other efficient debugging tools and techniques to validate and verify the forecasts and explanations provided by ChatGPT [19]. AI has made significant strides in the field of radiology where GPT-based models are providing new opportunities to enhance accuracy, efficiency, and patient outcomes. These models are extensively used for report generation, educational support, clinical decision support, patient communication, and data analysis [20]. The research was conducted that investigated how ChatGPT may be utilized as a learning tool and the advantages and disadvantages it provides to students and teachers in communication, business writing, and composition courses. The researchers ran 30 ChatGPT tests and discovered that it has the potential to replace search engines due to its accuracy and dependability in presenting information to pupils. It also enables teachers to incorporate technology into their classes and hold workshops to analyze and evaluate produced replies. However, the study discovered that unethical usage of ChatGPT by students might result in human unintelligence and unlearning, posing a problem for instructors in measuring learning results. The research recommends teachers minimize using theory-based questions as take-home evaluations, give thorough case-based and scenario-based assessment assignments, use plagiarism detection software, and use ChatGPT-produced replies as examples in classrooms [12].

After reflecting deeply on the literature on ChatGPT and its variety of uses across sectors it can be concluded that the research on the assessment of the utility of ChatGPT in organizations is limited. The research is mostly confined to understanding the uses of ChatGPT in the education sector, healthcare, and academic research. While there is a growing interest in using ChatGPT to improve business operations, there is a gap in understanding how it can be used in specific business contexts. Hence, this study attempts to uncover the potential benefits ChatGPT can provide to businesses. While some research has been undertaken on the use of ChatGPT in certain business contexts, such as healthcare or customer service, additional research on how it may be utilized in other sorts of enterprises or sectors is required. How may ChatGPT be utilized in the industrial, banking, or hospitality industries, for example? What are the potential advantages and disadvantages of utilizing ChatGPT in these situations? Further study is needed to understand the numerous applications of ChatGPT in various sectors and scenarios. Table 1 shows the explanation of different aspects of Benefits and their sub-benefits.

Table 1
Explanation of different aspects of Benefits and their sub-benefits

Benefits	Sub- Benefits		References
Cost Savings (CS)	CS1	Increased Efficiency within a business	[8,19,21,22]
	CS2	Improved Accuracy within a business	[13,23,24]
	CS3	automate repetitive tasks such as answering frequently asked questions	[8,25–27]
Enhanced customer engagement (ECE)	ECE1	providing quick, informative, and more natural responses	[3,10,17,21,28]
	ECE2	leads to a more positive experience for the customer	[29–31]
	ECE3	increased customer satisfaction and loyalty	[28,29,32]
Generate high-quality content (GHC)	GHC1	Save businesses time and resources for content creation	[8,24,28,33]
	GHC2	ability to generate human-like text	[3,23,34]
	GHC3	personalize customer interactions and tailor responses based on the customer's preferences	[3,10,21,34,35]

3. Research Methodology

Each indicator's weight was determined step-by-step using the preference selection index (PSI) and complex proportional assessment (COPRAS) techniques to prioritize them. The process flow for the research is shown in Fig. 1. Using PSI and COPRAS, the advantages of ChatGPT have been rated in terms of improving business operations. The benefits and their sub-benefits criteria for measurement aspect is given in Table 1. The actual metrics and the demographics of the decision-makers taken into account for the research study are provided in Table 2, respectively.

3.1. Data collection and sample

The data is collected through Google questionnaire-based survey from product designers, service engineers, data scientists, programmers, researchers, and business development. All respondents belong to different positions including executive, supervisor, manager, and senior manager as shown in Table 2. This study emphasizes that the selection of respondents for our study was based on a rigorous criterion of expertise

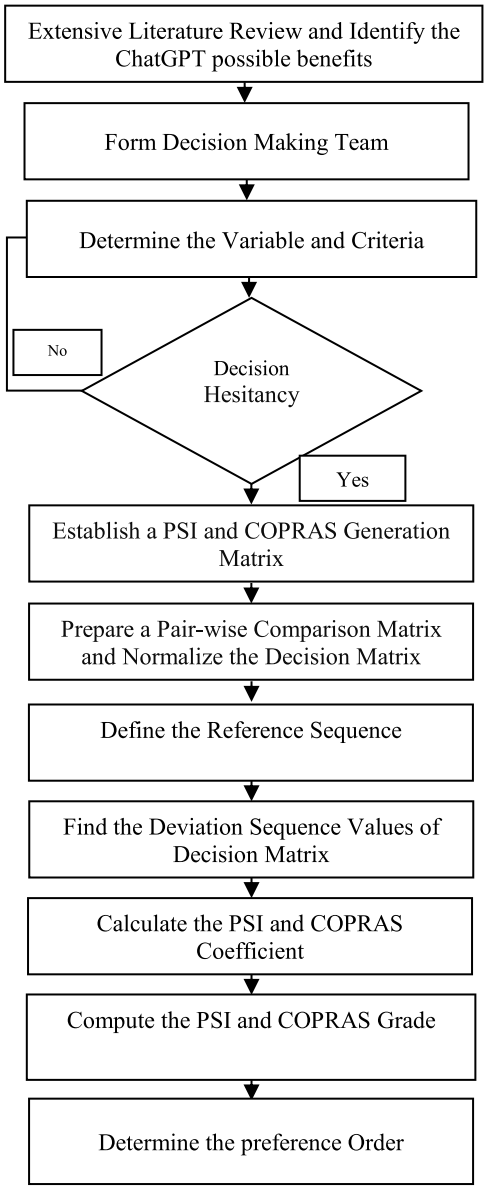

Fig. 1. The flow diagram of the research methodology

Table 2
Respondents' demographic details

Profile	*Classification*	*Count*
Sex	Female	8
	Male	7
Age	21-31	4
	32-41	7
	42-52	4
	Above 52	0
Denomination	Executive	3
	Supervisor	3
	Manager	3
	Senior manager	6
	Diploma	0
Education	Bachelors in Engineering	7
	Post Graduate in Computer Science	4
	Doctoral in technical education	4
	1-8 years	6
Present company tenure	9-17 years	5
	18-24 years	3
	above 24 years	1
	Product designer	3
Department of respondents	Service Engineer	4
	Data scientist	3
	Programmer	3
	Research and Business Development	2

within the field of generative AI technologies. Each of the fifteen respondents holds a distinguished track record in the field of generative AI technologies, and their insights are widely recognized as authoritative within the academic and professional community. In the context of our research objectives, we aimed to capture in-depth perspectives from a panel of recognized experts, allowing us to delve into nuanced aspects that are often challenging to access through larger-scale surveys ([36, 37]; et al., 2023; [38–41]). The intention was to prioritize the quality of responses over quantity, as these experts possess a wealth of knowledge and experience that greatly enriches our study [42–44]. The remark about the fifteen respondents is accurate, and this small sample size can help ensure the validity and strength of the study's findings. The sample is adequate in reflecting the population of interest and must be representative. The advice of experts would help to conclude. The outcomes are more likely to represent the group under study's actual features. With a small sample, it is important to recognize the restrictions and potential consequences of the study's sample size.

3.2. PSI Method

The PSI approach can be used to determine the objective weights of the various criteria. The following are the PSI method's steps: [45].

Step 1: Building the decision matrix (P) is the first step. This matrix is indicated by using Eq. 1.

$$P = \left[p_{ij}\right]_{m\times n} \tag{1}$$

The performance of the i th alternative on the j th criterion is shown in Eq. 1 by t_{ij}.

Step 2: Eq. 2 is used to perform the matrix's values' normalization.

$$p_{ij}^{*} = \frac{p_{ij}}{\max\left(p_{ij}\right)} \tag{2}$$

$$p_{ij}^{*} = \frac{\min\left(p_{ij}\right)}{p_{ij}} \tag{3}$$

Step 3: Using Eq. 4, the average values of the normalized matrix are calculated.

$$Q_{ij}^{*} = \frac{\sum_{i=1}^{m} P_{ij}^{*}}{m} \tag{4}$$

Step 4: Each alternative's preference variation value (δ_j) is calculated.

$$\delta_j = \sum_{i=1}^{m}\left[p_{ij} - Q_{ij}^{*}\right]^2 \tag{5}$$

Step 5: Calculated the preference value's deviation (θ_j).

$$\theta_j = \left|1 - \delta_j\right| \tag{6}$$

Step 6: The k_j criteria weights are computed.

$$k_j = \frac{\theta_j}{\sum_{j=1}^{n}\theta_j} \tag{7}$$

Step-7: Determine theP_j of for each option: – Each alternative'sP_j is provided as follows:

$$P_j = \sum_{i=1}^{m}\left[\delta_j \,\mathrm{x}\, k_j\right] \tag{8}$$

The significance ranks (priorities) of the alternatives are listed in increasing order of P_j value, i.e., the alternative with a higher $P_j$value has top importance than other alternatives.

3.3. COPRAS (Complex Proportional Assessment) Method

The "Complex Proportional Assessment" or COPRAS method was developed by Zavadskas and Kaklauskas in 1996. It was used to determine which alternative was superior to others and made it possible to compare them [46]. When more than one parameter needs to be considered in an evaluation, this method can be used to increase or decrease the number of criteria [47]. The COPRAS method ranks and evaluates choices in descending order according to their importance and utility [48]. The following steps are part of the COPRAS method:

Step 1. Decision matrix ($P = [a_{ij}]_{k\times l}$) is normalized by applying Eq. (9).

The normalized decision matrix is denoted by $N = [p_{ij}]_{k\times l}$. To compare all criteria, normalization seeks to produce various dimensionless values.

$$p_{ij} = a_{ij} \Big/ \sum_{j=1}^{l} a_{ij}\; i = 1, 2, \ldots, k; j = 1, 2, \ldots, l \tag{9}$$

Step 2. The weighted normalized decision matrix $Z = [z_{ij}]_{n\times m}$ was determined by applying Eq. (10).

$$z_{ij} = w_i a_{ij}\; i = 1, 2, \ldots, k; j = 1, 2, \ldots, l \tag{10}$$

Where p_{ij} is the normalized value of jth alternative according to ith criterion.

Table 3
Normalized Decision Matrix

CS	0.8	1	1	1	0.5	0.8	0.8	0.8	1	0.8	0.6	0.5	1	0.8	1
ECE	1	0.6	0.8	0.6	0.8	1	1	0.8	1	0.8	1	1	0.8	1	0.8
GHC	0.8	0.6	1	0.6	1	1	0.8	1	1	1	1	0.8	1	1	1

Table 4
The Results of PSI

δ_j	0.026	0.106	0.041	0.106	0.125	0.041	0.026	0.026	0.000	0.026	0.106	0.125	0.041	0.041	0.041
θ_j	0.973	0.893	0.958	0.893	0.875	0.958	0.973	0.973	1.000	0.973	0.893	0.875	0.958	0.958	0.958
k_j	0.068	0.063	0.067	0.063	0.061	0.067	0.068	0.068	0.070	0.068	0.063	0.061	0.067	0.067	0.067

Step 3. The weighted normalized values of the favorable and non-beneficial criteria were added. Calculating sums required the use of Eqs. (11) and (12).

$$Q_{+j} = \sum_{i=1}^{k} z_{+ij} \tag{11}$$

$$Q_{-j} = \sum_{i=1}^{k} z_{-ij} \tag{12}$$

where z_{+ij} and z_{-ij} are the weighted normalized values for the advantageous and unbeneficial criteria, respectively. The Q_{+j} the value increases and the Q_{-j} value decreases as the quality of the alternative increases. The values of Q_{+j} and Q_{-j} show how well each option has performed in terms of accomplishing its goals.

Step 4. The characteristics of the both positive and negative alternatives Q_{+j} and Q_{-j}, respectively, are described to establish the significance of the alternatives.

Step 5. The relative weighting or priorities of the options were determined. The priorities of the potential choices were established using E_j. With increasing E_j value, the alternative's importance grows. The relative relevance of a certain option reveals the extent to which it satisfies a desire. The best choice is the candidate alternative with the highest overall significance value (E_{max}). The comparative statistical significance of the j_{th} choice, E_j, was determined using Eq. (13).

$$E_j = Q_{+j} + \left(\left(Q_{-\min} \sum_{j=1}^{l} Q_{-j} \right) \Big/ \left(Q_{-j} \sum_{j=1}^{m} \left(Q_{-max} / Q_{-j} \right) \right) \right) \tag{13}$$

Where, $j = 1, 2, \ldots, l$ and $Q_{-\min}$ is the minimum value of Q_{-j}.

Step 6. The quantitative utility (U_j) for the j_{th} the alternative was calculated. An alternative's utility level is necessarily related to its relative significant level (E_j). One can assess an alternative's rank and degree of utility by comparing the efficiency rankings of all available possibilities. It is calculated using Eq. (14).

$$V_j = \left[\frac{E_j}{E_{\max}} \right] \times 100 \tag{14}$$

where $E_{\max}$ is the relative significance measure with the maximum value. The utility value of an option increases or decreases proportionally to its relative significance value. The more valuable V_j is, the higher the priority of the alternative. Depending on the utility ratings of the alternatives, a comprehensive ranking of the competing options can be created.

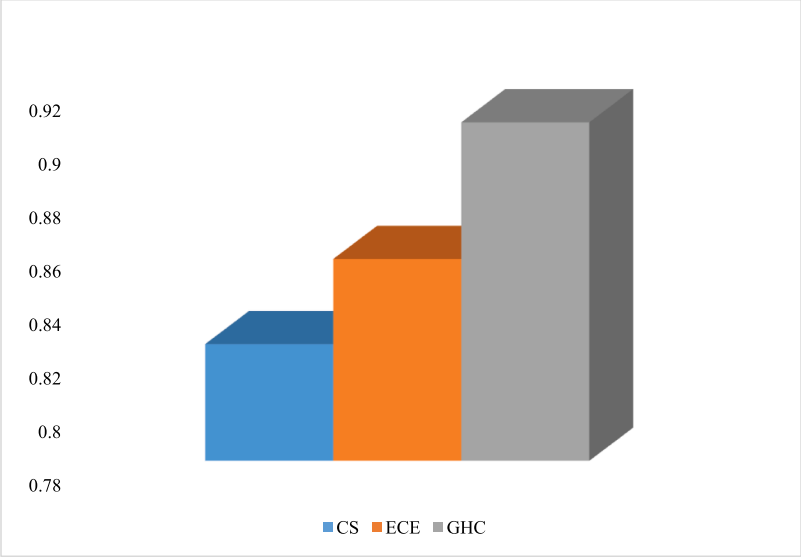

Fig. 2. Illustration of benefits parameters scores

Table 5
Final performance score

	P_j	Rank
CS	0.8236	3
ECE	0.8554	2
GHC	0.9063	1

4. Data Analysis and Results

4.1. PSI Results

Table 1 displays the PSI-based progress evaluation of the benefits indicators' ordering and prioritization. Think about how to prioritize signs for the PSI technique's validation based on other options. Eq. 1 is used to incorporate all collected data into a decision matrix. After the creation of the choice matrix for the benefits such as cost savings (CS), enhanced customer engagement (ECE), and generating high-quality content (GHC), respectively. For each indicator from Table 1, the normalized value was assessed using equations 2 through 5. The normalization decision matrix is displayed in Table 3. The values of δ_j, θ_j, and k_j were calculated using equations 5 through 7, respectively. The PSI results are shown in Table 4.

In light of scores of P_j, Fig. 2 shows the impact of performance reviews on the parameter. The consequences of different parametric values are shown in Fig. 2. Using Eq. 8, which highlights the actual results of the study, Table 5 provides the performance score for each of the three benefits. It demonstrates that option GHC has the highest performing score, with indicators ECE and CS coming in at second and last, respectively. As a result, based on the overall performance shown in Table 5.

4.2. COPRAS Results

Table 1 displays the results of utilizing COPRAS to evaluate the sub-benefits indicators' ranking and prioritization progress. In order to validate the COPRAS approach, consider prioritizing sub-benefits based on other options. After constructing the decision matrix for the sub-benefits such as increased efficiency within a business (CS1), improved accuracy within a business (CS2), automate repetitive tasks such as answering frequently asked questions (CS3), providing quick, informative, and more natural responses (ECE1), leads to a more positive experience for the customer (ECE2), increased customer satisfaction and loyalty (ECE3), Save businesses time and resources for content creation (GHC1), ability to generate human-like text (GHC2), and personalize customer interactions and tailor responses based on the customer's preferences (GHC3) respectively. Eq. (9) was used to calculate the normalized value for each of the sub-benefits from Table 1 that are shown in Table 6. By analyzing Eq. 10, Table 7 shows the weighted normalized choice matrix importance with indicators, with the enhanced optimal values illustrative of the value of all parameter weights.

The positive alternative sum value (Q_{+j}) was obtained using Eq. 11, while the negative alternative sum value was obtained using Eq. 12. (Q_{-j}). According to Eqs. 13 and 14, Table 8 displays the values of the two consolidated assessment scores, E_j, and V_j, along with ranking. Positive alternative sum value (Q_{+j}) and negative alternative sum value are used to evaluate this value data (Q_{-j}). Based on the two scores E_j, and V_j, Fig. 3 demonstrates the impact of performance assessments on the parameter. The consequences of various parameter settings are shown in Fig. 3. Table 8 summarizes the actual results of the study and provides the performance score for each of the nine sub-benefits. It demonstrates that criteria ECE3, GHC1, and CS2 have the poorest performance scores, whereas choice ECE1 has the highest performance score. As a result, based on the overall performance shown in Table 8.

5. Discussion and Findings

Providing quick, informative, and more natural responses (ECE1) under the category of Enhanced customer experience (ECE) is ranked first in the important features of ChatGPT that help boost business operations by accentuating customer satisfaction, improving customer retention, and ultimately, increasing revenue. ChatGPT may help organizations meet customer demands more efficiently and effectively by offering timely and informed responses to client inquiries or concerns, which can lead to improved levels of customer satisfaction [49]. Furthermore, the usage of natural language processing (NLP) technology may assist to make customer interactions feel more customized, which can aid in the development of better connections and increased customer loyalty [50]. Customers who feel satisfied and well-cared for are more inclined to do business with a firm again and may even suggest it to others. This can assist to boost client retention and attract new consumers, resulting in greater revenue and profitability for the company. Some examples of the aforementioned can be when customers need an instant response related to the day of the delivery of their

Table 6
Normalization of initial decision matrix for sub-benefits

CS1	0.097	0.128	0.121	0.097	0.097	0.097	0.133	0.069	0.147	0.121	0.100	0.118	0.083	0.125	0.121
CS2	0.097	0.128	0.061	0.097	0.097	0.161	0.100	0.103	0.118	0.091	0.100	0.147	0.111	0.063	0.152
CS3	0.129	0.103	0.121	0.129	0.129	0.065	0.167	0.138	0.088	0.121	0.133	0.118	0.139	0.125	0.091
ECE1	0.129	0.077	0.121	0.065	0.097	0.129	0.067	0.069	0.088	0.121	0.100	0.088	0.139	0.125	0.121
ECE2	0.097	0.128	0.121	0.161	0.161	0.097	0.100	0.138	0.147	0.121	0.100	0.088	0.139	0.125	0.121
ECE3	0.065	0.103	0.121	0.097	0.097	0.129	0.100	0.103	0.147	0.091	0.133	0.118	0.083	0.125	0.121
GHC1	0.161	0.128	0.121	0.129	0.129	0.129	0.100	0.103	0.118	0.121	0.133	0.088	0.111	0.094	0.091
GHC2	0.097	0.128	0.091	0.129	0.065	0.097	0.100	0.172	0.059	0.121	0.133	0.118	0.111	0.094	0.091
GHC3	0.129	0.077	0.121	0.097	0.129	0.097	0.133	0.103	0.088	0.091	0.067	0.118	0.083	0.125	0.091

Table 7
Weighted normalized decision matrix

CS1	0.0068	0.0064	0.0024	0.0019	0.0019	0.0029	0.0080	0.0021	0.0118	0.0024	0.0080	0.0094	0.0017	0.0125	0.0109
CS2	0.0068	0.0064	0.0012	0.0019	0.0019	0.0048	0.0060	0.0031	0.0094	0.0018	0.0080	0.0118	0.0022	0.0063	0.0136
CS3	0.0090	0.0051	0.0024	0.0026	0.0026	0.0019	0.0100	0.0041	0.0071	0.0024	0.0107	0.0094	0.0028	0.0125	0.0082
ECE1	0.0090	0.0038	0.0024	0.0013	0.0019	0.0039	0.0040	0.0021	0.0071	0.0024	0.0080	0.0071	0.0028	0.0125	0.0109
ECE2	0.0068	0.0064	0.0024	0.0032	0.0032	0.0029	0.0060	0.0041	0.0118	0.0024	0.0080	0.0071	0.0028	0.0125	0.0109
ECE3	0.0045	0.0051	0.0024	0.0019	0.0019	0.0039	0.0060	0.0031	0.0118	0.0018	0.0107	0.0094	0.0017	0.0125	0.0109
GHC1	0.0113	0.0064	0.0024	0.0026	0.0026	0.0039	0.0060	0.0031	0.0094	0.0024	0.0107	0.0071	0.0022	0.0094	0.0082
GHC2	0.0068	0.0064	0.0018	0.0026	0.0013	0.0029	0.0060	0.0052	0.0047	0.0024	0.0107	0.0094	0.0022	0.0094	0.0082
GHC3	0.0090	0.0038	0.0024	0.0019	0.0026	0.0029	0.0080	0.0031	0.0071	0.0018	0.0053	0.0094	0.0017	0.0125	0.0082

Table 8
Final performance score M_i for the criteria.

	E_j	V_j	Rank
CS1	0.17102625	99.802368	6
CS2	0.17087352	99.713244	9
CS3	0.1711948	99.900726	4
ECE1	0.17136492	100	1
ECE2	0.17116057	99.880749	5
ECE3	0.17092929	99.745785	7
GHC1	0.17092682	99.744346	8
GHC2	0.17125175	99.933955	3
GHC3	0.17127207	99.945815	2

orders. ChatGPT can provide an instant reply to the customers in a personalized way such as “Your order will be delivered in 3 days”. This helps customers in restoring confidence and trust in the organization. Similarly, the response by ChatGPT to the queries like “Can you tell me more about the return policy?” will be of the form “Our return policy allows customers to return items within 30 days of purchase for a full refund. However, some restrictions apply for certain products.” It can be readily observed that the response is sufficiently informative and provides adequate details to the customers regarding the return policy. Customers are more likely to engage with a business if they feel they are having a natural conversation, rather than interacting with a machine. ChatGPT’s NLP technology may help make interactions feel more human-like, which can lead to better customer connections [51]. For example, a consumer may enquire, "Do you provide any discounts?" and ChatGPT can answer naturally, "Sure, we offer a 10% discount to first-time clients." Do you want to know more?"

“Personalize customer interactions and tailor responses based on the customer’s preferences” (GHC3) under the category “Generate high-quality content” (GHC) acquires the second rank in the list of features of ChatGPT helpful in boosting business operations. Customers want to feel heard and valued by the companies they do business with [52]. By establishing personalized interactions with their customers and providing tailored responses based on preferences, organizations can instill confidence and trust in customers which will lead to increased customer satisfaction and loyalty. This will result in repeat purchases and positive word-of-mouth publicity and referrals thereby boosting business operations. When customers realize that they are being valued and respected they continue to do business with the companies. This leads to increased sales and revenue, as well as a larger customer base. Personalized interactions with customers can help businesses to improve their marketing efforts by providing deeper insights into consumer behavior and preferences. By understanding what customers need, companies can design more effective and targeted marketing campaigns leading to better conversion rates and higher returns on investment for marketing efforts. Organizations that can personalize customer interactions position themselves uniquely in the marketplace which differentiates these companies from competitors. This enables them to gain a competitive edge in the marketplace and unprecedented profits.

“Ability to generate human-like text (GHC2)” and “Improved Accuracy within a business (CS2)” under the categories of GHC and CS acquire third and fourth place respectively in the overall ranking of features of ChatGPT that are instrumental in boosting business operations. ChatGPT’s ability to create human-like text can help organizations to generate high-quality content more effectively and efficiently [53]. Whether it’s blog posts, social media updates, or product descriptions, ChatGPT can assist businesses to generate content that is more engaging, informative, and targeted to the intended audience. This can lead to increased website traffic, improved search engine rankings, and higher conversion rates [20]. By automating certain tasks, ChatGPT can prove immensely useful in cutting down costs that are incurred on services such as customer care. ChatGPT, for instance, can employ intelligent chatbots to provide customer assistance thereby eliminating the need to have humans do the job. This can help in the reduction of the cost that goes into paying the customer care executives. The automated services are comparatively more accurate than the traditional human-assisted services and thus the chances of errors are considerably reduced resulting in significant cost-savings. The ability of ChatGPT to respond in human-like text results in more personalized and engaging customer interaction. ChatGPT-powered technologies like chatbots can answer frequently asked questions (FAQs), provide recommendations, and resolve issues quickly and efficiently. By leveraging ChatGPT’S natural language processing capabilities, businesses can provide a more natural and human-like interaction, leading to higher customer satisfaction and retention rates [54]. Businesses may obtain a competitive edge in the market by exploiting ChatGPT’s capabilities. ChatGPT can assist businesses in improving their content creation, customer service, and accuracy, resulting in higher customer satisfaction and brand loyalty. Businesses may differentiate themselves from the competition and earn market share by delivering a superior client experience.

“Increased efficiency in the market (CS1)” and “Save businesses time and resources for content creation (GHC2)” are found to be the prominent features of ChatGPT that assist organizations in multiple ways to boost productivity. ChatGPT can help organizations leverage ChatGPT’s natural language processing capabilities to provide human-like responses to customers and make conversations more engaging and productive. The ability of the chatbot enabled by ChatGPT to handle customers’ queries and resolve issues quickly and efficiently can allow a significant amount of time for the employees to streamline and focus attention on addressing issues that are urgent thereby leading to better productivity [55]. Since this technology uses deep learning mechanisms

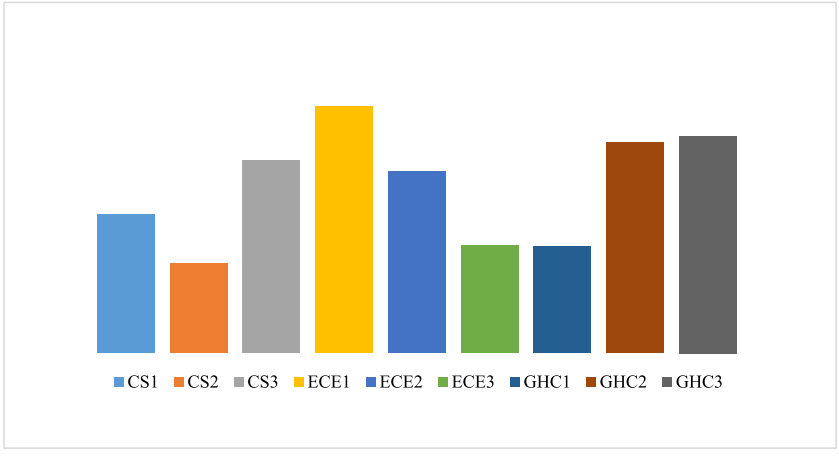

Fig. 3. Illustration of sub-benefits parameters scores

to design responses to queries based on historical data, it can be a useful resource to dig out important and relevant information pertaining to a given problem that occurs in organizations readily. In the case their limited time is available to make presentations to the client, ChatGPT's help can be used to create immensely interactive and informative ones to be presented to the clients. This is yet another way in which ChatGPT can enable employees to focus more on the tasks that require immediate attention. ChatGPT can produce high-quality written material on a wide range of topics. This can save organizations time and resources by eliminating the need for them to conduct their research and writing. ChatGPT may also edit and proofread current material to ensure that it is error-free and fits the requirements of the company. This can save businesses time and money by eliminating the need to recruit extra personnel or contractors to complete these activities. ChatGPT may assist enterprises with keyword research to help them optimize their content for search engines. This may save firms time and dollars by making their material more accessible to potential consumers. ChatGPT may also assist enterprises in improving the readability, clarity, and engagement of their existing material. Instead of having to generate new content from scratch, this can save businesses time and resources by boosting the efficacy of their existing material.

5.1. Implications of this study

Based on the findings of the current study, several practical implications are offered to the practitioners and management of business organizations. Some of the notable ones are as follows:

- ChatGPT may assist firms in more efficiently and successfully meeting customer expectations by offering rapid, informative, and natural solutions to client inquiries or problems. Management should think about using ChatGPT as a customer service tool to improve the customer experience and, eventually, revenue.
- Consumers who are happy and well-cared for are more likely to return to a company and may even promote it to others. ChatGPT's NLP technology may help make customer encounters feel more personalized, which can lead to stronger connections and higher customer loyalty. Managers should educate ChatGPT to respond to client inquiries in a customized manner, making them feel heard and appreciated.
- ChatGPT may give customers product and service information such as shipping timeframes, return policies, and discounts. Management should teach ChatGPT to deliver accurate and extensive product and service information to assist consumers to make informed selections. Management should routinely check ChatGPT's performance to verify that it is responding to consumer inquiries satisfactorily. They should also solicit consumer input to identify areas where ChatGPT might be enhanced.
- To offer consumers a consistent experience across all channels, ChatGPT should be connected with other customer care channels such as email and phone assistance. Customers shall receive the same quality of service regardless of the channel via which they contact the business.
- Managers and practitioners may save money by adopting ChatGPT to handle tasks like content development, editing, and proofreading instead of paying extra staff or contractors. This can result in substantial cost reductions for the firm.
- Managers and practitioners may obtain a competitive advantage in the marketplace by leveraging the possibilities of ChatGPT. This can assist firms in distinguishing themselves from the competition and attracting new clients.
- ChatGPT can assist corporate managers and practitioners in improving the efficiency of their operations. Organizations may minimize the time and effort necessary to perform activities by automating common processes and simplifying workflows, resulting in increased efficiency and productivity.
- The capacity of ChatGPT to handle numerous jobs at once makes it a great alternative for enterprises wishing to grow their operations. Organizations may increase their client base and operations without hiring extra staff by exploiting the advantages of ChatGPT.
- ChatGPT can assist company managers and practitioners in personalizing their interactions with clients. Organizations may tailor their offers and communications to better suit the requirements of their consumers by utilizing ChatGPT to collect data on client preferences and behavior.

5.2. Key contributions

The study emphasizes the potential benefits of incorporating ChatGPT, a popular chatbot built on a large-scale transformer-based language model, within companies. These advantages encompass enhanced customer service, multitasking capability for handling client inquiries, and operational cost savings. The study underscores the need for thorough analysis before integrating ChatGPT into enterprises. Factors such as domain-specific training data and potential errors in outcomes are highlighted as key considerations for successful deployment. The research draws from existing literature on ChatGPT, massive language models, and artificial intelligence to identify potential deployment areas. The utilization of PSI and COPRAS methodologies to evaluate benefits provides a structured approach for assessment. By elucidating current industry trends and potential advantages, the study offers valuable insights into the practical utilization of ChatGPT's capabilities to augment business operations and research endeavors.

6. Conclusions, limitations, and future scope of research

The study proposes to explore the benefits of ChatGPT that foster productivity in the operations of business organizations. Through extant literature review and consultation with experts, this research identified three benefits of ChatGPT which subsequently branched into nine sub-benefits.

The top three benefits revealed after analyzing the survey-based data belonged to the categories "Generate High-quality Content (GHC)" and "Enhanced Customer Engagement (ECE)". The analysis revealed that all three categories of benefits were significant in boosting business operations in their rite. "Providing quick, informative, and more natural responses (ECE1) under the category of Enhanced customer experience (ECE)", "Personalize customer interactions and tailor responses based on the customer's preferences" (GHC3), "Ability to generate human-like text (GHC2)", "Automate repetitive tasks such as answering frequently asked questions (CS3)", and "leads to a more positive experience for the customer (ECE2)" were the top 5 sub-benefits in the overall list of important benefits of ChatGPT in boosting business operations.

There are several limitations to this study which are mentioned below. The research makes no mention of any possible difficulties that may emerge during the implementation of ChatGPT in business processes. For example, problems concerning data protection, technical complexities, or client confidence may present difficulties. The research does not provide a complete cost-benefit analysis of ChatGPT implementation. The expense of implementing and maintaining the technology may be significant, and companies must balance the costs against the possible benefits. The study does not look into the possible risks of using ChatGPT to automate customer support. For example, if the chatbot is unable to correctly answer customer questions or duplicate a human-like exchange, it may create a negative customer experience.

The following could be the research's upcoming plan of action:

Conducting additional research into the feasibility of adopting ChatGPT in various business sectors, as well as finding best practices for tackling possible challenges. Conducting a thorough cost-benefit study to assist companies in making informed choices about ChatGPT implementation. One of the key advantages of ChatGPT fine-tuning lies in its ability to tailor the model's responses to specific industry domains or

niches. By exposing the model to domain-specific data and business jargon, further research can potentially create more contextually relevant and accurate responses. This fine-tuning process has the potential to significantly improve the model's ability to provide valuable insights, assist in decision-making, and enhance customer interactions within business applications. Moreover, conducting an additional study to investigate the possible risks and limitations of automating client support via ChatGPT, as well as developing strategies to handle these risks. The possibility of combining ChatGPT with other technologies such as machine learning, sentiment analysis, and voice recognition to provide a more complete and personalized user experience is being investigated. Evaluating ChatGPT's effect on client happiness, retention, and income, and developing methods to maximize the technology's impact.

Declaration of Competing Interest

The current manuscript titled "Analyzing the potential benefits and use cases of ChatGPT as a tool for improving the efficiency and effectiveness of business operations" is original work and never published before in any form anywhere. There is no conflict of interest associated with this manuscript with anyone to its publication. All the authors have given their significant contributions to this manuscript. I corresponding author to the manuscript declare that the manuscript has been read and approved by all the associated authors of the manuscript and they all consented to its submission.

References

[1] M.M. Mariani, I. Machado, S. Nambisan, Types of innovation and artificial intelligence: A systematic quantitative literature review and research agenda, Journal of Business Research 155 (2023), 113364.

[2] S.J.H. Shah, Chatbots for Business and Customer Support. Trends, Applications, and Challenges of Chatbot Technology, 2023, pp. 212–221, https://doi.org/10.4018/978-1-6684-6234-8.ch009.

[3] S.S. Biswas, Potential Use of Chat GPT in Global Warming, Annals of Biomedical Engineering (2023) 1–2, https://doi.org/10.1007/s10439-023-03171-8. Accepted.

[4] Mhlanga, D. (2023). Open AI in Education, the Responsible and Ethical Use of ChatGPT Towards Lifelong Learning. *Education, the Responsible and Ethical Use of ChatGPT Towards Lifelong Learning (February 11, 2023)*.

[5] J. Rudolph, S. Tan, S. Tan, ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? Journal of Applied Learning and Teaching 6 (1) (2023) https://doi.org/10.37074/jalt.2023.6.1.9.

[6] Y. Gao, W. Tong, E.Q. Wu, W. Chen, G. Zhu, F.Y. Wang, Chat with ChatGPT on Interactive Engines for Intelligent Driving, IEEE Transactions on Intelligent Vehicles (2023), https://doi.org/10.1109/TIV.2023.3252571.

[7] A. Shafeeg, I. Shazhaev, D. Mihaylov, A. Tularov, I. Shazhaev, Voice Assistant Integrated with Chat GPT, Indonesian Journal of Computer Science 12 (1) (2023), https://doi.org/10.33022/ijcs.v12i1.3146.

[8] A.S. George, A.H. George, A Review of ChatGPT AI's Impact on Several Business Sectors, Partners Universal International Innovation Journal 1 (1) (2023) 9–23.

[9] P. Tsigaris, J.A. Teixeira da Silva, Can ChatGPT be trusted to provide reliable estimates? Accountability in Research (2023) just-accepted.

[10] E.A. van Dis, J. Bollen, W. Zuidema, R. van Rooij, C.L. Bockting, ChatGPT: five priorities for research, Nature 614 (7947) (2023) 224–226.

[11] A.T. Gabrielson, A.Y. Odisho, D. Canes, Harnessing Generative AI to Improve Efficiency Among Urologists: Welcome ChatGPT, The Journal of Urology (2023) 10–1097.

[12] M.A. AlAfnan, S. Dishari, M. Jovic, K. Lomidze, ChatGPT as an Educational Tool: Opportunities, Challenges, and Recommendations for Communication, Business Writing, and Composition Courses, Journal of Artificial Intelligence and Technology (2023), https://doi.org/10.37965/jait.2023.01. Accepted.

[13] T.H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, V. Tseng, Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models, PLOS Digital Health 2 (2) (2023), e0000198.

[14] C. Zielinski, M. Winker, R. Aggarwal, L. Ferris, M. Heinemann, J.F. Lapeña, S. Pai, L. Citrome, Chatbots, ChatGPT, and Scholarly Manuscripts-WAME Recommendations on ChatGPT and Chatbots in Relation to Scholarly Publications, Afro-Egyptian Journal of Infectious and Endemic Diseases 13 (1) (2023) 75–79.

[15] J.V. Pavlik, Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education, Journalism & Mass Communication Educator (2023), 10776958221149577.

[16] M. Mijwil, M. Aljanabi, Towards Artificial Intelligence-Based Cybersecurity: The Practices and ChatGPT Generated Ways to Combat Cybercrime, Iraqi Journal For Computer Science and Mathematics 4 (1) (2023) 65–70.

[17] V. Taecharungroj, What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter, Big Data and Cognitive Computing 7 (1) (2023) 35.

[18] M. Dowling, B. Lucey, ChatGPT for (finance) research: The Bananarama conjecture, Finance Research Letters (2023), 103662.

[19] N.M.S. Surameery, M.Y. Shakor, Use Chat GPT to Solve Programming Bugs, International Journal of Information Technology & Computer Engineering (IJITC) 3 (01) (2023) 17–22. *ISSN: 2455-5290*.

[20] A. Lecler, L. Duron, P. Soyer, Revolutionizing radiology with GPT-based models: Current applications, future possibilities and limitations of ChatGPT. *Diagnostic and Interventional Imaging* (2023). Accepted, doi:10.1016/j.diii.2023.02.003.

[21] A. Haleem, M. Javaid, R.P. Singh, An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges, BenchCouncil transactions on benchmarks, standards and evaluations (2023), 100089.

[22] M. Halaweh, ChatGPT in education: Strategies for responsible implementation, Contemporary Educational Technology 15 (2) (2023).

[23] Y. Shen, L. Heacock, J. Elias, K.D. Hentel, B. Reig, G. Shih, L. Moy, ChatGPT and other large language models are double-edged swords, Radiology (2023), 230163.

[24] M. Sallam, ChatGPT Utility in Health Care Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns, Healthcare 11 (6) (2023) 887.

[25] S.B. Patel, K. Lam, ChatGPT: the future of discharge summaries? The Lancet Digital Health 5 (3) (2023) e107–e108.

[26] B.D. Lund, T. Wang, N.R. Mannuru, B. Nie, S. Shimray, Z. Wang, ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing, Journal of the Association for Information Science and Technology (2023), https://doi.org/10.1002/asi.24750. Accepted.

[27] S.A. Prieto, E.T. Mengiste, B. García de Soto, Investigating the use of ChatGPT for the scheduling of construction projects, Buildings 13 (4) (2023) 857.

[28] B. Rathore, Future of AI & Generation Alpha: ChatGPT beyond Boundaries, Eduzone: International Peer Reviewed/Refereed Multidisciplinary Journal 12 (1) (2023) 63–68.

[29] P. Korzynski, G. Mazurek, A. Altmann, J. Ejdys, R. Kazlauskaite, J. Paliszkiewicz, E. Ziemba, Generative artificial intelligence as a new context for management theories: analysis of ChatGPT, Central European Management Journal (2023), https://doi.org/10.1108/CEMJ-02-2023-0091. Accepted.

[30] T. Sakirin, R.B. Said, User preferences for ChatGPT-powered conversational interfaces versus traditional methods, Mesopotamian Journal of Computer Science 2023 (2023) 24–31.

[31] D. Singh, ChatGPT: A New Approach to Revolutionise Organisations, International Journal of New Media Studies (IJNMS) 10 (1) (2023) 57–63.

[32] H. Du, S. Teng, H. Chen, J. Ma, X. Wang, C. Gou, B. Li, S. Ma, Q. Miao, X. Na, P. Ye, Chat with ChatGPT on Intelligent Vehicles: An IEEE TIV Perspective, IEEE Transactions on Intelligent Vehicles (2023), https://doi.org/10.1109/TIV.2023.3253281. Accepted.

[33] S. Badini, S. Regondi, E. Frontoni, R. Pugliese, Assessing the Capabilities of ChatGPT to Improve Additive Manufacturing Troubleshooting, Advanced Industrial and Engineering Polymer Research (2023), https://doi.org/10.1016/j.aiepr.2023.03.003. Accepted.

[34] B.D. Lund, T. Wang, Chatting about ChatGPT: how may AI and GPT impact academia and libraries? Library Hi Tech News (2023) https://doi.org/10.1108/LHTN-01-2023-0009. Accepted.

[35] M. Aljanabi, ChatGPT: Future directions and open possibilities, Mesopotamian Journal of CyberSecurity (2023) 16–17. *2023*.

[36] S. Avikal, R. Pant, K.C.N. Kumar, V. Kumar, M. Ram, Prioritizing the Barriers of Manufacturing during COVID-19 using Fuzzy AHP, from the book title Advances in Soft Computing Applications, River Publishers, New York, USA, 2023, pp. 205–213, chapter 11ISBN 978-87-7022-817-6.

[37] S.M. Vadivel, A.H. Sequera, V. Kumar, V. Chandana, Performance Evaluation of Manufacturing Product Layout Design Using PROMETHEE II-MCDM Method, from the book title Intelligent Systems Design and Applications, Nature Switzerland (2023) 1–11, https://doi.org/10.1007/978-3-031-27440-4_24.

[38] A. Mittal, S. Sachan, V. Kumar, S. Vardhan, P. Verma, M.S. Kaswan, J.A. Garza-Reyes, Essential organizational variables for the Implementation of Quality 4.0: Empirical evidence from the Indian furniture industry, The TQM Journal (2023), https://doi.org/10.1108/TQM-06-2023-0189. Ahead-of-Print.

[39] V. Kumar, A. Mittal, P. Verma, J Antony, Mapping the TQM Implementation Approaches and their Impact on Leadership in Indian Tire Manufacturing Industry, The TQM Journal (2023), https://doi.org/10.1108/TQM-08-2022-0258. Ahead-of-Print.

[40] Singh, A., Kumar, V., & Verma, P. (2023), Sustainable Supplier Selection in a Construction Company: A new MCDM method based on Dominance-based Rough Set Analysis, Construction Innovation: Information, Process, Management, Ahead-of-Print. 10.1108/CI-12-2022-0324.

[41] A. Mittal, V. Kumar, P. Verma, A Singh, Evaluation of Organizational Variables of Quality 4.0 in Digital Transformation: The Study of an Indian Manufacturing Company, The TQM Journal (2022), https://doi.org/10.1108/TQM-07-2022-0236. Ahead-of-Print.

[42] G. Guest, A. Bunce, L. Johnson, How many interviews are enough? An experiment with data saturation and variability, Field Methods 18 (1) (2006) 59–82, https://doi.org/10.1177/1525822X05279903.

[43] R. Raj, V. Kumar, P. Verma, Big data analytics in mitigating challenges of sustainable manufacturing supply chain, Operations Management Research (2023) 1–15.

[44] V. Kumar, N.K. Sharma, A. Mittal, P. Verma, The Role of IoT and IIoT in Supplier and Customer Continuous Improvement Interface. Digital Transformation and Industry 4.0 for Sustainable Supply Chain Performance, Springer International Publishing, Cham, 2023, pp. 161–174.

[45] K. Maniya, M.G. Bhatt, A selection of material using a novel type decision-making method: Preference selection index method, Materials & Design 31 (4) (2010) 1785–1789.
[46] E.K. Zavadskas, A. Kaklauskas, T. Vilutiene, Multicriteria evaluation of apartment blocks maintenance contractors: Lithuanian case study, International Journal of Strategic Property Management 13 (4) (2009) 319–338.
[47] V. Podvezko, The comparative analysis of MCDA methods SAW and COPRAS, Engineering Economics 22 (2) (2011) 134–146.
[48] A. Kaklauskas, E.K. Zavadskas, J. Naimavicienė, M. Krutinis, V. Plakys, D. Venskus, Model for a complex analysis of intelligent built environment, Automation in construction 19 (3) (2010) 326–340.
[49] D. Gursoy, Y. Li, H. Song, ChatGPT and the hospitality and tourism industry: an overview of current trends and future research directions, Journal of Hospitality Marketing & Management 32 (5) (2023) 579–592.
[50] M. Javaid, A. Haleem, R.P. Singh, ChatGPT for healthcare services: An emerging stage for an innovative perspective, BenchCouncil Transactions on Benchmarks, Standards and Evaluations 3 (1) (2023), 100105.
[51] Sohail, S. S., Farhat, F., Himeur, Y., Nadeem, M., Madsen, D. Ø., Singh, Y., ... & Mansoor, W. (2023). Decoding ChatGPT: A Taxonomy of Existing Research, Current Challenges, and Possible Future Directions. Journal of King Saud University-Computer and Information Sciences, 101675.
[52] M.R. Kuchnik, Beyond Model Efficiency: Data Optimizations for Machine Learning Systems, Carnegie Mellon University, 2023. Doctoral dissertation.
[53] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, G. Kasneci, ChatGPT for good? On opportunities and challenges of large language models for education, Learning and Individual Differences 103 (2023), 102274.
[54] A. Haleem, M. Javaid, R.P. Singh, An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges, BenchCouncil transactions on benchmarks, standards and evaluations 2 (4) (2022), 100089.
[55] P. Budhwar, S. Chowdhury, G. Wood, H. Aguinis, G.J. Bamber, J.R. Beltran, A. Varma, Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT, Human Resource Management Journal (2023).

BenchCouncil Transactions on Benchmarks, Standards and Evaluations 3 (2023) 100138

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/

Research Article

MetaverseBench: Instantiating and benchmarking metaverse challenges

Hainan Ye [a], Lei Wang [b,*]

[a] University of Chinese Academy of Sciences, Beijing, 100049, China
[b] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

ARTICLE INFO

Keywords:
Metaverse
Systems
Benchmarks

ABSTRACT

The rapid evolution of the metaverse has led to the emergence of numerous metaverse technologies and productions. From a computer systems perspective, the metaverse system is a complex, large-scale system that integrates various state-of-the-art technologies, including AI, blockchain, big data, and AR/VR. It also includes multiple platforms, such as IoTs, edges, data centers, and diverse devices, including CPUs, GPUs, NPUs, and 3D glasses. Integrating these technologies and components to build a holistic system poses a significant challenge for system designers. The first step towards building the metaverse is to instantiate and evaluate the challenges and provide a comprehensive benchmark suite. However, to the best of our knowledge, no existing benchmark defines the metaverse challenges and evaluates state-of-the-art solutions from a holistic perspective. In this paper, we instantiate metaverse challenges from a system perspective and propose MetaverseBench, a holistic and comprehensive metaverse benchmark suite. Our preliminary experiments indicate that the existing system performance needs to catch up to the requirements of the metaverse by two orders of magnitude on average.

1. Introduction

In recent years, there has been increasing commercial interest in the metaverse. While the metaverse is still a developing concept, the term was first coined in Neal Stephenson's novel "Snow Crash" [1] published in 1992, referring to a shared virtual reality inhabited by millions of users with its economy, laws, and social interactions. For a long time, the metaverse was seen more as science fiction than something achievable until recently.

On the one hand, technologies enabling the metaverse have made considerable progress, including but not limited to artificial intelligence, blockchain, and extended reality. Specifically, artificial intelligence, especially deep learning and reinforcement learning, which have advanced significantly since the 2010s, has been crucial for developing the metaverse and is expected to be fundamental for realizing it. With the rise of blockchain technology, decentralization has become a vital feature of the metaverse. Improvements in devices and wearable technologies have also spurred growing interest in virtual and augmented reality among the general public. On the other hand, since 2020 and the global COVID-19 pandemic, online industries like online education have grown explosively. Analysts estimated the global online education market size at $210.1 billion in 2021 and predicted it would reach $848.12 billion by 2030 [2]. Online offices, gaming, and other industries have also seen similar growth. The rapid growth of these industries not only drives the development of relevant technologies but also promotes the evolution of the metaverse.

The metaverse is a complex interdisciplinary concept encompassing extensive technological domains and presenting challenges surpassing the capabilities of existing computing, storage, network, and other infrastructure. For example, Raja Koduri [3] has pointed out that providing real-time access to immersive computing for billions of people would require an increase in computing power of at least one thousand times from the current state-of-the-art, with real-time response latency of fewer than ten milliseconds. Therefore, the first step in designing a system that can meet the metaverse requires building a quantitative benchmark for metaverse systems.

However, existing benchmarks typically focus on specific technological domains. For example, MLPerf [4] and AIBench [5] aim to benchmark deep learning systems, while BigDataBench [6] aims to benchmark big data systems. The interdisciplinary nature of the metaverse means that existing benchmarks can only cover certain aspects of its related technological domains. Furthermore, the entanglement of various technologies significantly complicates the metaverse system. Therefore, Zhan [7] claimed that it is critical to propose a benchmark suite that quantitatively defines the challenges of the metaverse system and explores and evaluates state-of-the-art and state-of-practice solutions. Such a benchmark suite is necessary to systematically assess the metaverse system and address how far different technologies are from realizing the metaverse within the current computing, storage, and network capabilities.

* Corresponding author.
E-mail addresses: yehainan22@mails.ucas.ac.cn (H. Ye), wanglei_2011@ict.ac.cn (L. Wang).

https://doi.org/10.1016/j.tbench.2023.100138
Received 19 June 2023; Received in revised form 27 August 2023; Accepted 27 August 2023
Available online 4 September 2023

In this study, firstly, we summarize various definitions and concepts of metaverse by investigating existing literature, and we propose a comprehensive and sophisticated conceptual system of the metaverse from the perspective of computer systems. Secondly, we present a methodology based on the aforementioned conceptual system for benchmarking the metaverse system. Finally, we introduce an implementation of a benchmark suite based on this methodology, named MetaverseBench. Our contributions are as follows.

(1) We propose the metaverse conceptual system from the computer systems perspective, including three key aspects: components, technological domains, and specifications. The fundamental components include the access system, avatar, environment, and activity. We have summarized nine relevant technological domains: artificial intelligence, big data, extended reality, blockchain, cloud computing, edge computing, and networking. Furthermore, we distill five specifications to which the metaverse should adhere: automatic computing, immersive experience, decentralized architecture, ubiquitous access, and hyperspace interaction.

(2) We propose a benchmarking methodology for the metaverse system, which combines our conceptual system with the scenario-based approach proposed in [8]. Considering the complexity of the metaverse system, firstly, we build a typical metaverse application scenario and analyze its workflow, extracting critical paths, modules, and algorithms. Next, we select representative workloads based on nine technological domains to determine candidate workloads. Finally, we combine the results from the previous two steps to acquire the final workloads representing the designated scenario.

(3) We propose MetaverseBench, a benchmark suite for evaluating metaverse systems that conform to our conceptual system. Now, MetaverseBench comprises eight workloads corresponding to four components and nine domains. We also conduct experiments using MetaverseBench on a state-of-the-practice platform. The experimental results suggest that the existing platform requires an average of two orders of magnitude of performance improvements to support the metaverse.

This study is structured as follows: Section 2 reviews representative definitions of the metaverse. Section 3 introduces the metaverse conceptual system. Section 4 presents the methodology for benchmarking metaverse systems. MetaverseBench is presented in Section 5. Preliminary experiments under MetaverseBench are discussed in Section 6. Section 7 concludes related work, while Section 8 outlines the conclusions and plans for further research.

2. The metaverse definitions

A forward-looking research project that cannot be ignored, and the earliest systematic research project about the metaverse, is the "Metaverse Roadmap (MVR)" initiated by the Acceleration Studies Foundation (ASF) around 2006. In 2007, ASF published "Metaverse Roadmap: Pathways to the 3D Web", which provides a comprehensive overview of the potential of the metaverse and the pathways that may lead to its realization to report their research. In this study, we dig into ASF's report as a beginning. To explore and summarize up-to-date definitions and concepts of the metaverse, we investigate extensive literature, especially those published in recent years.

The definition of metaverse originates from a single 3D virtual world, gradually deriving into multiple interconnected virtual worlds and the fusion of reality and virtuality. In the ASF's report, John et al. [9] adopted the definition of the metaverse as "the convergence of virtually enhanced physical reality and physically persistent virtual space". Dionisio et al. [10] viewed the metaverse as "an integrated network of 3D virtual worlds". Lee et al. [11] considered the metaverse "a virtual environment blending physical and digital spaces". Ning et al. [12] defined the metaverse as "a new type of Internet application and social form that integrates a variety of new technologies" existing as a hyperspatiotemporal virtual world. PARK and KIM [13] summarized and compared the definitions of fifty-four papers published from 1992 to 2021 that specifically described the metaverse. Moreover, following the idea that the social value of Generation Z constructed the core of the contemporary metaverse, they proposed a new definition referring to the metaverse as "a three-dimensional virtual world where avatars engage in political, economic, social, and cultural activities". From the digital economy perspective, YANG et al. [14] viewed the metaverse as "a complete and self-consistent economic system, a complete chain of the production and consumption of digital items". Wang et al. [15] defined the metaverse as "a computer-generated world with a consistent value system and an independent economic system linked to the physical world". Dwivedi et al. [16] agreed with describing the metaverse as "the layer between you and reality". More specifically, the metaverse is viewed as "a 3D virtual shared world where all activities can be carried out with the help of augmented and virtual reality services". The summary is in Table 1.

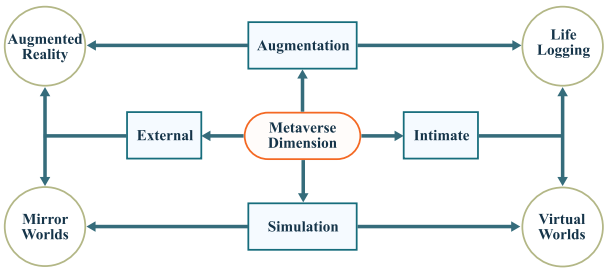

Fig. 1. Metaverse dimensions and categories.

Reviewing the definitions described above, we find that the metaverse involves virtuality and relies heavily on reality, constructing a bridge between the virtual and physical worlds. From this point of view, the version adopted by ASF [9] (i.e., "the convergence of virtually enhanced physical reality and physically persistent virtual space") elaborates the metaverse concisely and precisely. In the rest of this article, we use this definition.

3. Metaverse systems: Components, domains, and specifications

Researchers who engage in the metaverse debates are interested in identifying the essential concepts necessary for its construction. We propose a conceptual system encompassing three aspects: the components that make up the metaverse, the technological domains that enable the realization of the metaverse, and the specifications to which the metaverse should adhere. Despite numerous proposals for metaverse concepts in recent years, a comprehensive conceptual system that covers all three aspects has yet to be put forward. In this section, we will review the state-of-the-art concepts of the metaverse and present our conceptual system for the metaverse.

3.1. State-of-the-art concepts of metaverse

The ASF's report [9] published in 2007 is the first effort to provide a systematic viewpoint for analyzing the metaverse. As shown in Fig. 1, according to different dimensions determining how the metaverse evolves, John et al. [9] categorized the metaverse into the following four scenarios: "virtual worlds, mirror worlds, augmented reality, and lifelogging".

Dionisio et al. [10] argued that realism, ubiquity, interoperability, and scalability were decisive areas enabling the metaverse. Among the four areas, realism allows users to feel fully immersed; ubiquity facilitates users to access via various devices and maintains the identities of users; interoperability enables interaction across multiple virtual worlds; and scalability allows the metaverse to accommodate a massive number of users. Lee et al. [11] proposed a metaverse

Table 1
Representative metaverse definitions.

Definition	Year	Refs.
"The convergence of virtually enhanced physical reality and physically persistent virtual space."	2007	[9]
"An integrated network of 3D virtual worlds."	2013	[10]
"A virtual environment with duality blending physical and digital spaces."	2021	[11]
"A new type of Internet application and social form exists as a hyper spatiotemporal virtual world."	2021	[12]
"A three-dimensional virtual world where avatars engage in political, economic, social, and cultural activities."	2022	[13]
"A complete and self-consistent economic system, a complete chain of the production and consumption ofdigital items."	2022	[14]
"A computer-generated world with a consistent value system and an independent economic system linked tothe physical world."	2022	[15]
"A 3D virtual shared world where all activities can be carried out with the help of augmented and virtualreality services."	2022	[16]

Table 2
Key concepts of the metaverse.

Concept	Corresponding to the concept system	Refs.
Virtual world, mirror world, augmented reality, lifelogging. Avatar, environment.	N/A Component	[9]
Realism, ubiquity, interoperability, scalability.	Specification	[10]
Avatar, content creation, virtual economy. Social acceptability, security, privacy, trust, and accountability.	Component N/A	[11]
Multi-technology, sociality, hyper spatiotemporality.	Specification	[12]
Hardware, software, content.	Component	[13]
Economy, digital creation, digital asset, digital market, digital currency.	Component	[14]
Digital avatar, virtual environment, virtual goods/services. Immersiveness, hyper spatiotemporality, sustainability, interoperability, scalability,heterogeneity.	Component Specification	[15]
Immersive, boundless, connected.	Specification	[16]

ecosystem composed of "six user-centric factors: avatar, content creation, virtual economy, social acceptability, security, and privacy, and trust and accountability" to enable a self-sustaining, persistent, and shared realm. While an avatar is a vital element representing physical users, content creation and virtual economy are, respectively, activities and derivatives. Moreover, social acceptability, security, privacy, trust, and accountability correspond to social norms and regulations in the physical world. According to Ning et al. [12], the metaverse was characterized by multi-technology (as an internet application), sociality (as a social form), and hyper spatiotemporality (as a virtual world). PARK and KIM [13] also considered avatars as one of the core concepts of the metaverse. In addition, they divided the metaverse into hardware, software, and contents from the component perspective. Hardware refers to physical devices and sensors, software refers to recognition and rendering, and contents refer to scenarios and stories. YANG et al. [14] paid the most attention to the economy, claiming it to be the fundamental component of the metaverse. Furthermore, they stated that digital creation, digital assets, digital markets, and digital currency were the four components of the metaverse economy system. Wang et al. [15] proposed an architecture of metaverse integrating the human, physical, and digital worlds, in which digital avatars, virtual environments, and virtual goods/services supported the interconnected virtual worlds. They further refined six critical characteristics of the metaverse: immersiveness, hyper spatiotemporality, sustainability, interoperability, scalability, and heterogeneity. Dwivedi et al. [16] conceptualized metaverse building on contributions from twenty individual researchers. According to the conceptualization, the metaverse holds immersive, boundless, and connected features. Additionally, they aligned with the categories of metaverse scenarios presented in the ASF's report [9].

Based on the above discussions, we have summarized the keywords of state-of-the-art concepts in Table 2, categorized according to components, domains, and specifications. Despite the numerous studies on metaverse concepts, it is clear that a comprehensive and sophisticated conceptual system still needs to be improved.

3.2. Metaverse conceptual systems

We propose a comprehensive and sophisticated conceptual system of the metaverse, covering the three aspects of system components, technological domains, and specifications. Components are the essential elements of the conceptual system; the technological domains are the implemented technological, and specifications are the implemented standards. There is no real metaverse conceptual system, and the science fiction movie "Ready Player One" [17] explores the concept of a metaverse system, as the film takes place in a highly advanced virtual space called the "OASIS". So, in this section, we take OASIS as an example to elaborate on the conceptual system.

3.2.1. Components

According to the metaverse concepts and considering the aspect of system components, we divide a metaverse system into four critical components: access systems, avatars, environments, and activities (see Fig. 2).

Access Systems. An access system serves as a bridge between real users and the objective environment of the metaverse. While similar in functionality to the user login system of a game scenario, the metaverse access system is far more complicated in terms of access approaches and user experience. The access system of OASIS plays a crucial role in allowing users to enter and interact within the virtual world determining who can access the OASIS, how they can access it, and what permissions and privileges they have within the virtual environment. In its initial stage, the access system is expected to include two subsystems: core access and auxiliary access, each composed of corresponding hardware and software parts. Specifically, the core access subsystem is derived from wearable devices, with VR/AR/MR glasses serving as its most essential component, providing visual perception in the metaverse; the auxiliary access subsystem is necessary to meet the vast computing power need of the metaverse, with various end devices such as smartphones, tablets, laptops, and desktops providing auxiliary storage and computation capabilities, turning out to maximizes user convenience.

Avatars. An avatar is a digital representation of a real user in the metaverse, carrying their character role and identity. While the term avatar gained popularity after the movie "Avatar" was released, it has been widely used in account-based platforms for a long time. In recent years, various companies, led by Apple, have introduced capabilities for building avatars that are much more sophisticated, vivid,

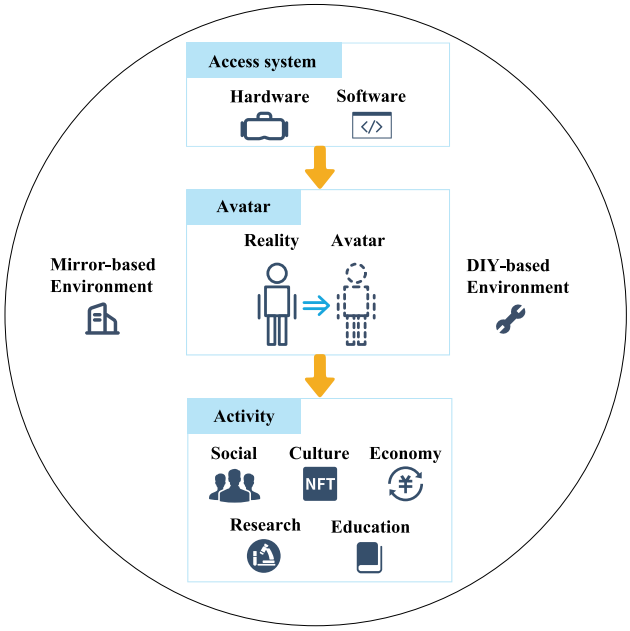

Fig. 2. Metaverse components.

and immersive than those created before. The concept of avatars in OASIS demonstrates their powerful role in enhancing user experiences within a metaverse. Avatars in the OASIS offer users a means of self-expression, enabling them to take on virtual personas and participate in diverse activities within the virtual reality universe. Considering the implementation approaches, theoretically, an avatar can mirror a real user, generally called a digital twin, or be a virtual character based on creation, called a digital native. Avatars in the metaverse should support digital twins and digital natives to satisfy different requirements from various application scenarios. In scenarios requiring accurate identity recognition, digital twins are suitable. In contrast, in entertainment scenarios like virtual games, digital natives and the fusion of digital twins and natives suggest broader application prospects.

Environments. Similar to the physical environment in the real world, the metaverse also requires a corresponding setting to carry out all the activities of the avatars. The metaverse environment is a 3D digital space designed to look and feel like a real-world environment. For example, OASIS is depicted as an expansive, interconnected virtual world featuring countless planets, zones, and domains. Each area within OASIS offers unique themes, landscapes, and challenges for users to explore. Considering the implementation approaches, the metaverse environment can mirror the real-world environment or be a completely DIY (Do-It-Yourslef) virtual environment. Similar to the implementation of avatars, the fusion of mirror-based and DIY-based approaches is also reasonable. And the need for these different types of metaverse environments is also to satisfy various application scenarios. Specifically, a DIY-based process is essential for building sufficiently immersive environments in gaming, learning, and work scenarios. In contrast, a mirror-based climate allows users to achieve almost the same experience as in the real world in systems such as sightseeing.

Activities. Just like humans conduct different kinds of activities in the real world, in the metaverse, interactions between other avatars and between avatars and the environment yield activities too. Activities within OASIS (the metaverse) are central to the plot and serve as the primary focus of the movie's narrative. These diverse and engaging activities reflect the vast possibilities that a fully realized metaverse system can offer. This study classifies the metaverse activities into four categories: sociality, economy, culture and entertainment, and education and research. Social activities are the most basic everyday activities in the metaverse, eliminating spatial constraints and language barriers. Economic activities involve concepts such as digital currency, digital assets, and digital market [15], with high reliance on decentralization and interoperability. Typical cultural and entertaining activities include literary and artistic creation, cultural tourism, and playing electronic games. For education and research, the metaverse provides platforms enabling immersive learning and teaching and interoperable experimental environments by applying extended reality and various sensors.

3.2.2. Domains

From the perspective of technological domains, we summarize the following nine elements as fundamental infrastructure that enable the realization of the metaverse and discuss several examples of how each element is applied in different components of the metaverse.

Artificial Intelligence. Relevant technologies based on deep neural networks have experienced tremendous progress in the last decade to become powerful driving forces for the development of the metaverse. For access systems, AI-based biometric identification technology can be applied to verify the identity of users as they attempt to log into the metaverse. For avatars in the metaverse, AI-based personalization algorithms can be used to create avatars that are more personalized to the individual user based on factors such as their preferences, interests, and behavior within the metaverse. This can make interactions with avatars more engaging and meaningful. For the metaverse environment, AI-based generation and reconstruction techniques can automatically create and populate vast and diverse backgrounds within the metaverse. For activities in the metaverse, AI-based natural language processing technology can enable users to interact with others and the metaverse environment, which helps create a more intuitive and user-friendly experience. It would allow users to quickly and easily access the necessary functions and features.

Big Data. The realization of the metaverse poses daunting challenges for the storage, transmission, and processing of big data, due to which big data is a necessary element. For access systems, big data can be used to monitor and analyze user access patterns and identify potential security threats, such as unauthorized access attempts or suspicious behavior. This can help to identify and prevent security breaches and enable the implementation of more effective access control mechanisms. For avatars in the metaverse, by collecting and analyzing data on user preferences, behaviors, and interests within the metaverse, it is possible to build detailed user profiles that can be used to personalize the avatar experience. This can include avatar appearance, behavior, interaction style, and the content and activities presented to the user. For environment and activities in the metaverse, it is possible to gain insights into how users interact with the environment and what features and activities are most popular by analyzing large datasets on user behavior and preferences. This can inform the development of new social features and activities and enable the creation of more engaging and interactive user experiences.

Data Security And Privacy. Data security and privacy are critical considerations in the metaverse, as users interact and engage within virtual environments that collect and process vast amounts of personal and sensitive information. For the metaverse access systems, robust access control mechanisms need to be implemented to restrict data access to authorized personnel only, and role-based access controls should be utilized to ensure that individuals can only access the data necessary for their specific roles. Data encryption must be adopted for other metaverse components to protect data during transmission and storage. Moreover, by utilizing anonymization and pseudonymization, we can minimize the use of personally identifiable information whenever possible, further protecting user identities and reducing the risk of data breaches.

Extended Reality. Extended reality refers to a family of technologies, including augmented reality (AR), virtual reality (VR), and mixed reality (MR). These relevant technologies generally function as

wearable devices. However, it limits human perception mainly to vision and hearing. The metaverse will gradually expand users' perception boundaries and bring more interactive possibilities. For the access system and environment of the metaverse system, users can access the metaverse and interact with the metaverse environment more intuitively and naturally without the need for traditional input devices such as keyboards or controllers. Furthermore, users can immerse themselves in the metaverse as if they were physically present in that environment. For avatars users can design and try on virtual clothing and accessories for their avatars, allowing for greater personalization and self-expression. For activities in the metaverse, extended reality can be applied to various activities within the metaverse, enhancing the user experience and making it more immersive, intuitive, and engaging. For example, users can access immersive educational or training content, allowing for more effective learning and skill development in a safe and controlled environment.

Brain–Computer Interface. While extended reality relies on additional external devices to function, the brain–computer interface (BCI) allows users to interact with the metaverse through neural signals. BCI can be applied in the metaverse in various ways, enhancing the user experience and interaction within the virtual world. BCI provides an alternative access method for the metaverse access systems that is more convenient and direct than the XR-based method. For avatars, BCI enables users to control their avatars within the metaverse using their neural signals directly. Instead of relying on traditional input devices like keyboards or controllers or XR-based devices, users can move, interact, and perform actions within the virtual world using their thoughts or intentions. For environment and activities in the metaverse, BCI can provide a more immersive and natural way of interacting with the metaverse environment. Users can perform actions in the metaverse, such as picking up objects or navigating through the virtual space, by simply thinking about those actions, creating a more intuitive and embodied experience.

Blockchain. Blockchain is expected to be used to establish the decentralized network in the metaverse. For access system by leveraging blockchain's ability to create a decentralized and secure identity system, users can have greater control over their digital identity and access to the metaverse environment without relying on a central authority or platform. For avatars, blockchain can be applied to manage the ownership and control of avatars, enabling users to have complete control over their avatars. For the metaverse environment, creating decentralized marketplaces with blockchain is possible. This can allow users to trade virtual assets directly with each other without the need for intermediaries or third-party platforms. For activities in the metaverse, blockchain can be applied to create non-fungible tokens (NFTs), representing unique and valuable digital assets such as virtual real estate, digital art, and collectibles. NFTs can be traded on blockchain-based marketplaces, providing users with a new way to engage in economic activities in the metaverse.

Cloud Computing. Cloud computing provides on-demand availability of computer system resources, especially data storage and computing power, without direct active management by users themselves. Scalable application scenarios and massive amounts of data in the metaverse require extremely huge computing and storage capacities, making cloud computing necessary. As fundamental infrastructure, cloud computing can be applied to the metaverse in specific ways. Firstly, cloud computing provides the scalability needed to accommodate the increasing number of users and their interactions within the metaverse, allowing for seamless user experiences even during peak usage times. Secondly, cloud-based data processing services can process and analyze large volumes of data generated within the metaverse, including user interactions, social behaviors, and market trends. Lastly, cloud-based content distribution services can distribute content, including 3D models, textures, and other digital assets, to users within the metaverse, improving the user experience and reducing latency.

Edge Computing. Edge computing is a distributed computing paradigm that brings computation and data storage closer to the data sources. This is expected to improve response latency and save bandwidth. Since the metaverse poses daunting challenges to computing and response delays, entirely using edge and end devices to provide auxiliary storage and computing capabilities is a promising solution. To be specific, on the one hand, edge computing can reduce the latency in the metaverse by placing computing resources closer to the end-user, thus reducing the round-trip time between the user's device and the central server. This can lead to a more responsive experience in the metaverse. On the other hand, edge computing can reduce the bandwidth requirements for the metaverse by processing data locally at the edge instead of sending it back to the central server. This can lead to cost savings for both end-users and service providers.

Network. Networks play a crucial role in enabling connection and communication within the metaverse. First, the metaverse relies on high-speed connectivity to help seamless communication and interaction between users. Therefore, networks must be designed to support high-bandwidth applications and low-latency connections. Moreover, the metaverse is expected to accommodate a large number of users, and as such, networks must be designed to be highly scalable to handle high traffic and data volumes. Last, deploying wireless networks is significant to enable users to access the metaverse anytime and anywhere.

3.2.3. Specifications

Although the exact specifications to which the metaverse should conform may evolve, we must summarize existing key elements that would be valuable for designing and evaluating metaverse systems. In this study, the following five aspects are considered.

Autonomic Computing. The metaverse is a persistent virtual world that remains available and accessible to users at all times, even when they are not logged in. In other words, the metaverse should be a self-running system parallel to the real world.

Immersive Experience. The metaverse offers a high level of immersion, allowing users to feel fully present in the digital space through the use of advanced graphics, haptic feedback, and other sensory experiences. Immersive experiences heavily rely on wearable devices such as AR/VR glasses, but the metaverse should expand its boundaries to include touch, smell, taste, and other perception approaches.

Decentralized Architecture. The metaverse is designed to operate distributed and decentralized without being controlled by any single entity or organization. In a decentralized metaverse, no single company or organization has complete control over the platform, the digital assets, or the user data.

Ubiquitous Access. The metaverse is designed to be accessible and available to users from anywhere, at any time, and through any device. In a ubiquitous metaverse, users can seamlessly transition between the physical and virtual worlds and between devices such as smartphones, computers, and VR headsets.

Hyperspace Interaction. The metaverse enables seamless communication and interaction between applications, platforms, and digital spaces, even between the metaverse and the physical world. In an interoperable metaverse, users can transfer and use digital assets, identities, and experiences across different environments and contexts. In a hyperspace-enabled metaverse, users can move from one virtual environment to another without noticeable lag or disruption, creating a seamless and immersive experience.

We summarize our metaverse conceptual system in Fig. 3.

4. Benchmarking the metaverse

According to Zhan [7], definition, instantiation, and measurement are three essential processes of a benchmark (see Fig. 4). In this section, we propose the problem definition of the metaverse benchmark and introduce our methodology for benchmarking the metaverse. First, we define the problem of benchmarking the metaverse as: **Quantifying design/tune challenges for the metaverse conceptual system.** Then, we introduce our metaverse benchmark methodology.

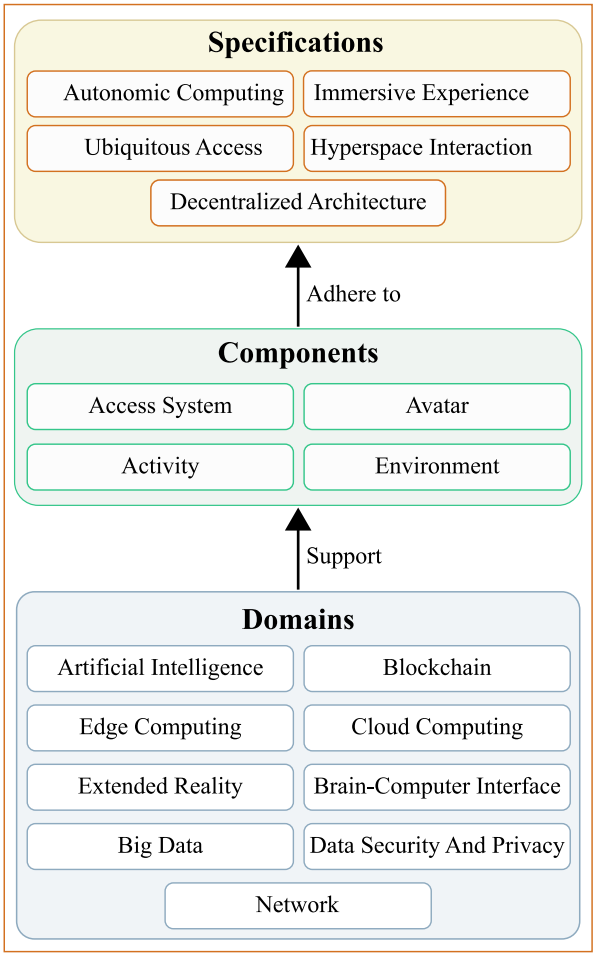

Fig. 3. Metaverse conceptual system.

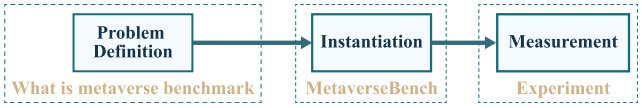

Fig. 4. Metaverse benchmark roadmap.

4.1. The metaverse benchmark methodology

We combine the metaverse conceptual system with the scenario-based approach proposed by Gao et al. [8] to build our methodology for the metaverse benchmark. As depicted in Fig. 5, firstly, we select three representative application scenarios of the metaverse. Secondly, we determine the common critical path by analyzing how each metaverse component functions in these scenarios and summarize several vital elements. Thirdly, viewing the essential elements from the perspective of underlying technologies, we determine candidate workloads by extracting representative ones corresponding to the metaverse domains. Finally, for the benchmark implementation, we further build a reduced set from the set of candidate workloads for the critical path of corresponding scenarios. Moreover, we refer to the specifications part of the metaverse concept system to determine the final workloads and related metrics.

4.2. The metaverse scenarios

Office, education, and entertainment are three primary activities in human society. The Internet era has accelerated the forms of these activities to shift from offline to online. What is certain is that they will still constitute the most basic application scenarios in the metaverse.

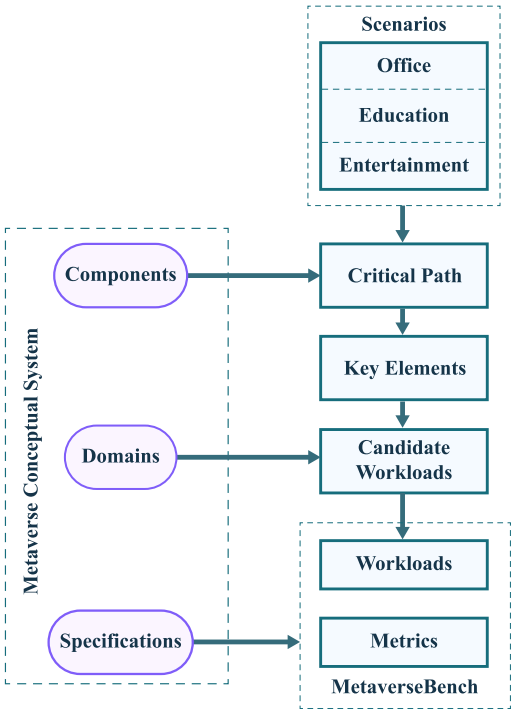

Fig. 5. Metaverse benchmark methodology.

Office. Metaverse offers a more immersive experience for online offices than traditional internet applications, allowing individuals to work remotely through extended reality while interacting with colleagues in the shared online space as avatars. Various companies have developed related products and services, such as Microsoft's virtual collaboration platform Mesh [18] and Meta's Horizon Workrooms [19]. By utilizing VR devices, Horizon Workrooms enables users to bring their desks, computers, and keyboards into the virtual world for work.

Education. Generally, the education industry involves three elements: teachers, students, and learning environments. In the metaverse education scenario, students could interact with their environment, collaborate with classmates, and engage in experiential learning activities. For example, they could explore historical landmarks, visit foreign countries, or participate in simulated experiments that might not be possible in the physical world.

Entertainment. The entertainment industry is anticipated to play a significant role in the metaverse. For instance, the movie "Ready Player One" showcases a game experience that shatters traditional geographical restrictions, facilitates instant scene switching, accommodates unlimited user capacity, and provides a low-cost immersive experience. Another example is Faye Wong's Fantastic Music concert in 2016, which utilized VR to give a three-dimensional online experience that went viral online.

4.3. The critical path and key elements

Upon examining the three primary application scenarios of office, education, and entertainment in the metaverse, we summarize the critical path of these scenarios as follows: users are granted access to the metaverse environment through the access system; following successful authority, users then operate within the metaverse environment in the form of avatars; interactions between avatars and the overall environment precipitate a range of activities encompassing various social, economic, and cultural fields. Considering how the

Table 3
Candidate workloads for MetaverseBench.

Key elements	Workloads
User authentication	Fingerprint recognition Face recognition Voice recognition
Creation and rendering of models	3D reconstruction Graphics rendering
Technologies facilitating interactions	Semantic segmentation Image classification Object detection Image generation Text classification Question answering Machine translation Speech recognition
Decentralized networks	Proof of Work Proof Of Stake Proof of Space
Big data processing	Read, write Sort, grep Data caching Media streaming

metaverse components function in the critical path, we can get the key elements of each component. The part that most affects the user experience in the access system is **user authentication**. For avatars and environments, sophisticated models enable users to feel immersive, so we focus on the **creation and rendering of models**. Activities are the results of interactions between different avatars and between avatars with environments. We focus on not only **technologies that can help facilitate those interactions** but also the **performance of decentralized networks** based on blockchain to ensure a consistent experience. Moreover, as the essential part, **the capabilities for storage, computing, and transmission of big data** are also considered.

5. MetaverseBench

We present MetaverseBench as our solution instantiation for the metaverse benchmark. For most benchmark suites, workloads, datasets, and metrics are the three fundamental elements that apply to MetaverseBench. Moreover, selecting datasets and metrics depends on specific workloads; the paramount step is determining the workloads.

We adopt a three-step process to determine the workloads. Firstly, we follow the critical path and key elements of general scenarios. We select representative workloads from the nine technological domains described in the conceptual system to form a set of candidate workloads. Secondly, we conduct an in-depth analysis of general scenarios and build the mapping from specific workloads to them by reviewing the workflow of metaverse components. Finally, we extract and refine workloads that cover the critical path and the components from candidates to reflect general metaverse scenarios as realistically as possible.

5.1. Candidate workloads

We utilize the critical elements concluded in Section 4.3 and refer to the metaverse technological domains to obtain the candidate workloads. In particular, we refer to typical benchmarks in these domains for selection. Table 3 lists our candidate workloads according to the essential elements.

Various methods have been applied for user authentication, such as username/password authentication, token-based authentication, and multi-factor authentication. Biometric authentication is being adopted widely since it offers a more secure, convenient, and reliable experience than traditional methods. We take different types of biometric identification, such as fingerprint recognition, face recognition, and voice recognition, as part of our candidate workloads. Creation and rendering of models comprise various processes within which 3D modeling using specialized software like Blender [20], real-time streaming for loading models, and real-time rendering with GPUs are fundamental. We respectively include 3D reconstruction, media streaming, and rendering performance across multiple graphics APIs (OpenGL, Vulkan, DirectX, etc.) into the candidate workloads. As for technologies facilitating interactions, we focus on underlying AI-based algorithms. Specifically, we select representative algorithms in computer vision: semantic segmentation, image classification, object detection, and image generation, and in natural language processing, text classification, question answering, machine translation, and speech recognition as candidate workloads. Decentralized networks based on blockchain involve various aspects such as consensus protocols, smart contracts, governance mechanisms, security management, etc. Mainly, consensus algorithms are what we are concerned about the most. Therefore, relevant tasks like "Proof of Work" (PoW) and "Proof of Stake" (PoS) are included in our candidate workloads. Moreover, we include representative workloads: read, write, sort, grep, and data caching, for evaluating capabilities of storage, computing, and transmission provided by the overall hardware and software infrastructure for handling big data.

5.2. Scenarios mapping and selected workloads

Corresponding to three scenarios (Office, Education, and Entertainment), we abstract the minimum set of workloads from candidate workloads. We consider mapping from the workloads to general metaverse scenarios. By building the mapping, we can ensure that our benchmark suite accurately reflects real-world scenarios and the performance requirements of the metaverse. We follow the critical path across different components to construct the mapping. Firstly, we choose the face recognition workload to inspect user authentication. According to a report from Frost&Sullivan [21], face recognition held over twenty percent of the global biometrics market in 2021. Therefore, face recognition suits three scenarios and is a representative workload for the access systems of the metaverse. Secondly, we choose the 3D reconstruction workload to inspect the creation and rendering of models. As mentioned in Section 3.2, the creation of avatars and environments in the metaverse both support the mirror-based approach, which heavily relies on the application of 3D reconstruction. Moreover, model creation consumes considerable resources in the components of avatars and environments. It is representative to choose 3D reconstruction to represent the model creation of avatars and environments. Thirdly, we use machine translation and speech recognition workloads to represent technologies facilitating interactions in the metaverse. The two technologies are state-of-the-art interaction technologies and can be utilized to break boundaries among users of different native languages and cultural backgrounds to maximize user experience. Fourthly, we use the Proof of Work (PoW) workload to represent decentralized networks. The choice of consensus algorithm depends on the goals of the blockchain network. The blockchain space has evolved to include various consensus algorithms that address different scalability, efficiency, and security considerations. Since PoW is historically significant due to its role in Bitcoin's creation [22], we include it to check the performance of decentralized networks. Lastly, we include sort, grep, and media streaming workloads for big data processing. The sort workload is indispensable in data organization, promoting searching efficiency, ensuring aggregation and analysis, and helping data deduplication. The grep workload is crucial for quick data retrieval, filtering, extraction, and cleansing. Both operations are essential for efficiently managing and processing big data. For media streaming workload, on the one hand, users need to utilize it to access the metaverse environment. On the other hand, users can conduct various activities like online meetings, attending classes, and watching videos in scenarios like office, education, and entertainment by media streaming.

Table 4
Workloads of MetaverseBench.

No.	Workloads	Key elements	Components	Metrics
1	Face recognition	User authentication	Access system	Accuracy; latency
2	3D reconstruction	Creation and rendering of models	Avatar; environment	IoU; latency
3	Machine translation	Technologies facilitating interactions	Activity	BLEU
4	Speech recognition	Technologies facilitating interactions	Activity	WER; F1-score
5	Proof of work	Decentralized net	Access system	Block confirmation time
6	Sort			
7	Grep	Big data processing	Environment; activity	Throughput
8	Media streaming			

We summarize the final selected eight workloads of MetaverseBench as shown in Table 4. We inspect different evaluation metrics for various workloads. For face recognition, we care about recognition accuracy and latency. The latency is necessary for Real-Time applications and user experience, and the recognition accuracy is essential for reliable identification. We also need to strike a balance between recognition accuracy and latency. For 3D reconstruction, except for latency, we check the intersection of union (IoU) for evaluating model quality. The latency is a critical performance metric for 3D reconstruction algorithms, especially when considering real-time or time-sensitive applications. The IoU is a widely used metric for measuring the accuracy of 3D reconstruction results, particularly in the context of comparing the reconstructed 3D model to a ground truth or reference model. For machine translation, bilingual evaluation understudy (BLEU) is adopted. The BLEU metric is an essential metric for evaluating the quality of machine translation systems. It is widely used in natural language processing and machine translation. We use word error rate (WER) and F1-score for speech recognition. Using WER and F1-score together allows a comprehensive assessment of speech recognition workload. While WER focuses on word-level errors and overall accuracy, the F1-score accounts for precision and recall, providing insights into how well the system handles correctly and incorrectly recognized words. For proof of work (PoW), block confirmation time is what we are concerned about. It refers to the time it takes for a new block to be added to the blockchain after being successfully mined by a miner, involving trade-offs between security, throughput, user experience, and the economics of the blockchain. For sort, grep, and media streaming, we inspect throughput to reflect the performance of storage, computing, and transmission. Throughput reflects the rate at which the workloads can process data and is often measured in terms of records per second or data size per unit of time.

6. Preliminary experiments

To illustrate the challenges of the metaverse to the point, we construct a concise metaverse scenario based on MetaverseBench and summarize the challenges based on evaluations. In the concise metaverse scenario, we only consider the minimum system requirement, which is constructed by four workloads corresponding to the metaverse components. Although the concise scenario cannot completely summarize the whole picture of the metaverse system, preliminary experiments on designated workloads can quickly clarify the gaps between the performance of state-of-the-art systems and that of metaverse systems.

6.1. The concise scenario

To construct the concise scenario, we choose four workloads from MetaverseBench: face recognition, 3D reconstruction, media streaming, and sort. As demonstrated in Fig. 6, face recognition is adopted to represent the metaverse access system; 3D reconstruction is adopted to describe the construction of both avatars and the metaverse environment; media streaming is adopted as a typical workload in all kinds of activities in the metaverse; sort is adopted to represent overall data processing of the metaverse system.

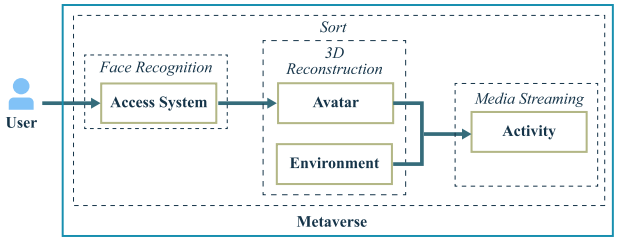

Fig. 6. Workloads in the concise scenario.

In the concise scenario, we assumed that a single server node serves a thousand users. The overall design target is to satisfy the concurrency, which is also the typical Internet service mode. We examined the latency or throughput for each workload to determine whether the system could meet the requirements. Specifically, we focus on latency for face recognition and 3D reconstruction, while for media streaming and sorting, we focus on throughput.

We further set the baseline performance for four workloads. For face recognition workload, we refer to smartphones' face recognition unlocking process and set its latency requirement to be no more than 3 s for a single user. For the 3D reconstruction workload, given that the latency of mainstream XR devices is generally under 50 ms and Apple Vision Pro makes it as low as 12 ms, it is reasonable that we set its latency requirement to be ten milliseconds. To enable an immersive experience, we take 4K videos as media sources for the media streaming workload. Specifically, the video parameters are resolution 3840*2160, frame rate 24 fps, video codec H.264, and bitrate 40 Mbps. Therefore, the throughput requirement of the media streaming workload is 5 MB per second per user. For the sort workload, the throughput requirement of the metaverse system is 1 GB per second. In our preliminary experiments, we only evaluated the performance of the state-of-the-practice system for each workload separately. All performance requirements are summarized in Table 5.

6.2. The preliminary results

We conducted preliminary experiments on a single server equipped with Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40 GHz CPU and NVIDIA Tesla V100 GPU. The testbed is a typical state-of-the-practice platform.

Face Recognition. We conducted offline inference on the LFW (Labeled Faces in the Wild) dataset [23]. At the same time, the pre-trained model used was a TensorFlow implementation of Google FaceNet [24] with the architecture of Inception ResNet v1, which was trained with VGGFace2 dataset [25] under the V100 GPU. The inference process took 9 min and 13 s on 13233 images and showed an accuracy of more than 0.99. This implied that the platform could deal with nearly 24 images per second. On the other hand, the latency requirement of the metaverse system for face recognition is 3 s. Therefore, it suggested that a state-of-the-practice solution is enough to meet the need.

Table 5
Preliminary results.

Workloads	Metrics	State-of-the-practice results	Requirements	Gaps
Face Recognition	Latency under Accuracy	0.05 s under 0.995 per user	3 s under 0.99 per user	No gap
3D Reconstruction	Latency under IoU	1800 s under 0.9 per scene	0.01 s under 0.8 per scene	180,000 times
Media Streaming	Throughput	0.15MB/s per user	5MB/s per user	33.3 times
Sort	Throughput	20.3MB/s per 1000 users	1GB/s per 1000 users	50.4 times

3D Reconstruction. Since scene reconstruction could be more challenging than object reconstruction due to the need to reconstruct multiple objects and their relationships within the scene, we focus on scene reconstruction for now. Specifically, we conducted an evaluation on SceneNet dataset [26] with the open source POCO pre-trained model [27], which was trained with ShapeNet dataset [28]. The inference results showed that the POCO model achieved a considerably fine reconstruction quality, while the average time consumed per scene was about 30 min. On the other hand, the latency requirement of the metaverse system for 3D reconstruction is ten milliseconds, and the gap is 180,000 times.

Media Streaming. We conducted this workload experiment with docker images released by CloudSuite [29]. During the running process, media streaming created four concurrent clients while each client held no more than 500 sessions (each session represented one user). The total throughput for all clients was about 292.4 MB per second. In other words, the throughput was 0.15 MB per second per user. On the other hand, the throughput requirement of the metaverse system for Media streaming is 5 MB per second per user, and the gap is about 33.3 times.

Sort. We adopted results in BigDataBench [6] as a reference to evaluate the state-of-the-practice performance of our big data workloads. BigDataBench conducted sort operations using a 32 GB unstructured Wikipedia data set of 4,300,000 English articles on a typical state-of-the-art processor, Intel Xeon E5645. The results showed the throughput was about 20.3 MB per second. On the other hand, the throughput requirement of the metaverse system for the sort workload is 1 GB per second, so the gap is about 50.4 times.

6.3. Summary

Table 5 summarizes the gaps between the performance requirements of the metaverse system and those of the state-of-the-practice system. Our evaluations show that to achieve the performance requirements of the metaverse system; the state-of-the-practice system performance needs to catch up by two orders of magnitude on average. The smallest gap is face recognition, whose state-of-the-art performance can meet the requirement, and the most significant gap is the one of 3D reconstruction, which is five orders of magnitude. So, state-of-the-art technology needs more revolutions to achieve the performance requirements of the metaverse system. Besides performance, we also conclude some requirements for metaverse system designs.

Fitting Various application scenarios. The metaverse involves a lot of application domains. Many real-life activities, such as business, social, education, finance, medicine, meetings, and games, can be mapped to the virtual world. These different application domains have other application characteristics and technical requirements. So, the metaverse system should define a set of standard interfaces and specifications to fit these different domains.

Providing Stronger interaction. In the metaverse, the ways of interaction will be more diverse. Users can issue instructions through handheld devices, head-mounted devices, etc.; the machine can also capture the user's actions and language through cameras and microphones. In addition, the user's brain turbulence, heart rate, blood pressure, breathing, and environmental information can also be obtained through sensors. Different types of precision sensors make the interaction between the user and the machine smoother. At the same time, through smart glasses, seats, projection equipment, and other output devices, technologies such as virtual reality and augmented reality can be used to make users immersive.

Using more edge or end devices to implement stronger interactions the metaverse collects user and environmental data through tremendous and heterogeneous sensors. These collected data have various formats, including images, videos, voices, etc. These data need to be quickly and accurately identified and processed accordingly. Traditional Internet applications often send user requests to servers, parse requests, and process data in servers. In the metaverse, some simple sensor data processing tasks can be performed in end and edge devices, while complex tasks are sent to the server for processing. Currently, task allocation and scheduling are not limited to servers in the data center but must be performed on ends or edges.

7. Related work

Although benchmark evaluations exist for the related technological domains involved in the metaverse, such as XRBench [30] for evaluating the performance of Machine-Learning hardware for future Extended Reality systems and BigDataBench [6] for evaluating big data systems and architectures, creating benchmarks for complete metaverse systems remains uncharted territory. The Hyperledger Foundation [31] introduced Hyperledger Caliper, a benchmark tool for blockchain. At the same time, in 2020, Dimitri et al. [32] developed BCTMark, a generic framework for benchmarking blockchain on an emulated network in a reproducible way. Additionally, there are benchmarks available for Artificial Intelligence, such as MLPerf [4] and AIBench [5]. However, building benchmarks against complete metaverse systems is still uncharted.

8. Conclusion and plan

Metaverse is a rapidly iterative interdisciplinary comprehensive concept, due to which building benchmarks for corresponding hardware and software systems is still an emerging subject. In this paper, firstly, we proposed a definition of the metaverse from the perspective of system composition: a metaverse system is composed of four subsystems, which are the access system, avatar, environment, and activity. Based on this definition, we investigated and analyzed nine specific related technological domains and explored the requirements and challenges of each component and the corresponding technological domains. Finally, combining system composition, technological domains, and requirements, we proposed our metaverse benchmark methodology. Furthermore, based on this methodology, we built a preliminary metaverse benchmark. We conducted experiments and evaluations on several relevant workloads, including face recognition, 3D reconstruction, big data sorting, and media streaming.

Regarding benchmark construction, this paper focuses on the workload abstraction of multiple individual real-world tasks. The challenges brought by the subsystems composed of multi-tasks and the entire system consisting of various subsystems greatly exceed that of a single task. We plan to build the respective metaverse subsystem based on different workload abstractions. Then, we will complete the complete metaverse system. Finally, we will construct the metaverse system-oriented metaverse benchmark suite.

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] N. Stephenson, Snow Crash, Bantam Books, United States, 1992.
[2] FnF, E-learning market size, share global analysis report, 2022 – 2030, 2023, https://www.fnfresearch.com/e-learning-market. (Accessed 5 June 2023).
[3] R. Koduri, Powering the metaverse, 2021, https://www.intel.com/content/www/us/en/newsroom/opinion/powering-metaverse.html. (Accessed 20 April 2023).
[4] P. Mattson, C. Cheng, G. Diamos, C. Coleman, P. Micikevicius, D. Patterson, H. Tang, G.Y. Wei, P. Bailis, V. Bittorf, et al., Mlperf training benchmark, in: Proceedings of Machine Learning and Systems, Vol. 2, 2020, pp. 336–349.
[5] W. Gao, F. Tang, L. Wang, J. Zhan, C. Lan, C. Luo, Y. Huang, C. Zheng, J. Dai, Z. Cao, et al., AIBench: an industry standard internet service AI benchmark suite, 2019, arXiv preprint arXiv:1908.08998.
[6] L. Wang, J. Zhan, C. Luo, Y. Zhu, Q. Yang, Y. He, W. Gao, Z. Jia, Y. Shi, S. Zhang, et al., Bigdatabench: A big data benchmark suite from internet services, in: 2014 IEEE 20th International Symposium on High Performance Computer Architecture, HPCA, IEEE, 2014, pp. 488–499.
[7] J. Zhan, A BenchCouncil view on benchmarking emerging and future computing, BenchCouncil Trans. Benchmarks Stand. Eval. (2022) 100064.
[8] W. Gao, F. Tang, J. Zhan, X. Wen, L. Wang, Z. Cao, C. Lan, C. Luo, X. Liu, Z. Jiang, Aibench scenario: Scenario-distilling ai benchmarking, in: 2021 30th International Conference on Parallel Architectures and Compilation Techniques, PACT, IEEE, 2021, pp. 142–158.
[9] J. Smart, A metaverse roadmap: Pathways to the 3D web, 2007, https://www.academia.edu/266307/A_Metaverse_Roadmap_Pathways_to_the_3D_Web_2007. (Accessed 3 March 2023).
[10] J.D.N. Dionisio, W.G. Burns III, R. Gilbert, 3D virtual worlds and the metaverse: Current status and future possibilities, ACM Comput. Surv. 45 (3) (2013) 1–38.
[11] L.H. Lee, T. Braud, P. Zhou, L. Wang, D. Xu, Z. Lin, A. Kumar, C. Bermejo, P. Hui, All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda, 2021, arXiv preprint arXiv:2110.05352.
[12] H. Ning, H. Wang, Y. Lin, W. Wang, S. Dhelim, F. Farha, J. Ding, M. Daneshmand, A survey on metaverse: the state-of-the-art, technologies, applications, and challenges, 2021, arXiv preprint arXiv:2111.09673.
[13] S.M. Park, Y.G. Kim, A metaverse: taxonomy, components, applications, and open challenges, IEEE Access 10 (2022) 4209–4251.
[14] Q. Yang, Y. Zhao, H. Huang, Z. Xiong, J. Kang, Z. Zheng, Fusing blockchain and AI with metaverse: A survey, IEEE Open J. Comput. Soc. 3 (2022) 122–136.
[15] Y. Wang, Z. Su, N. Zhang, R. Xing, D. Liu, T.H. Luan, X. Shen, A survey on metaverse: Fundamentals, security, and privacy, IEEE Commun. Surv. Tutor. (2022).
[16] Y.K. Dwivedi, L. Hughes, A.M. Baabdullah, S. Ribeiro-Navarrete, M. Giannakis, M.M. Al-Debei, D. Dennehy, B. Metri, D. Buhalis, C.M. Cheung, et al., Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy, Int. J. Inf. Manage. 66 (2022) 102542.
[17] S. Spielberg, Ready Player One, Warner Bros, United States, 2018.
[18] Microsoft, Microsoft mesh, 2021, https://www.microsoft.com/en-us/mesh. (Accessed 20 April 2023).
[19] Meta, Meta horizon workrooms, 2021, https://forwork.meta.com/horizon-workrooms/. (Accessed 20 April 2023).
[20] Blender, Blender, 2023, https://www.blender.org/. (Accessed 5 June 2023).
[21] askci, 2022 Global biometrics market size and forecast analysis of segmented industry market size, 2022, https://www.askci.com/news/chanye/20220203/1618231744677.shtml. (Accessed 25 August 2023).
[22] S. Nakamoto, Bitcoin: A peer-to-peer electronic cash system, Decentralized Bus. Rev. (2008).
[23] G.B. Huang, M. Mattar, T. Berg, E. Learned-Miller, Labeled faces in the wild: A database forstudying face recognition in unconstrained environments, in: Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition, 2008.
[24] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.
[25] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, A. Zisserman, Vggface2: A dataset for recognising faces across pose and age, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, IEEE, 2018, pp. 67–74.
[26] A. Handa, V. Pătrăucean, S. Stent, R. Cipolla, Scenenet: An annotated model generator for indoor scene understanding, in: 2016 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2016, pp. 5737–5743.
[27] A. Boulch, R. Marlet, Poco: Point convolution for surface reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 6302–6314.
[28] A.X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al., Shapenet: An information-rich 3d model repository, 2015, arXiv preprint arXiv:1512.03012.
[29] parsa-epfl, CloudSuite, 2016, https://github.com/parsa-epfl/cloudsuite. (Accessed 3 March 2023).
[30] H. Kwon, K. Nair, J. Seo, J. Yik, D. Mohapatra, D. Zhan, J. Song, P. Capak, P. Zhang, P. Vajda, et al., XRBench: An extended reality (XR) machine learning benchmark suite for the metaverse, 2022, arXiv preprint arXiv:2211.08675.
[31] M. Dabbagh, M. Kakavand, M. Tahir, A. Amphawan, Performance analysis of blockchain platforms: Empirical evaluation of hyperledger fabric and ethereum, in: 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology, IICAIET, IEEE, 2020, pp. 1–6.
[32] D. Saingre, T. Ledoux, J.-M. Menaud, BCTMark: a framework for benchmarking blockchain technologies, in: 2020 IEEE/ACS 17th International Conference on Computer Systems and Applications, AICCSA, IEEE, 2020, pp. 1–8.

BenchCouncil Transactions on Benchmarks, Standards and Evaluations 3 (2023) 100139

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/

Research article

Mind meets machine: Unravelling GPT-4's cognitive psychology

Sifatkaur Dhingra [a,b], Manmeet Singh [c], Vaisakh S.B. [c], Neetiraj Malviya [d], Sukhpal Singh Gill [e,*]

[a] Department of Psychology, Nowrosjee Wadia College, Pune, India
[b] Jindal Institute of Behavioural Sciences, O.P. Jindal Global University, Delhi-NCR, India
[c] Indian Institute of Tropical Meteorology, Pune, India
[d] Defence Institute of Advanced Technology, Pune, India
[e] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK

ARTICLE INFO

Keywords:
Artificial intelligence
Large language models
ChatGPT
Cognitive psychology
GPT

ABSTRACT

Cognitive psychology delves on understanding perception, attention, memory, language, problem-solving, decision-making, and reasoning. Large Language Models (LLMs) are emerging as potent tools increasingly capable of performing human-level tasks. The recent development in the form of Generative Pre-trained Transformer 4 (GPT-4) and its demonstrated success in tasks complex to humans exam and complex problems has led to an increased confidence in the LLMs to become perfect instruments of intelligence. Although GPT-4 report has shown performance on some cognitive psychology tasks, a comprehensive assessment of GPT-4, via the existing well-established datasets is required. In this study, we focus on the evaluation of GPT-4's performance on a set of cognitive psychology datasets such as CommonsenseQA, SuperGLUE, MATH and HANS. In doing so, we understand how GPT-4 processes and integrates cognitive psychology with contextual information, providing insight into the underlying cognitive processes that enable its ability to generate the responses. We show that GPT-4 exhibits a high level of accuracy in cognitive psychology tasks relative to the prior state-of-the-art models. Our results strengthen the already available assessments and confidence on GPT-4's cognitive psychology abilities. It has significant potential to revolutionise the field of Artificial Intelligence (AI), by enabling machines to bridge the gap between human and machine reasoning.

1. Introduction

Cognitive psychology aims to decipher how humans learn new things, retain knowledge, and recall it when needed. Cognitive psychologists seek to understand how the mind works by conducting studies on people's thoughts and actions and by using other experimental methods like brain imaging and computer modelling [1]. Understanding the human mind and developing our cognitive skills to excel in a variety of areas is the ultimate objective of cognitive psychology [2]. Fig. 1 shows the different fields of cognitive psychology under different subfields such as common sense, mathematical reasoning, logical reasoning and others. Language models have come a long way since the first statistical models for modelling language were introduced [3]. With the advent of deep learning and the availability of large amounts of data [4], recent years have seen a rapid evolution of language models that have achieved human-like performance on many language tasks. Large Language Models (LLMs) are a type of Artificial Intelligence (AI) framework that have garnered significant attention in recent years due to their remarkable language processing capabilities [5–10]. These models are trained on vast amounts of text data and are able to generate coherent, human-like responses to natural language queries. One of the key features of LLMs is their ability to generate novel and creative responses to text-based prompts, which has led to their increasing use in fields such as chatbots, question answering systems, and language translation. An example of the prompts from different datasets used in this study are shown in Fig. 2. The use of self-attention has been a key factor in this success, as it allows for more efficient and accurate modelling of long-range dependencies within the input sequence, resulting in better performance compared to traditional Recurrent Neural Network (RNN)-based models [11]. LLMs have demonstrated impressive performance on a wide range of language tasks, including language modelling, machine translation, sentiment analysis, and text classification. These capabilities have led to the increased use of LLMs in various fields, including language-based customer service, virtual assistants, and creative writing.

One of the key areas measuring intelligence in humans, other species and machines is the cognitive psychology [12]. There are

* Correspondence to: School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, UK.
E-mail addresses: sifatkaurd13@gmail.com (S. Dhingra), manmeet.cat@tropmet.res.in (M. Singh), vaisakh.sb@tropmet.res.in (Vaisakh S.B.), neetirajmalviya@gmail.com (N. Malviya), s.s.gill@qmul.ac.uk (S.S. Gill).

https://doi.org/10.1016/j.tbench.2023.100139
Received 23 August 2023; Received in revised form 28 August 2023; Accepted 28 August 2023
Available online 1 September 2023

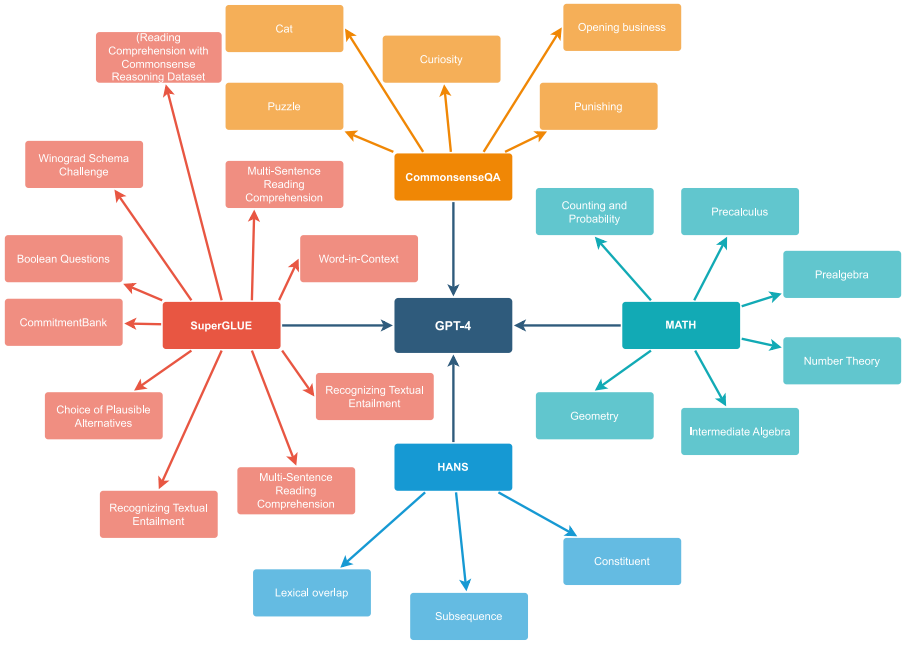

Fig. 1. Datasets used in the study with the different categories contained in them.

several tasks that are considered to be the benchmarks for testing cognitive psychology. Some of them are text interpretation, computer vision, planning and reasoning [13–15]. For cognitive psychology to work, we rely on a complex and potent social practise: the attribution and assessment of thoughts and actions [16]. The scientific psychology of cognition and behaviour, a relatively recent innovation, focuses primarily on the information-processing mechanisms and activities that characterise human cognitive and behavioural capabilities [17]. Researchers have attempted to create systems that could use natural language to reason about their surroundings [18] or that could use a world model to get a more profound comprehension of spoken language [19]. The report introducing Generative Pre-trained Transformer 4 (GPT-4) [20] has tested the HellaSwag [21] and WinoGrande [22] datasets for cognitive psychology. Although, these tests are relevant, they lack the sophistication required to understand deep heuristics of GPT-4. Hellaswag entails the task of finishing a sentence and WinoGrande involves identifying the correct noun for the pronouns in a sentence, which are quite simple. Other tasks and standardised datasets [23] which test the psychology are needed in order to perform a comprehensive assessment of cognitive psychology for GPT-4. Moreover GPT-4 needs to go through complex reasoning tasks than just predicting the last word of the sentence such as in Hellaswag, to emerge as a model capable of high-level intelligence. [24] note that SuperGLUE [25], CommonsenseQA [26], MATH [27] and HANS [28] are four such datasets that are needed to be tested for a comprehensive cognitive psychology evaluation of AI models. In this study, we evaluate the performance of GPT-4 on the SuperGLUE, CommonsenseQA, MATH and HANS datasets. This is a work in progress and we are performing continuous tests with the other datasets as suggested by [24]. Our study can be used to build up higher-order psychological tests using GPT-4.

The rest of the paper is structured as follows: Section 2 presents the datasets and methodology. Section 3 discusses the experimental results. Section 4 concludes the paper.

2. Datasets and methodology

In this study, four datasets have been used to test the cognitive psychology capabilities of GPT-4. The four datasets are CommonsenseQA, MATH, SuperGLUE and HANS. They are described as below:

2.1. CommonsenseQA

CommonsenseQA is a dataset composed for testing commonsense reasoning. There are 12,247 questions in the dataset, each with 5 possible answers. Workers using Amazon's Mechanical Turk were used to build the dataset. The goal of the dataset is to evaluate the commonsense knowledge using CONCEPTNET to generate difficult questions. The language model tested in the CommonsenseQA paper has an accuracy of 55.9% whereas the authors report that human accuracy on the dataset is around 89%.

2.2. MATH

The MATH dataset includes almost 12,500 problems from scholastic mathematics contests. Machine learning models take a mathematical problem as input and produce an answer-encoding sequence, such as $frac23$. After normalisation, their answers are distinct, therefore MATH may be evaluated using exact match instead of heuristic metrics like BLEU. Problems in seven different areas of mathematics, including geometry, are categorised by complexity from 1 to 5, and diagrams can be expressed in text using the Asymptote language. This allows for a nuanced evaluation of problem-solving skills in mathematics across a wide range of rigour and content. Problems now have comprehensive, detailed, step-by-step answers. To improve learning and make model outputs more interpretable, models can be trained on these to develop their own step-by-step solutions. The MATH dataset presents a significant challenge, with accuracy rates for big language models ranging from 3.0% to 6.9%. Models attain up to 15% accuracy on the least difficulty level and can develop step-by-step answers that are coherent and on-topic even when erroneous, suggesting that they do possess some mathematical knowledge despite their low accuracies. The results of human evaluations on MATH show that it may be difficult for humans as well; a computer science Ph.D. student who does not really like mathematics scored about 40%, while a three-time IMO gold medallist scored 90%.

2.3. SuperGLUE

SuperGLUE is an updated version of the GLUE benchmark that includes a more challenging set of language understanding tasks. Using

Dataset/Task	Sample Prompt	GPT4 response
Commonsense reasoning	She caught the first snowflake of Sher life on her tongue, she was naively excited to be spending a Swhat in a region that actually had snow? question concept: punishing choices: SA:cloud B:december C:air D:africa E:winterwhat in a region that actually had snow? question concept: punishing choices:	E:winter
Algebra	Problem: Find $a+b+c$ if the graph of the equation $y=ax^2+bx+c$ is a parabola with vertex $(5,3)$, vertical axis of symmetry, and contains the point $(2,0)$. Level: Level 5 Type: Algebra	$a+b+c = -\frac{7}{3}$.
Prealgebra	Problem: What is the greatest integer x for which $\frac79 > \frac{x}{13}$? Level: Level 3 Type: Prealgebra Solution:	the greatest integer x that satisfies the inequality is $x = 10$.
Geometry	Problem: $ABCD$, a rectangle with $AB = 12$ and $BC = 16$, is the base of pyramid P, which has a height of 24. A plane parallel to $ABCD$ is passed through P, dividing P into a frustum F and a smaller pyramid P'. Let X denote the center of the circumsphere of F, and let T denote the apex of P. If the volume of P is eight times that of P', then the value of XT can be expressed as $\frac{m}{n}$, where m and n are relatively prime positive integers. Compute the value of $m + n$.	Finally, we have $m + n = 108 + 10 = \boxed{118}$.
SuperGLUE	Sentence1 Tanks were developed by Britain and France, and were first used in combat by the British during a battle. Sentence2 Tanks were developed by Britain and France, and were first used in combat by the British during a battle with German forces. Entailment or Non-Entailment	Non entailment

Fig. 2. Examples of sample prompts and the respective responses of GPT4 on CommonsenseQA, MATH and SuperGLUE datasets.

the gap between human and machine performance as a metric, SuperGLUE improves upon the GLUE benchmark by defining a new set of difficult Natural Language Understanding (NLU) problems. About half of the tasks in the SuperGLUE benchmark have fewer than 1k instances, and all but one have fewer than 10k examples, highlighting the importance of different task formats and low-data training data problems. As compared to humans, SuperGLUE scores roughly 20 points worse when using BERT as a baseline in the original study. To get closer to human-level performance on the benchmark, the authors argue that advances in multi-task, transfer, and unsupervised/self-supervised learning approaches are essential.

2.4. HANS

The strength of neural networks lies in their ability to analyse a training set for statistical patterns and then apply those patterns to test instances that come from the same distribution. This advantage is not without its drawbacks, however, as statistical learners, such as traditional neural network designs, tend to rely on simplistic approaches that work for the vast majority of training samples rather than capturing the underlying generalisations. The loss function may not motivate the model to learn to generalise to increasingly difficult scenarios in the same way a person would if heuristics tend to produce mostly correct results. This problem has been observed in several applications of AI. Contextual heuristics mislead object-recognition neural networks in computer vision, for example; a network that can accurately identify monkeys in a normal situation may mistake a monkey carrying a guitar for a person, since guitars tend to co-occur with people but not monkeys in the training set. Visual question answering systems are prone to the same heuristics. This problem is tackled by HANS (Heuristic Analysis for Natural Language Inference (NLI) Systems), which uses heuristics to determine if a premise sentence entails (i.e., suggests the truth of) a hypothesis sentence. Neural NLI models have been demonstrated to learn shallow heuristics based on the presence of specific words, as has been the case in other fields. As not often appears in the instances of contradiction in normal NLI training sets, a model can categorise all inputs containing the word not as contradiction. HANS prioritises heuristics that are founded on elementary syntactic characteristics. Think about the entailment-focused phrase pair below:

Premise: The judge was paid by the actor.

Hypothesis: The actor paid the judge.

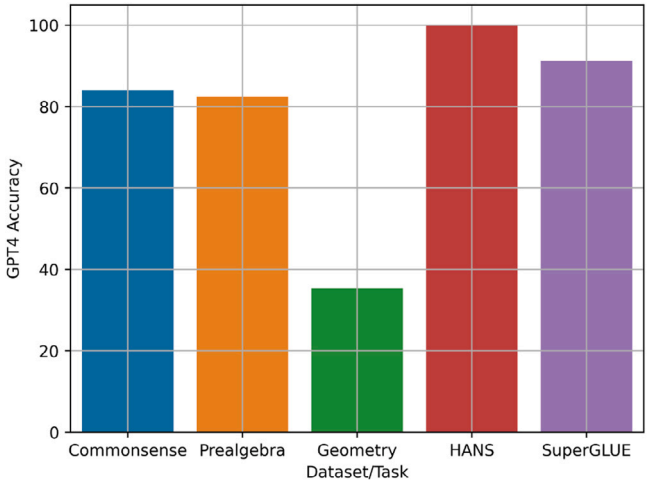

Fig. 3. Accuracy of GPT4 on cognitive psychology tasks.

An NLI system may accurately label this example not by deducing the meanings of these lines but by assuming that the premise involves any hypothesis whose terms all occur in the premise. Importantly, if the model is employing this heuristic, it will incorrectly classify the following as entailed even when it is not.

Premise: The actor was paid by the judge.

Hypothesis: The actor paid the judge.

HANS is intended to detect the presence of such faulty structural heuristics. The authors focus on the lexical overlap, subsequence, and component heuristics. These heuristics are not legitimate inference procedures despite often producing correct labels. Rather than just having reduced overall accuracy, HANS is meant to ensure that models using these heuristics fail on specific subsets of the dataset. Four well-known NLI models, including BERT, are compared and contrasted using the HANS dataset. For this dataset, all models significantly underperformed the chance distribution, with accuracy just exceeding 0% in most situations.

2.5. Methodology

We test the four datasets as described above to test the cognitive psychology capabilities of GPT-4. The model is accessed using the ChatGPT-Plus offered by OpenAI. We evaluate these models as shown in the results and discussion section on accuracy metric. Accuracy is a fundamental metric used to evaluate the performance of large language models, especially when applied to psychology datasets. It measures the proportion of predictions that the model gets right out of all the predictions it makes. In the realm of psychology, where understanding human behaviour and cognition is paramount, the accuracy of a model can be crucial. A high accuracy indicates that the model is adept at capturing the nuances of psychological data. When testing large language models on psychology datasets, accuracy can help researchers and practitioners gauge how well the model understands and processes psychological concepts, theories, and patterns. As the field of artificial intelligence evolves, striving for higher accuracy on psychology datasets ensures that models remain relevant and effective in interpreting complex human behaviours and emotions. While accuracy is vital, it is equally important to ensure that the models are tested and trained in an ethical manner, respecting the privacy and sensitivity of psychological data.

3. Experimental results

We will first discuss the human and machine skill of the different models traditionally used in the datasets used to test cognitive psychology. As compared to humans, SuperGLUE scores roughly 20 points worse when using BERT as a baseline in the original study. To get closer to human-level performance on the benchmark, the authors argue that advances in multi-task, transfer, and unsupervised/self-supervised learning approaches are essential. The language model tested in the CommonsenseQA paper has an accuracy of 55.9% whereas the authors report that human accuracy on the dataset is around 89%. The accuracy of humans on HANS dataset ranged from 76%–97% and the authors show that the BERT model performed below 10% on the non-entailment category. The human performance on MATH varied from 40%–90% and GPT-2/GPT-3 showed accuracies below 10%.

Fig. 3 shows that GPT-4 has an accuracy of 83.2% on CommonSenseQA, data, we find that GPT-4 has an accuracy of around 84%, 82% on prealgebra, 35% on geometry, 100% on HANS and 91.2% on SuperGLUE. It is to be noted that the perfect results on HANS data might be because all the examples used are of non-entailment, as the model might be memorising this particular heuristic. The experiments to generate GPT-4 results with mixed data from HANS are ongoing.

3.1. Comparison assessing the cognitive abilities of GPT-3: A state-of-the-art model

In a previous study, researchers [29] draw a parallel between the historical case of "Clever Hans", a horse believed to solve mathematical problems, and the modern interpretation of large-scale machine learning models, particularly GPT-3. The researchers employed systematic investigations and psychological experimentation to assess GPT-3's decision-making, information search, deliberation, and causal reasoning abilities. The results indicated that GPT-3 could solve certain tasks similarly or even better than humans. However, its performance was inconsistent, especially when minor changes were made to the tasks. The study found that GPT-3 performed well in certain tasks, such as gambles and a multiarmed bandit task, but lacked in areas like directed exploration and causal reasoning. The researchers emphasised the importance of understanding how these models solve tasks and suggested that future models would benefit from active interaction with the world. The study also highlighted the potential of cognitive psychology methods in understanding the behaviour of deep learning models.

4. Conclusions

GPT-4, which is a state-of-the-art large language model, is a revolution in the field of psychology since it gives psychologists unprecedented resources to use in their studies and work. This sophisticated AI model offers psychologists and psychiatrists to learn more about the human mind and come up with novel treatment theories and approaches. It provides an avenue for improved efficacy of psychological therapies and allowing professionals to spend more time with clients, leading to deeper and more fruitful therapeutic bonds. The potential applications of GPT-4 can only be realised if the model is thoroughly tested on basic tests of reasoning and cognition. Cognitive psychology enables the humans to perform various activities [30] in their personal and professional lives. We show that the performance of GPT-4 greatly surpasses the language model used in the original studies from where the different datasets are sourced, thus it can make a tool of day-to-day utility for psychologists. This development can lead to cascading benefits in addressing the mental health challenges faced by today's society.

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.

Acknowledgements

We would like to thank the Editor-in-Chief, area editor and anonymous reviewers for their valuable comments, useful suggestions to improve the quality of the paper.

References

[1] R. Núñez, M. Allen, R. Gao, C. Miller Rigoli, J. Relaford-Doyle, A. Semenuks, What happened to cognitive science? Nat. Hum. Behav. 3 (8) (2019) 782–791.
[2] L.W. Barsalou, Cognitive Psychology: an Overview for Cognitive Scientists, Psychology Press, 2014.
[3] M.C. Frank, Baby steps in evaluating the capacities of large language models, Nat. Rev. Psychol. (2023) 1–2.
[4] S.S. Gill, M. Xu, C. Ottaviani, P. Patros, R. Bahsoon, A. Shaghaghi, M. Golec, V. Stankovski, H. Wu, A. Abraham, et al., AI for next generation computing: Emerging trends and future directions, Internet Things 19 (2022) 100514.
[5] S. Harrer, Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine, eBioMedicine 90 (2023) 104512.
[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901.
[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) 9.
[9] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, et al., The malicious use of artificial intelligence: Forecasting, prevention, and mitigation, 2018, arXiv preprint arXiv:1802.07228.
[10] V. Liévin, C.E. Hother, O. Winther, Can large language models reason about medical questions? 2022, arXiv preprint arXiv:2207.08143.
[11] F. Tang, W. Gao, SNNBench: End-to-end AI-oriented spiking neural network benchmarking, BenchCouncil Trans. Benchmarks Stand. Eval. 3 (1) (2023) 100108.
[12] J. Zhao, M. Wu, L. Zhou, X. Wang, J. Jia, Cognitive psychology-based artificial intelligence review, Front. Neurosci. 16 (2022) 1024316.
[13] H. Singh, et al., Metaheuristics for scheduling of heterogeneous tasks in cloud computing environments: Analysis, performance evaluation, and future directions, Simul. Model. Pract. Theory 111 (2021) 102353.
[14] K. Bansal, et al., DeepBus: Machine learning based real time pothole detection system for smart transportation using IoT, Internet Technol. Lett. 3 (3) (2020) e156.
[15] D. Chowdhury, et al., CoviDetector: A transfer learning-based semi supervised approach to detect Covid-19 using CXR images, BenchCouncil Trans. Benchmarks Stand. Eval. 3 (2) (2023) 100119.
[16] A. Madaan, A. Yazdanbakhsh, Text and patterns: For effective chain of thought, it takes two to tango, 2022, arXiv preprint arXiv:2209.07686.
[17] S. Singh, I. Chana, M. Singh, The journey of QoS-aware autonomic cloud computing, IT Prof. 19 (2) (2017) 42–49.
[18] J. McCarthy, A basis for a mathematical theory of computation, in: Studies in Logic and the Foundations of Mathematics, Vol. 26, Elsevier, 1959, pp. 33–70.
[19] T. Winograd, Understanding natural language, Cogn. Psychol. 3 (1) (1972) 1–191.
[20] OpenAI, GPT-4 technical report, 2023.
[21] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, Y. Choi, HellaSwag: Can a machine really finish your sentence? 2019, arXiv preprint arXiv:1905.07830.
[22] K. Sakaguchi, R.L. Bras, C. Bhagavatula, Y. Choi, Winogrande: An adversarial winograd schema challenge at scale, Commun. ACM 64 (9) (2021) 99–106.
[23] Y. Li, J.L. McClelland, Systematic generalization and emergent structures in transformers trained on structured tasks, 2022, arXiv preprint arXiv:2210.00400.
[24] R. Shiffrin, M. Mitchell, Probing the psychology of AI models, Proc. Natl. Acad. Sci. 120 (10) (2023) e2300963120.
[25] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, Superglue: A stickier benchmark for general-purpose language understanding systems, Adv. Neural Inf. Process. Syst. 32 (2019).
[26] A. Talmor, J. Herzig, N. Lourie, J. Berant, Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2018, arXiv preprint arXiv:1811.00937.
[27] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt, Measuring mathematical problem solving with the math dataset, 2021, arXiv preprint arXiv:2103.03874.
[28] R.T. McCoy, E. Pavlick, T. Linzen, Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference, 2019, arXiv preprint arXiv:1902.01007.
[29] M. Binz, E. Schulz, Using cognitive psychology to understand GPT-3, Proc. Natl. Acad. Sci. 120 (6) (2023) e2218523120.
[30] G. Aher, R.I. Arriaga, A.T. Kalai, Using large language models to simulate multiple humans, 2022, arXiv preprint arXiv:2208.10264.

BenchCouncil Transactions on Benchmarks, Standards and Evaluations 3 (2023) 100137

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/

Review article

Algorithmic fairness in social context

Yunyou Huang [a], Wenjing Liu [a], Wanling Gao [b], Xiangjiang Lu [a], Xiaoshuang Liang [a], Zhengxin Yang [b], Hongxiao Li [b], Li Ma [a], Suqin Tang [a,*]

[a] Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, No. 15 Yucai Road, Qixing District, Guilin 541004, Guangxi, China

[b] Research Center for Advanced Computer Systems, Institute of Computing Technology, Chinese Academy of Sciences, No. 6 Kexueyuan South Road, Haidian District, 100190, Beijing, China

ARTICLE INFO

Keywords:
Social fairness
Bias
Discrimination
Fairness measure
Fairness datasets
Fairness algorithms

ABSTRACT

Algorithmic fairness research is currently receiving significant attention, aiming to ensure that algorithms do not discriminate between different groups or individuals with similar characteristics. However, with the popularization of algorithms in all aspects of society, algorithms have changed from mere instruments to social infrastructure. For instance, facial recognition algorithms are widely used to provide user verification services and have become an indispensable part of many social infrastructures like transportation, health care, etc. As an instrument, an algorithm needs to pay attention to the fairness of its behavior. However, as a social infrastructure, it needs to pay even more attention to its impact on social fairness. Otherwise, it may exacerbate existing inequities or create new ones. For example, if an algorithm treats all passengers equally and eliminates special seats for pregnant women in the interest of fairness, it will increase the risk of pregnant women taking public transport and indirectly damage their right to fair travel. Therefore, algorithms have the responsibility to ensure social fairness, not just within their operations. It is now time to expand the concept of algorithmic fairness beyond mere behavioral equity, assessing algorithms in a broader societal context, and examining whether they uphold and promote social fairness. This article analyzes the current status and challenges of algorithmic fairness from three key perspectives: fairness definition, fairness dataset, and fairness algorithm. Furthermore, the potential directions and strategies to promote the fairness of the algorithm are proposed.

1. Introduction

Currently, the fairness of algorithms has drawn a lot of attention in many fields, such as recidivism prediction [1], item recommendation [2], and outcome prediction [3] et al. Numerous studies have demonstrated the prevalence of unfairness in decision-making algorithms and algorithm-based systems [4–10]. In response, researchers have been actively working towards eliminating algorithmic unfairness through the development of fairness measures [11–13], the creation of fairness datasets [14–16], and the proposal of fair algorithms [17–19], among other approaches.

Research on algorithmic fairness can be categorized according to two different principles: whether to consider the long-term impact and whether to consider non-technical factors [22,23]. Table 1 shows the explanation of the terminology used in this paper. Based on the first principle, algorithmic fairness can be divided into static fairness and dynamic fairness [17,23–27]. For example, when a loan application algorithm tackles discrimination in selection rates between races, it is classified as static fairness research, but if it also takes into account the long-term effects (such as credit score change) of its decisions on the underlying population, it is categorized as dynamic fairness research [28]. According to the second principle, algorithmic fairness can be classified as technical fairness, social fairness, and sociotechnical fairness [29–36]. For example, when a loan application research optimizes mathematical rate-related fairness measures (such as equalized odds), it is classified as technical fairness research, when it pursues regulation of non-algorithmic factors (for example, making a norm to uphold the developer, user, and executor of algorithms [37]), it is classified as social fairness research, and when it addresses discrimination against different races from both technical and non-technical perspectives, it is classified as sociotechnical fairness research [22].

The central idea behind algorithmic fairness in current literature is to minimize discrimination by algorithms or systems that use algorithms, both against different groups and against individuals who are similar to each other. However, according to the definition of infrastructure — the myriad structures that underpin modern society, the algorithm has become an important social infrastructure [38]. Thus,

* Corresponding author.
E-mail address: sqtang@mailbox.gxnu.edu.cn (S. Tang).

https://doi.org/10.1016/j.tbench.2023.100137
Received 24 July 2023; Received in revised form 23 August 2023; Accepted 24 August 2023
Available online 1 September 2023

Table 1
Terminology and Explanation of Terms.

Term	Definition
Fairness [20]	"Fairness means the absence of any biases based on an individual's inherent or acquiredcharacteristics that are irrelevant within the specific decision context."
Static Fairness [21]	Without considering changes in the environment, only the current state is taken into account.Usually, "static fairness provides a one-time fair solution based on optimizing fairness constraints."
Dynamic Fairness [21]	It is an ongoing process that requires considering environmental changes,learning, and adapting to those changes to maintain fairness in decision-making.
Social Fairness [22]	Society maintains fairness by continuously striving to balance various forces.
Technical Fairness [22]	Efforts are made to utilize fairness metrics and other approaches to measure biasesin algorithms, seeking technological means to mitigate algorithmic discrimination against different subgroups or individuals.
Sociotechnical Fairness [22]	"The outcomes of a system are influenced by the interplay betweentechnical structure and social structure, as well as the interplay between instrumental values and humanistic values."
Process Fairness	Emphasizing the fairness in the process of decision-making or allocation,without being concerned about the actual outcomes.
Outcome Fairness	Focus on whether the actual outcome is fair.
Group Fairness [4]	A certain group should receive equal treatment as privileged groups or the overall population.
Individual Fairness [4]	"Similar individuals should be treated similarly."

algorithmic fairness research should not only strive to be fair but also bear the responsibility of creating a fair society, otherwise may lead to seemingly fair algorithms causing societal unfairness or creating new forms of unfairness. For example, face recognition provides audit services for all railways and aviation in China, supports the normal operation of railways and aviation, and becomes an important social infrastructure. For fairness of algorithmic behavior itself, even if the failure rate of facial recognition for individuals with facial impairments is the same as that of normal individuals, due to their heightened psychological sensitivity, they may avoid using facial recognition technology out of fear of recognition failure. This, to some extent, harms the interests of this group. For social fairness algorithmic, individuals with facial impairments should be treated specially (for example, adjust the algorithm recognition threshold) to protect them from non-technical discrimination in public due to facial defects.

In this paper, we first review and analyze existing fairness definitions (problem instantiation), fairness datasets (problem instantiation), and fairness algorithms (solution instantiation) to summarize the progress of algorithm fairness [39,40]. Then extend fairness from the algorithm level to the social level across the entire life cycle of the algorithm. As shown in Fig. 1, for the problem definition, fairness in a social environment demands not only fair behavior from algorithms and systems as societal infrastructure but also their contribution to promoting social fairness. Hence, we emphasize that the assessment of fairness extends beyond the behavior of algorithms or algorithm-based systems and includes fairness within society. For the problem instantiation, in addition to subjects, the instantiation of fairness problems should also involve algorithm developers, users, and executors. Further, in order to reflect the real fairness of society, the instances of the problem must also maintain the consistency of characteristics with the real society. For the solution instantiation, the design of the algorithm not only needs to consider optimizing the fairness metric but also needs to be able to detect the degree of social fairness and then adjust the behavior of the algorithm. Additionally, we highlight the need to restructure the algorithmic fairness benchmark in light of the new algorithmic fairness techniques above to advance algorithmic fairness research.

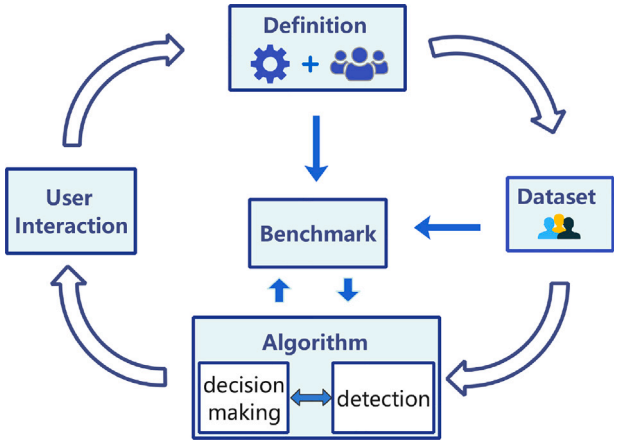

Fig. 1. The extension of algorithm fairness.

The paper is structured as follows. Section 2 reviews the definition of algorithmic fairness. Section 3 reviews data used in recent algorithmic fairness research. Section 4 reviews the fair algorithms. Section 5 extends the algorithmic fairness in the social context. Section 6 draws a concluding remark.

2. Fairness definitions: Problem definition

The concept of fairness has been widely debated in moral and political discussions, but it lacks a consistent definition [41]. With the increasing integration of AI in various domains, ethical and moral concerns have arisen, leading scientists to explore ways of incorporating fairness into algorithmic systems [42]. Currently, fairness in algorithms is defined as the overall performance of an algorithm or algorithm-based system in treating individuals or groups, as assessed by

fairness metrics. In the subsequent sections, we discuss the definition and metrics related to algorithmic fairness.

2.1. Fairness definition

The presence of diverse preferences and perspectives across different cultural backgrounds makes it challenging to establish a universal definition that applies to all individuals [43]. Broadly, fairness means that there are no biases towards an individual's inherent or acquired characteristics during the decision-making process [44]. Fairness can be divided into two categories: process fairness and outcome fairness [45], depending on whether the focus is on the fairness of the decision-making process itself or its resulting outcomes [46]. Ensuring fairness throughout the entire decision-making process is particularly challenging due to factors such as the black-box nature of deep learning algorithms. So current research primarily concentrates on outcome fairness. In terms of outcome fairness, it can be further divided into group fairness and individual fairness based on the goals of algorithmic fairness [45,46]. Distributive individual fairness holds that outcomes should be fair at the individual level, while group fairness holds that outcomes should be fair across groups [47]. Although fairness can be classified based on objectives, different researchers have different views on what the outcome of fairness should be, which we call fairness concepts [45]. The most influential concepts include Consistent Fairness and Calibrated Fairness. Consistent Fairness holds that similar individuals or diverse groups with similarities should obtain similar outcomes, while Calibrated Fairness requires that an individual's (or group's) outcome value should be proportional to their merit [45]. The above definitions are all based on the behavior of the algorithm itself without considering the existence of social unfairness. In Section 5.1, we will expand the definition of fairness: social fairness not only requires maintaining fairness in algorithmic behavior but also considers eliminating social unfairness and promoting social equity.

2.2. Fairness metrics

The definition of algorithmic fairness is intertwined with its measurement metrics. The fairness is determined by the values of fairness metrics, and the design of fairness metrics relies on the definition of fairness. In Table 2, we have provided a list of commonly used fairness metrics, categorized into individual fairness and group fairness. Individual fairness lacks a simple and executable definition, making it often difficult to achieve a consensus. On the other hand, group fairness, due to its simplicity and quantifiability, is widely utilized in fairness research. All these metrics can be useful in bias mitigation tasks when dealing with protected attributes, where A represents the protected attribute. The true label is denoted as Y, and the predicted label is denoted as $\hat{Y}$, where 0 represents negative outcomes and 1 represents positive outcomes [4]. Probabilities are represented as P.

However, it is important to note that discussions surrounding algorithmic fairness extend beyond technical metrics. The social objectives of deploying a model, the group of individuals affected by the model's decisions, and the available decision space for decision-makers to interact with the model's predictions must also be considered [48]. Different stakeholders have varying objectives, and the selection of fairness metrics must consider various application scenarios.

In summary, while fairness metrics can serve as useful tools for mitigating task biases, it is crucial to adopt a holistic approach to algorithmic fairness by considering the social context and the needs of all relevant stakeholders.

3. Fairness datasets: Problem instantiation

A dataset, which is an instance of a problem or task, is a fundamental component for the development of data-driven machine learning algorithms as it reflects the essential characteristics of a problem or task. In recent years, many datasets have been utilized for algorithmic fairness research [43,55]. In addition, due to the long-term impact of fairness, several simulators have also been developed to address the limitations of static datasets in fairness research [23]. In the following subsections, we discuss the efforts related to datasets and simulators in the context of algorithmic fairness.

3.1. Dataset

In Table 3, we have compiled a list of 10 fairness datasets and described their protective attributes and other characteristics. However, some of the datasets used in equity research may exhibit biases against protected attributes such as gender, race, and age. These biases can have adverse effects on vulnerable groups. For example, the UCI Adult dataset includes three protected attributes — gender, age, and race. However, an analysis of the dataset shows that high-income men outnumber women in almost all relationship statuses, and there is also some racial bias present [43]. Although this dataset is commonly used for categorical tasks such as predicting income levels, the $50k threshold is set inappropriately, leading to biases against Blacks and women [43].

In practice, addressing these biases may involve expanding the dataset, changing the labels of some data points, or weighting the protected attributes. For example, Retiring Adult, a reconstructed UCI Adult dataset, has been created using real data from the American Community Survey (ACS) Public Use Microdata Sample (PUMS) [15]. This dataset encompasses income as a continuous variable, enabling a more realistic prediction of whether an individual earns over $50k annually. Moreover, the dataset includes forecasting tasks for various applications such as income, employment, health, transportation, and housing.

Furthermore, Ilias et al. [16] proposed a benchmark set of legal texts covering multiple regions and languages. They adopted a competency-centered equity approach with the goal of ensuring that each group had sufficient resources to achieve similar performance levels. Ultimately, this approach is centered on the important factors of how individuals are treated in the legal process, making it an equitable approach.

Although researchers have made many improvements to fairness datasets, fairness datasets are typically static and difficult to support the development of dynamic fairness algorithms. Moreover, currently collected fairness datasets do not consider how to maintain social fairness characteristics, making it challenging to support the development of algorithms that are oriented towards social equity.

3.2. Simulator

In the pursuit of achieving dataset fairness, simulators have been proposed as a valuable tool, particularly for long-term and dynamic scenarios. D'Amour et al. [23] have developed an open-source software framework called ml-evenness-gym, an extension of OpenAI's Gym, to examine the long-term impact of the existing fair decision-making system. The framework employs a Markov decision process (MDP) where an agent chooses an action at each step to influence the state of the environment. The environment presents an observation to the agent, which is then used to determine the next action. This iterative process continues until the environment reaches an end state. Similarly, Xueru et al. [56] adopted a partially observed Markov decision process (POMDP) framework to model sequential decisions in different situations. They consider a discrete-time sequential decision process applicable to a particular population, where the effects of decisions made in each time step are reflected in the population characteristics

Table 2
Common fairness metrics.

Type	Name	Mathematical expression	Meaning
Group Fairness	Statistical Parity [44]	$P(\hat{Y}\|A=0) = P(\hat{Y}\|A=1)$	The unprotected groupand the protected group have an equal proportion of favorable outcomes.
	Equalized Odds [13]	$P(\hat{Y}\|A=0,Y=y) = P(\hat{Y}\|A=1,Y=y)$	Individuals from different group should have an equal chance of being correctly classified as positive (true positive) and incorrectly classified as positive(false positive).
	Equal Opportunity [13]	$P(\hat{Y}=1\|A=0,Y=1) = P(\hat{Y}=1\|A=1,Y=1)$	The true positive rateis equal across protected and unprotected groups.
	Treatment Equality [49,50]	$\frac{P(\hat{Y}=1\|A=1,Y=0)}{P(\hat{Y}=0\|A=1,Y=1)} = \frac{P(\hat{Y}=1\|A=0,Y=0)}{P(\hat{Y}=0\|A=0,Y=1)}$	The ratio of false positiverate to false negative rate is the same between different populations.
	Test Equality [51]	$P(Y=1\|A=0,\hat{Y}) = P(Y=1\|A=1,\hat{Y})$	The probability ofindividuals in both the protected and unprotected groups belonging to the positive class is equal.
Individual Fairness	Fairness ThroughU-nawareness [44,52,53]	\	An algorithm is considered fairas long as it does not explicitly use any protected attribute A in the decision-making process.
	Fairness ThroughAwareness [44,54]	\	For a given task-specificsimilarity measure(inverse distance), any two similar individuals should receivesimilar outcomes.
	Counterfactual Fairness [44,53]	$P(\hat{Y}_{A\leftarrow a}(U)=y\|X=x,A=a) = P(\hat{Y}_{A\leftarrow a'}(U)=y\|X=x,A=a)$	"If a decision remains consistenttowards an individual in both the actual world and a counterfactual world where the individual belongs to a different demographic group, then that decision is considered fair towards the individual."

Table 3
Overview of real-world datasets for fairness.

Dataset name		#Instances (cleaned)	Class	Domain	Protected attributes	Collection location
Law School [58]		20, 798	Binary classification	Education	Male, race	USA
UCI adult dataset [59]		45, 222		Finance	Sex, race, age	
Diabetes [60]		45, 715		Healthcare	Gender	
Dutch Census [61]		60, 420		Social	Sex	The Netherlands
Diversity in faces dataset [62]		1, 000, 000	Face recognition	Facial images	–	–
Credit Card Clients [63]		30, 000	Binary classification	Finance	Sex, marriage, education	Taiwan
Bank Marketing [64]		45, 211			Age, marital	Portugal
COMPAS Recid [65]		6, 172		Criminology	Race, sex	USA
COMPAS Viol Recid [65]		4, 020				
Retiring Adult: 2018 PUMS [15]	ACSIncome	1, 599, 229		Finance	Sex, race, age	
	ACSPublicCoverage	1, 127, 446		Healthcare		
	ACSMobility	620, 937		Housing		
	ACSEmployment	2, 320, 013		Employment		
	ACSTravelTime	1, 428, 642		Transportation		

in subsequent time steps. This work successfully addresses the limitations of using limited long-term dynamic datasets. In addition, some simulation methods utilize data augmentation techniques to address discrepancies in data sets. For instance, Iosifidis et al. [57] have used oversampling and SMOTE to generate pseudo-instances in minority communities.

Simulators have been widely used in the research of dynamic fairness algorithms to compensate for the limitations of fairness datasets, which cannot adequately support dynamic development. However, current simulators cannot provide personalized simulations for participants, leading to lower accuracy. Additionally, existing simulators do not consider the interaction between the simulated algorithm system and the participants, making it difficult to support research on algorithms focused on social fairness.

4. Fairness algorithms: Solution instantiation

In recent years, a multitude of algorithms have emerged with the aim of reducing bias and discrimination in the behavior of algorithms and systems. These innovative approaches have been introduced to address this pressing issue and ensure fairer outcomes [45,82]. Table 4 summarizes the pre-process, in-process, and post-process mechanisms for algorithmic fairness. Pre-process mechanisms, while applicable to any classification algorithm, may compromise interpretability. In-process mechanisms effectively address accuracy and fairness in the objective function, but are closely tied to the algorithm. Conversely, post-process mechanisms can be used with any classification algorithm but often yield inferior results due to their delayed application.

4.1. Pre-process

Pre-processing mechanisms play a crucial role in preparing data for machine learning algorithms. Their purpose is to minimize or eradicate bias and unfairness within the data. These methods are employed prior to feeding the data into the algorithms, ensuring that the subsequent analysis and modeling are based on a more equitable and unbiased foundation. Typically, these methods encompass techniques that focus on manipulating the distribution of protected variables within the sample or applying specific transformations to the data. The goal is to ensure that the input data remains impartial and unbiased, thereby

Table 4
Methods for fairness.

Paper	Stage	Scheme	Datasets	Evaluation Measure
[66]	Pre-process	Transformation	Adult	Discrimination=0.11, AUC=0.78
[67]	Pre-process	Reweighing	Adult	p%-rule=100%, Accuracy=82%
[68]	Pre-process	Causal Methods	NYCSF	FACE=0.273
[69]	Pre-process	Adversarial learning	Adult	EMD=0.001, Avg.Score=0.239
[70]	Pre-process	Adversarial learning	Adult	Risk Difference=0.0411, Balanced Error Rate (BER)=0.3862
[71]	In-process	Regularization	Adult, Crime and Communities, COMPAS, Default,Law School, Sentencing	–
[72]	In-process	Regularization	COMPAS	Benefits=0.97, Accuracy=68%
[73]	In-process	Adversarial learning	Adult	–
[74]	In-process	Adversarial learning	Adult	FPR=0.0248, FNR=0.4492
[75]	In-process	constraint optimization	Bank Marketing	Accuracy=87%, p%-rule=45%
[76]	In-process	constraint optimization	Use data from 3 real conferences	Paper Score(PS)=1.65, The assigned papers per reviewer(RA)=0.63
[13]	Post-process	Threshold	FICO	Profit=99.3%
[77]	Post-process	Threshold	COMPAS	–
[78]	Post-process	Transformation	Adult	PSE<3.7, Accuracy=73.8%
[79]	Post-process	Transformation	COMPAS	NDE=(0.95,1.05), Accuracy=67.8%
[80]	Post-process	Calibration	–	–
[81]	Post-process	Calibration	Racial Faces in the Wild	Accuracy=90.58%

enabling machine learning algorithms to generate fair and equitable decisions [83].

Du et al. [66] introduced a convex optimization approach to learn data transformations that aim to control group discrimination, limit distortion in individual data samples, and preserve utility. Krasanakis et al. [67] proposed a novel approach called CULEP for mitigating bias in binary classifiers. It uses an iterative reweighting process to recognize sources of bias and diminish their impact without affecting features or labels. The approach encapsulates both fairness- and classifier-related information and allows for a more precise stochastic analysis. Khademi et al. [68] made significant contributions by introducing two novel definitions of group causal relations from a causal perspective. These definitions were developed using causal methods and were designed to effectively quantify group fairness.

Recently, adversarial learning techniques have been utilized by researchers to generate fair samples. Feng et al. [69] proposed a framework for learning a latent representation of attributes through adversarial learning, preprocessing the data, and preserving useful information while preventing useless information as much as possible. Wu et al. [70] introduced a unified framework called FairGAN for generating data that meets various fairness requirements while having good utility.

4.2. In-process

The main concept of the in-process approach is to incorporate fairness considerations into the model optimization process during machine learning training, with the aim of addressing issues of unfairness resulting from dataset bias [83]. This method involves integrating fairness metrics into the model's objective function, allowing for the simultaneous optimization of performance and fairness during training.

Regularization is a common technique used in the in-process stage of fairness mechanisms. Berk et al. [71] introduced a flexible regularizer incorporating individual and group penalty mechanisms into the framework. Heidari et al. [72] addressed the fairness problem in welfare measures by enhancing the penalty for the fair benefit function. Adversarial learning methods are also rapidly developing. Celis et al. [73] employed an adversarial learning paradigm to design fair classifiers by introducing fairness objectives to enhance model performance. Similarly, Zhang et al. [74] utilized adversarial learning to mitigate bias, which is flexible and applicable to various definitions of fairness.

Constrained optimization is also a popular method. Zafar et al. [75] designed a classifier to maximize accuracy while adhering to fairness constraints to ensure that algorithmic decisions do not have unfair effects on certain sensitive attribute groups. Kobren et al. [76] proposed a novel formulation for the paper matching problem. The proposed algorithm, FAIRIR, simultaneously optimizes the global objective, obeys local fairness constraints, and satisfies lower and upper bounds on reviewer loads to ensure more balanced allocation.

4.3. Post-process

The post-processing mechanisms discussed in this passage are applied to machine learning models during the prediction and evaluation stages. These mechanisms aim to adjust the model's output to enhance fairness.

Threshold adjustment and transformation are commonly used methods for improving fairness in machine learning models. Hardt et al. [13] and Corbett et al. [77] have implemented different decision thresholds for various groups to enhance equal opportunities. Chiappa et al. [78] proposed a path-specific approach to address fairness issues. The approach corrects individual decisions by removing unfair information caused by sensitive attributes while preserving the remaining fair information along a specific path. The method is demonstrated using linear models and graphical causal models. To address the problem of fair statistical inference based on results, Nabi et al. [79] formulated the existence of discrimination as the presence of specific path effects (PSE). This refers to a path of effect determined by mediation analysis and can assist in understanding the mechanisms and reasons for discrimination.

Compared to threshold adjustment and transformation, calibration of prediction results is a method that can adjust the bias of predictions to make them closer to the true values. Hebert et al. [80] proposed a multi-calibration approach that considers the predictions of multiple calibration predictors to reduce bias and ensure fairness and accuracy. Salvador et al. [81] proposed a fair calibration method due to the recognition bias of facial recognition technology towards minority groups. This method improves the model's accuracy and generates fair calibration probabilities, thereby reducing the unfair treatment of minority groups.

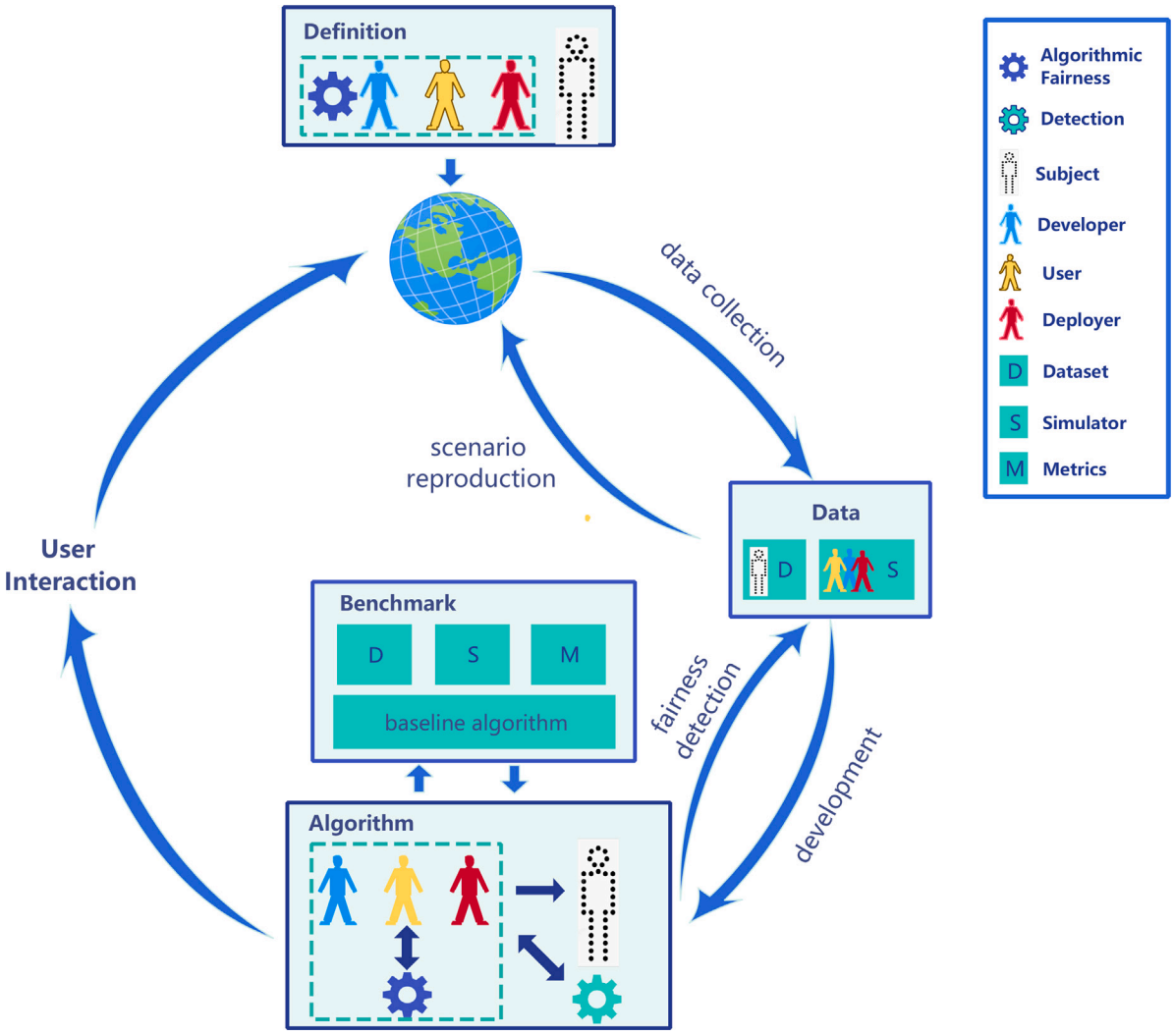

Fig. 2. The lifecycle of fairness in social content. We have expanded three stages of the development of fairness algorithms. Firstly, in the definition stage of fairness algorithms, besides including the definition of algorithmic behavior itself, it is essential to consider relevant social factors during the development process, such as developers, users, and executors. Additionally, the dynamic changes of subjects should also be taken into account. Secondly, in the instantiation stage of fairness problems, besides collecting data related to the fairness definition of the algorithm, it is necessary to model and develop simulators for each individual to simulate the dynamic interaction between the algorithm and the subjects. At this stage, the characteristics of the dataset should align with those of the real world, enabling the dataset to simulate the real world. Lastly, in the development stage of fairness algorithms, we emphasize that fairness algorithms should be capable of detecting and perceiving the level of social fairness and be able to dynamically adjust their behavior accordingly. Furthermore, to promote the development of algorithmic fairness, we need to reconstruct the current fairness benchmarks based on the aforementioned changes.

5. Fairness in social context

As social infrastructure, algorithms are not only responsible for their own behavioral fairness but also for alleviating unfairness and maintaining fairness in society. Fig. 2 presents the fairness lifecycle in the social context and identifies where algorithmic fairness can contribute. In the following subsections, current gaps in algorithmic fairness across different stages of the algorithm lifecycle are identified, and recommendations are provided for ensuring fairness in the social context.

5.1. Fairness definitions: Problem definition

The current approach to algorithmic fairness only considers the behavior of algorithms or algorithm-based systems towards people. However, fairness is a social attribute, and the impact of algorithm or system behavior on social fairness should be considered. Otherwise, a fair algorithm may risk undermining social fairness. To illustrate this, we can take the example of the MRI-PET-based diagnostic model proposed by Janghel et al. [84].

Under the current definition of algorithmic fairness, the algorithm's fairness is defined by the algorithm's prediction accuracy, sensitivity, and specificity in different groups based on gender, age, ethnicity, and disease. However, this definition has a significant flaw: the developer and user of the diagnostic model require all patients to undergo costly MRI and PET examinations, which is unfair to most patients, especially those who are healthy.

Current research has expanded the definition of algorithmic fairness to address this shortcoming by considering social perspectives. For example, Huang et al. [85] proposed a model that can tailor diagnostic strategies to patient-specific conditions. Algorithmic fairness can be defined as the prediction accuracy, sensitivity, and specificity of the system, composed of developers, users, executors, and algorithms, in different groups based on gender, age, ethnicity, income, and disease, ultimately promoting social fairness. While this approach considers humans in the loop, it also focuses only on the fairness of the algorithm-based system behavior itself.

A fair algorithm that does not take into account social inequalities can perpetuate or exacerbate social inequities. For instance, due to

Table 5
Differences between in the social context and current research in Fairness Definitions, Fairness Datasets, Fairness Algorithms.

	In The Social Context	Current Researchs
Fairness Defintions	The behavior exhibited jointly bythe algorithm and its associated social factors is equitable, and it contributes to the enhancement of societal fairness.	The behavior of the algorithm itself is fair.
Fairness Datasets	In addition to the data concerningthe target objects,the dataset also incorporates information involving developers,users, deployers,and their interactions with the system, enabling the data to recreate real-world scenarios.	Collecting data of the algorithm'sor system's target objects.
Fairness Algorithms	With the involvement of developers, users,deployers, and other individuals, this falls under the "human in the loop" mode. Moreover, it allows for the dynamic adjustment and evaluation of the algorithm's fairness.	Focusing solely on the fairnessof the algorithm itself.

the uneven distribution of medical resources, the diagnosis of irreversible Alzheimer's disease presents considerable inequity. In low-income areas, there is even a lack of specialized outpatient clinics for Alzheimer's disease, leading to a significant number of undiagnosed patients and missed early intervention. A fair Alzheimer's disease diagnosis algorithm alone will not alleviate this inequity; we need to consider improving the diagnosis accuracy of low-income groups and reducing the resource requirements of diagnosis strategies in similar situations of high-income groups.

Therefore, the definition of algorithmic fairness in specific tasks needs to focus on promoting social fairness by ensuring the prediction accuracy, sensitivity, and specificity of a system composed of algorithms, developers, users, and executors in different groups, and ultimately slowing down social discrimination in the short and long term.

5.2. Fairness datasets: Problem instantiation

Developing data-driven solutions requires using datasets to instantiate the problem and developing algorithms on the dataset. Researchers in different fields have collected datasets that can represent fairness problems in their respective fields to develop fairness algorithms. Researchers have also proposed various simulators to supplement the current static dataset to restore the dynamic and long-term nature of the fairness problem.

However, the behavior of an algorithm or algorithm-based system is not solely determined by the algorithm itself. Still, it should also include the behavior of developers, users, and executors, collectively called the "human-in-the-loop". Fairness datasets do not collect data on these social roles and cannot fully represent fairness problems in the real world. Collecting data on these social roles to create appropriate case examples for the problem will become a new direction for future research on algorithmic fairness datasets.

Furthermore, algorithms, as infrastructure, should be responsible for promoting social fairness. Unlike traditional fair datasets with loose inclusion criteria during data collection, future fair datasets must truly reflect the degree of social fairness, thereby supporting research and development to promote social equity algorithms. Maintaining fairness in real-world characteristics during data collection will become an urgent problem to be addressed. In order to ensure that datasets remain consistent with the real world, it is necessary to consider stratifying participants during data collection to select more representative individuals. Additionally, a dynamic updating mechanism for the dataset needs to be established to ensure that its characteristics continuously align with the real world. Furthermore, there is a need to develop Fairness Metrics Tool to measure the level of fairness in both the real world and the dataset. These Fairness Metrics Tool will guide the updating process of the dataset.

5.3. Fairness algorithms: Solution instantiation

While fairness algorithms have made progress in addressing algorithmic bias, they have primarily focused on the technical aspects of algorithmic fairness. However, it is important to consider the role of humans in the algorithm-based system, particularly in human-in-the-loop scenarios. The interaction between human behavior and algorithm behavior is complex and requires a more comprehensive approach. This can be achieved by modeling human roles and optimizing fairness alongside the algorithm, which has become a crucial research direction in algorithmic fairness.

In addition to being fair, algorithms should also improve social fairness. Social fairness is not static, and continued protection of vulnerable groups can inadvertently create new injustices. To achieve social fairness, algorithms must detect social fairness and dynamically adjust decision-making behaviors based on the degree of social fairness. This approach will ultimately improve social fairness while maintaining it over time. In this social context, social fairness detection and fairness dynamic game modeling will become crucial extensions of algorithmic fairness research. In order to promote social fairness, future research should focus on enhancing the current fairness algorithms by incorporating features such as dynamism, interactivity, and detectability.

5.4. Fairness benchmark

Fairness benchmarks have garnered considerable attention as a driver of algorithmic innovation. Currently, existing datasets containing sensitive information are often used as benchmark datasets for algorithmic fairness. To evaluate the long-term fairness of the algorithm, researchers have combined simulators and benchmark datasets as the evaluation benchmark of the algorithm. However, current simulators lack data on human responses to decisions of algorithms, and the accuracy of the simulation is difficult to guarantee. In the future, system–human interaction and long-term human behavior data will play an essential role in fairness benchmark research.

As shown in the Table 5, for algorithmic fairness benchmarks in the social context, the algorithmic fairness problem definition, algorithmic fairness dataset construction (problem instantiation), and fairness algorithm baseline (solution instantiation) are different from current algorithmic fairness benchmarks. Based on the above chapters on fairness definition, fairness instantiation, and fairness algorithm design, it is necessary to redesign the current fairness benchmark to promote the innovation of algorithmic fairness research in the social context. The newly introduced benchmark should be capable of concretely formulating the problem, instantiating fairness issues, adhere to the new concepts mentioned in Sections 5.1, 5.2, and 5.3, and offering a standardized and quantifiable evaluation approach.

6. Conclusion

In summary, to promote algorithmic fairness in the social context, it is important to consider the interaction between humans and algorithms and to incorporate subject-based definitions of social fairness into algorithm design. Additionally, collecting and simulating interaction data between humans and algorithms or systems, as well as addressing real-world characteristics and maintenance of social fairness are crucial. Finally, a dynamic fairness algorithm that combines subject-system interaction modeling and fairness detection, as well as benchmark refactoring in the social context, are important research directions. By addressing these challenges, we can make progress towards creating algorithms that promote social fairness and contribute to a more fair society.

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.

Acknowledgments

This work was supported by the Project of Guangxi Science and Technology, China (No. GuiKeAD20297004 to Y.H) and the National Natural Science Foundation of China (Grant No. 61967002 to S.T.).

References

[1] W. Dieterich, C. Mendoza, T. Brennan, COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity, Vol. 7, Northpointe Inc, 2016, p. 1, 7.4.
[2] Y. Wu, J. Cao, G. Xu, FASTER: A dynamic fairness-assurance strategy for session-based recommender systems, ACM Trans. Inf. Syst. (2023).
[3] H. Estiri, Z.H. Strasser, S. Rashidian, J.G. Klann, K.B. Wagholikar, T.H. McCoy Jr., S.N. Murphy, An objective framework for evaluating unrecognized bias in medical AI models predicting COVID-19 outcomes, J. Am. Med. Inf. Assoc. 29 (8) (2022) 1334–1341.
[4] Z. Chen, J.M. Zhang, F. Sarro, M. Harman, A comprehensive empirical study of bias mitigation methods for machine learning classifiers, ACM Trans. Softw. Eng. Methodol. (2023).
[5] A. Aggarwal, P. Lohia, S. Nagar, K. Dey, D. Saha, Black box fairness testing of machine learning models, in: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2019, pp. 625–635.
[6] S. Biswas, H. Rajan, Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline, in: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2021, pp. 981–993.
[7] J. Chakraborty, S. Majumder, T. Menzies, Bias in machine learning software: Why? how? what to do? in: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2021, pp. 429–440.
[8] M. Hort, J.M. Zhang, F. Sarro, M. Harman, Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods, in: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2021, pp. 994–1006.
[9] S. Udeshi, P. Arora, S. Chattopadhyay, Automated directed fairness testing, in: Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, 2018, pp. 98–108.
[10] P. Zhang, J. Wang, J. Sun, G. Dong, X. Wang, X. Wang, J.S. Dong, T. Dai, White-box fairness testing through adversarial sampling, in: Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering, 2020, pp. 949–960.
[11] A. Wang, O. Russakovsky, Directional bias amplification, in: International Conference on Machine Learning, PMLR, 2021, pp. 10882–10893.
[12] Y. Hirota, Y. Nakashima, N. Garcia, Quantifying societal bias amplification in image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13450–13459.
[13] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Advances in Neural Information Processing Systems, Vol. 29, 2016.
[14] A. Fabris, S. Messina, G. Silvello, G.A. Susto, Algorithmic fairness datasets: the story so far, Data Min. Knowl. Discov. 36 (6) (2022) 2074–2152.
[15] F. Ding, M. Hardt, J. Miller, L. Schmidt, Retiring adult: New datasets for fair machine learning, in: Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 6478–6490.
[16] I. Chalkidis, T. Pasini, S. Zhang, L. Tomada, S.F. Schwemer, A. Sø gaard, FairLex: A multilingual benchmark for evaluating fairness in legal text processing, 2022, arXiv preprint arXiv:2203.07228.
[17] Y. Hu, L. Zhang, Achieving long-term fairness in sequential decision making, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 9, 2022, pp. 9549–9557.
[18] T. Hashimoto, M. Srivastava, H. Namkoong, P. Liang, Fairness without demographics in repeated loss minimization, in: International Conference on Machine Learning, PMLR, 2018, pp. 1929–1938.
[19] Q. Zhang, J. Liu, Z. Zhang, J. Wen, B. Mao, X. Yao, Mitigating unfairness via evolutionary multi-objective ensemble learning, IEEE Trans. Evol. Comput. (2022).
[20] N.A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D.C. Parkes, Y. Liu, How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 99–106.
[21] Y. Ge, S. Liu, R. Gao, Y. Xian, Y. Li, X. Zhao, C. Pei, F. Sun, J. Ge, W. Ou, Y. Zhang, Towards Long-Term Fairness in Recommendation, Association for Computing Machinery, New York, NY, USA, 2021.
[22] M. Dolata, S. Feuerriegel, G. Schwabe, A sociotechnical view of algorithmic fairness, Inf. Syst. J. 32 (4) (2022) 754–818.
[23] A. D'Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, Y. Halpern, Fairness is not static: deeper understanding of long term fairness via simulation studies, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 525–534.
[24] V. Guardieiro, M.M. Raimundo, J. Poco, Enforcing fairness using ensemble of diverse Pareto-optimal models, Data Min. Knowl. Discov. (2023) 1–29.
[25] C. Makri, A. Karakasidis, E. Pitoura, Towards a more accurate and fair SVM-based record linkage, in: 2022 IEEE International Conference on Big Data (Big Data), IEEE, 2022, pp. 4691–4699.
[26] S. Liu, Y. Ge, S. Xu, Y. Zhang, A. Marian, Fairness-aware federated matrix factorization, in: Proceedings of the 16th ACM Conference on Recommender Systems, 2022, pp. 168–178.
[27] A. Weber, B. Metevier, Y. Brun, P.S. Thomas, B.C. da Silva, Enforcing delayed-impact fairness guarantees, 2022, arXiv preprint arXiv:2208.11744.
[28] L.T. Liu, S. Dean, E. Rolf, M. Simchowitz, M. Hardt, Delayed impact of fair machine learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 3150–3158.
[29] S. Ahmadian, A. Epasto, M. Knittel, R. Kumar, M. Mahdian, B. Moseley, P. Pham, S. Vassilvitskii, Y. Wang, Fair hierarchical clustering, Adv. Neural Inf. Process. Syst. 33 (2020) 21050–21060.
[30] J. Cho, G. Hwang, C. Suh, A fair classifier using kernel density estimation, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 15088–15099.
[31] W. Alghamdi, H. Hsu, H. Jeong, H. Wang, P.W. Michalak, S. Asoodeh, F.P. Calmon, Beyond adult and compas: Fairness in multi-class prediction, 2022, arXiv preprint arXiv:2206.07801.
[32] liobait, Indr, Measuring discrimination in algorithmic decision making, Data Min. Knowl. Discov. 31 (4) (2017) 1060–1089.
[33] S. Yao, B. Huang, Beyond parity: Fairness objectives for collaborative filtering, 2017.
[34] E.S. Jo, T. Gebru, Lessons from archives: Strategies for collecting sociocultural data in machine learning, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 306–316.
[35] C. Kuhlman, L. Jackson, R. Chunara, No computation without representation: Avoiding data and algorithm biases through diversity, 2020, arXiv preprint arXiv:2002.11836.
[36] D. Saha, C. Schumann, D. Mcelfresh, J. Dickerson, M. Mazurek, M. Tschantz, Measuring non-expert comprehension of machine learning fairness metrics, in: International Conference on Machine Learning, PMLR, 2020, pp. 8377–8387.
[37] S. Mohamed, M.-T. Png, W. Isaac, Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence, Philos. Technol. 33 (2020) 659–684.
[38] S. Thacker, D. Adshead, M. Fay, S. Hallegatte, M. Harvey, H. Meller, N. O'Regan, J. Rozenberg, G. Watkins, J.W. Hall, Infrastructure for sustainable development, Nat. Sustain. 2 (4) (2019) 324–331.
[39] J. Zhan, A BenchCouncil view on benchmarking emerging and future computing, in: BenchCouncil Transactions on Benchmarks, Standards and Evaluations, Elsevier, 2022, 100064.
[40] J. Zhan, Three laws of technology rise or fall, in: BenchCouncil Transactions on Benchmarks, Standards and Evaluations, Elsevier, 2022, 100034.
[41] B. Goldman, R. Cropanzano, "Justice" and "fairness" are not the same thing, J. Organ. Behav. 36 (2) (2015) 313–318.
[42] J. Susskind, Future Politics: Living Together in a World Transformed By Tech, Oxford University Press, 2018.
[43] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, E. Ntoutsi, A survey on datasets for fairness-aware machine learning, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 12 (3) (2022) e1452.
[44] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (6) (2021) 1–35.

[45] Y. Wang, W. Ma, M. Zhang, Y. Liu, S. Ma, A survey on the fairness of recommender systems, ACM Trans. Inf. Syst. 41 (3) (2023) 1–43.
[46] M.K. Lee, A. Jain, H.J. Cha, S. Ojha, D. Kusbit, Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation, Proc. ACM Hum.-Comput. Interact. 3 (CSCW) (2019) 1–26.
[47] Y. Ge, S. Liu, R. Gao, Y. Xian, Y. Li, X. Zhao, C. Pei, F. Sun, J. Ge, W. Ou, et al., Towards long-term fairness in recommendation, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 445–453.
[48] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, K. Lum, Algorithmic fairness: Choices, assumptions, and definitions, Annu. Rev. Stat. Appl. 8 (2021) 141–163.
[49] R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, Fairness in criminal justice risk assessments: The state of the art, Sociol. Methods Res. 50 (1) (2021) 3–44.
[50] Y. Li, H. Chen, S. Xu, Y. Ge, J. Tan, S. Liu, Y. Zhang, Fairness in recommendation: A survey, 2022, arXiv preprint arXiv:2205.13619.
[51] C. Simoiu, S. Corbett-Davies, S. Goel, The problem of infra-marginality in outcome tests for discrimination, 2017.
[52] N. Grgic-Hlaca, M.B. Zafar, K.P. Gummadi, A. Weller, The case for process fairness in learning: Feature selection for fair decision making, in: NIPS Symposium on Machine Learning and the Law, Vol. 1, No. 2, Barcelona, Spain, 2016, p. 11.
[53] M.J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.
[54] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 2012, pp. 214–226.
[55] M. Zilka, B. Butcher, A. Weller, A survey and datasheet repository of publicly available US criminal justice datasets, Adv. Neural Inf. Process. Syst. 35 (2022) 28008–28022.
[56] X. Zhang, R. Tu, Y. Liu, M. Liu, H. Kjellstrom, K. Zhang, C. Zhang, How do fair decisions fare in long-term qualification? Adv. Neural Inf. Process. Syst. 33 (2020) 18457–18469.
[57] V. Iosifidis, E. Ntoutsi, Dealing with bias via data augmentation in supervised learning scenarios, Jo Bates Paul D. Clough Robert Jäschke 24 (2018) 11.
[58] L.F. Wightman, LSAC National Longitudinal Bar Passage Study, LSAC Research Report Series, ERIC, 1998.
[59] A. Asuncion, D. Newman, UCI Machine Learning Repository, Irvine, CA, USA, 2007.
[60] B. Strack, J.P. DeShazo, C. Gennings, J.L. Olmo, S. Ventura, K.J. Cios, J.N. Clore, Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records, BioMed Res. Int. 2014 (2014).
[61] P. Van der Laan, The 2001 census in the Netherlands, in: Conference the Census of Population, 2000.
[62] M. Merler, N. Ratha, R.S. Feris, J.R. Smith, Diversity in faces, 2019, arXiv preprint arXiv:1901.10436.
[63] I.-C. Yeh, C.-h. Lien, The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Expert Syst. Appl. 36 (2) (2009) 2473–2480.
[64] S. Moro, P. Cortez, P. Rita, A data-driven approach to predict the success of bank telemarketing, Decis. Support Syst. 62 (2014) 22–31.
[65] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, in: Ethics of Data and Analytics, Auerbach Publications, 2016, pp. 254–264.
[66] F. du Pin Calmon, D. Wei, B. Vinzamuri, K.N. Ramamurthy, K.R. Varshney, Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis, IEEE J. Sel. Top. Sign. Proces. 12 (5) (2018) 1106–1119.
[67] E. Krasanakis, E. Spyromitros-Xioufis, S. Papadopoulos, Y. Kompatsiaris, Adaptive sensitive reweighting to mitigate bias in fairness-aware classification, in: Proceedings of the 2018 World Wide Web Conference, 2018, pp. 853–862.
[68] A. Khademi, S. Lee, D. Foley, V. Honavar, Fairness in algorithmic decision making: An excursion through the lens of causality, in: The World Wide Web Conference, 2019, pp. 2907–2914.
[69] R. Feng, Y. Yang, Y. Lyu, C. Tan, Y. Sun, C. Wang, Learning fair representations via an adversarial framework, 2019, arXiv preprint arXiv:1904.13341.
[70] X. Wu, D. Xu, S. Yuan, L. Zhang, Fair data generation and machine learning through generative adversarial networks, in: Generative Adversarial Learning: Architectures and Applications, Springer, 2022, pp. 31–55.
[71] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, A. Roth, A convex framework for fair regression, 2017, arXiv preprint arXiv:1706.02409.
[72] H. Heidari, C. Ferrari, K. Gummadi, A. Krause, Fairness behind a veil of ignorance: A welfare analysis for automated decision making, Adv. Neural Inf. Process. Syst. 31 (2018).
[73] L.E. Celis, V. Keswani, Improved adversarial learning for fair classification, 2019, arXiv preprint arXiv:1901.10443.
[74] B.H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 335–340.
[75] M.B. Zafar, I. Valera, M.G. Rogriguez, K.P. Gummadi, Fairness constraints: Mechanisms for fair classification, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 962–970.
[76] A. Kobren, B. Saha, A. McCallum, Paper matching with local fairness constraints, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 1247–1257.
[77] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, Algorithmic decision making and the cost of fairness, in: Proceedings of the 23rd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2017, pp. 797–806.
[78] S. Chiappa, Path-specific counterfactual fairness, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01, 2019, pp. 7801–7808.
[79] R. Nabi, I. Shpitser, Fair inference on outcomes, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, No. 1, 2018.
[80] U. Hébert-Johnson, M. Kim, O. Reingold, G. Rothblum, Multicalibration: Calibration for the (computationally-identifiable) masses, in: International Conference on Machine Learning, PMLR, 2018, pp. 1939–1948.
[81] T. Salvador, S. Cairns, V. Voleti, N. Marshall, A. Oberman, Faircal: Fairness calibration for face verification, 2021, arXiv preprint arXiv:2106.03761.
[82] D. Pessach, E. Shmueli, A review on fairness in machine learning, ACM Comput. Surv. 55 (3) (2022) 1–44.
[83] S. Caton, C. Haas, Fairness in machine learning: A survey, 2020, arXiv preprint arXiv:2010.04053.
[84] R. Janghel, Y. Rathore, Deep convolution neural network based system for early diagnosis of Alzheimer's disease, IRBM 42 (4) (2021) 258–267.
[85] Y. Huang, N. Wang, S. Tang, L. Ma, T. Hao, Z. Jiang, F. Zhang, G. Kang, X. Miao, X. Guan, et al., OpenClinicalAI: Enabling AI to diagnose diseases in real-world clinical settings, 2021, arXiv preprint arXiv:2109.04004.