Augmenting Human Intelligence: Evaluate the Performance of Conversational Agents for Effective Human-Machine Teaming

Jacopo Caratti
Seminar Human-Computer Interaction
Fribourg, Switzerland
jacopo.caratti@students.unibe.ch

Dr. Simon Ruffieux
Seminar Human-Computer Interaction
Fribourg, Switzerland
simon.ruffieux@unifr.ch

ABSTRACT

In this paper, we address the challenge of evaluating the performance of Conversational Agents (CAs) in the domain of Human Machine Teaming (HMT). We first discuss the current evaluation methods and expose the lack of a defined procedure to assess the performance of a CA. To overcome this challenge, we propose the performance metric *P*, which incorporates general and domain-specific measures to provide a complete assessment of a CA. In addition, we propose a visual representation to aid in the interpretation of the results. To validate the usefulness of *P* as a comprehensive HMT metric for assessing the performance of the machine, the human, and their interaction, we outline an experiment on two different CAs that demonstrates its assessment capabilities.

Author Keywords

Artificial Intelligence; Human-Centered AI; Human-Machine Teaming; Conversational Agent; Evaluation; Performance

INTRODUCTION

The evolution of Artificial Intelligence (AI) has been nothing short of remarkable. From on of its first inceptions in 1959 [17] to present day, AI has been lauded as one of the greatest discoveries in human history, with some likening its impact to that of fire and the wheel. Indeed, AI has changed the way we view not only technology but also the entire world, sparking a paradigm shift for society [1].

Towards more Human-Centered AI

Previously, interactions between intelligent systems and humans were viewed through two predominant lenses: the rationalistic and design perspectives [22]. The rationalistic approach, pioneered by John McCarthy in his research paper *Programs With Common Sense* [17], considered humans as "cognitive machines", whose behavior and thinking could be replicated by machines. Conversely, Douglas Engelbart's design approach acknowledged that AI is an algorithmic and statistical tool to solve problems intended for humans [5].

However, these perspectives fell short in accounting for the complexities of real-world scenarios. To address this gap, modern approaches to AI have become more human-centered, with technology engineering now designed keeping the needs and capabilities of users in mind. This new perspective introduced by Donald A. Norman in [19] considers the individual as a whole person, taking into account their unique background, desires, ambitions, needs, interests, and lifestyle within a specific cultural context [5].

Hybrid Intelligence vs. Human-Machine-Teaming

The integration of machines and humans is becoming increasingly prevalent: from finance to healthcare, and everything in between. This growth has given rise to various methods of combining machine and human intelligence, two examples being Hybrid Intelligence (HI) and HMT.

Both approaches acknowledge the unique strengths and weaknesses of humans and machines [12]. However, they differ in their approaches to integration. Machines excel at recognizing patterns, machine learning, reasoning, and optimization but lack general world knowledge, common sense, collaboration, adaptability, responsibility, and explanation. Conversely, humans bring experience, ethics, legal and social concerns, collaboration, and flexibility to the table [1].

HI seeks to leverage the unique capabilities of both human and machine intelligence, creating a new, higher level of intelligence that is greater than the sum of its parts [1], where both benefit from the interaction. Conversely, HMT focuses on machines assisting humans in completing tasks. Tasks are assigned to each agent, with their results being combined to produce a final solution beneficial for humans, in a more one-way profit's flow scenario [18]. Nevertheless, the goal remains the same: to improve outcomes by leveraging the strengths of both human and machine intelligence.

IBM Watson for Oncology is a HI system that uses advanced algorithms to provide oncologists with evidence-based treatment options and insights. It has revolutionized oncology by accessing clinical knowledge, offering unparalleled accuracy and precision to treatment planning [23]. Watson learns from

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

human experts [10] and adapts to new challenges, improving its performance. It integrates human and machine intelligence, making it an excellent example of healthcare achievements.

Centaurus Chess is a remarkable HMT system where teams composed of both humans and computers collaborate to compete against other teams. With each member responsible for distinct aspects of the game, their collaboration results in sophisticated moves. In 2005, 3 Chess grand masters lost a game against Zacks HMT, composed of 2 average players using 3 conventional laptops [13]. This synergy between human and machine intelligence produces a force that rivals even supercomputers [9].

The line between HI and HMT can be blurry, however we must remember that the goal is not to replace human intelligence but use technology to amplify it. The paper [1] highlights this point, emphasizing the importance of creating collaborative, adaptive, responsible, and explainable AI systems. By leveraging the unique strengths of both humans and machines, we can design systems that not only enhance human performance but also ensure ethical and responsible use of technology.

Need for Metrics

However, since HMT is a relatively new and wide research field, there remains a lack of metrics to fully comprehend the behavior and effectively assess the performance of HMT systems [11].

It's crucial to remember that, despite their impressive capabilities, machines are still mere machines. Just like humans, they are not infallible, and therefore, it is essential to exercise vigilance when integrating them into your team. However, the effectiveness of CAs in the domain of HMT systems is not solely dependent on the quality of the AI system. It also relies on the quality of human involvement and the interaction between humans and machines.

As CAs such as ChatGPT become more widespread and accessible, it is necessary to develop metrics and methodologies to effectively evaluate this synergy among all its components involved, to ensure a safe and optimal use of the system.

In this paper, we investigate and analyze existing metrics in order to explore how they can be adapted for the CAs' scenario in the HMT domain, allowing then to assess and compare their performances with greater precision and clarity.

RELATED WORKS

This section begins with general insights on the evaluation of HMT systems, followed by a closer look at CAs assessment, and concluding with methods for measuring the effectiveness of the advanced ChatGPT. Our order of discussion progresses from general to specific contexts, allowing to understand how general techniques are adapted to suit specific situations.

HMT Evaluation

Researchers proposed the PRODEC method [8]: an innovative cyclic approach designed to define performance metrics of HMT systems built on top of the MOHICAN project [7]. PRODEC has been specifically applied to the domain of military aviation, revealing the crucial role of trust and collaboration between pilots and machines for an optimal performance.

However, quantifying the levels of trust and collaboration in HMT systems is not a straightforward task, given their more subjective nature. To address this, the study delves into the key factors that contribute to these essential elements of an HMT system, and explores strategies for measuring them.

As a result, the researchers identified four fundamental criteria for the effective measurement of trust and collaboration in an HMT system. These criteria provide a comprehensive framework that encompasses this intricate interplay between human and machine.

- Effectiveness = tasks completion, resources consumption, risk management, eyes-tracking
- Efficiency = total interaction time
- Reliability/Robustness = bugs
- Situation awareness/Mental workload = index evaluating human workload

By focusing on these crucial metrics, PRODEC provides a valuable framework to improve the performance of HMT systems in this and - by adapting the metrics - other domains, leading to systems that can operate at their full potential.

CA Evaluation

Evaluating the performance of a CA requires specific methods that vary according to the type of CA being evaluated. The paper [3] discusses three classes of general evaluation methods that can be applied to any CA.

The most accurate evaluation method is Human-Based Evaluation, where humans interact with the CA and provide feedback on its quality, fluency, appropriateness, and sensibleness. Metrics specific to open-domain and closed-domain CA are used to evaluate user experience, coherence, engagement, domain coverage, topical depth, topical diversity, dialogue duration, information transmission, and speech action contents. While human evaluation is the most accurate, it can also be very expensive, and cheaper methods are welcome.

Widely used is Machine-Based Evaluation, which relies on metrics such as BLEU [20], ROUGE [15], and METEOR [6] to evaluate translations. However, difficulties in defining the measurement goal poses a problem. In fact, machine evaluations assume that a good CA's answer should overlap with the ground-truth, which may not always be the case: depending on the context, a human answer could be completely different than the ground truth but still be valid.

To address this issue, Machine-Learning-Based Evaluation has been introduced, with methods such as ADEM [16] and RUBER [21] trying to predict human judgments.

In the end, they say that by defining and analyzing domain-specific metrics, we can obtain more accurate and comprehensive evaluations of CAs. It can also help designers and developers identify areas of improvement and enhance the

CA's capabilities to meet the specific needs of their intended domains. Why is this important? Because CAs may be designed to perform specific tasks, and these tasks can vary across domains. For example, a CA designed to assist in healthcare has different objectives and measures of success than a CA built for customer service in retail. In healthcare, a CA's objective may be to assist patients in managing their health conditions, while in retail on enhancing customers experience. Thus, measures of success for a healthcare CA may include patient satisfaction and accuracy of health information, while for a retail CA, they may include customers satisfaction and purchases facilitation. Thus, it becomes more evident that generic metrics may not capture the nuances and intricacies of domain-specific tasks.

ChatGPT Evaluation

In this segment, we delve into two different methods for evaluating the performance of ChatGPT, offering a deeper understanding of how this CA performs in various contexts.

ChatGPT in Statistics

Researchers selected 15 statistics open questions of varying difficulty from the Math Courses/Statistics 101 platform [2] to assess the capabilities of ChatGPT. The model's responses were then compared to ground-truth answers from experts, allowing researchers to analyze its accuracy. The aim was to see how ChatGPT could handle different statistical tasks.

Interestingly, results showed that ChatGPT was very good at answering fundamental operations and process analysis questions, but struggled with more straightforward calculations: it explained the correct method to compute the mean of a 9 integers set but wrongly solved the computation.

This happens when algorithms are biased on the training data set, posing a potential threat to academic integrity. Therefore, it is essential to ensure the ethical and responsible use of AI models: they can simulate human attitudes but cannot replicate creativity or critical thinking.

However, despite these limitations, the model still managed to provide enough correct answers to achieve a sufficient grade.

ChatGPT in Ophthalmology

ChatGPT has been subjected to a rigorous testing regime, wherein it was tasked with answering a series of multiple-choice questions from the Ophthalmic Knowledge Assessment Program [4]. These questions, varying in difficulty and spanning diverse subject matter, were all equipped with ground-truth answers.

Researchers conducted two comprehensive evaluations, each consisting of 260 questions, which bore witness to ChatGPT's precision rates of 55.8% and 42.7%, respectively.

Even in this study it emerged clear that ChatGPT is more competent in general questions: it was more accurate in general medicine rather than the intricate fields of ophthalmology, ophthalmic pathology and intraocular tumors.

Undeniably, the results of this study are promising, but researchers suggest that domain-specific training may be required to further elevate the performance of this AI model.

CONCEPTION

The main focus of these metrics is evaluating the synergy between humans and CAs, rather than assessing their individual performance. While individual metrics like accuracy for machines or IQ for humans can measure their capabilities, these have already been extensively studied. Thus, by closely analyzing the techniques proposed in the previous section, this work focuses on defining key aspects specifically designed to assess human-CA performance.

At first, some metrics are domain-specific, while others are more generally applicable to any kind of HMT. These metrics serve different purposes and can provide valuable insights into different aspects of the CA's performance.

Domain-specific metrics are tailored to the specific domain in which the CA operates. These metrics are designed to evaluate the performance in completing tasks that are specific to that domain. In computer science, we may measure task **accuracy** by comparing tasks' responses to predefined ground-truth responses.

On the other hand, general HMT metrics are more broadly applicable and can be used to evaluate the system performance across different domains. An example is **efficiency**, which can be objectively measured by tracking the time it takes to complete a task. A second one is **ineffectiveness**, a subjective metric based on participants' mental workload, measured by a questionnaire like NASA-TLX [14]. It provides insight into how well the CA meets users' needs and if improvements are needed to reduce cognitive load.

Considering both domain-specific and general metrics gives researchers a nuanced and holistic understanding of the CA's performance, even allowing for performance comparison and user experience improvements.

Based on the discussion, we propose a visualization and a mathematical formula for the performance, which are based on accuracy A, efficiency E, and ineffectiveness I of the system. To ensure consistency, each metric M_i is firstly scaled to a 0-1 range based on its minimum and maximum values. This technique is also known as *min-max normalization* $\hat{M}_i$, and allows to standardize the values and better compare results across different systems.

$$\hat{M}_i = \frac{\mu_{i,M} - min(M_i)}{max(M_i) - min(M_i)}$$

These three metrics can be mapped - each to a separate axis - to the 3d space, allowing us to produce a visualization of the system performance offering an easy and comprehensive view of strengths and weaknessess of the system.

Figure 1 showcases hypothetical systems Green and Orange. Green demands a high mental workload and considerable time to generate responses. However, it generates high quality responses, making it suitable for use in applications where accuracy is paramount. Differently, Orange is faster and requires much lower mental workload from the user. However,

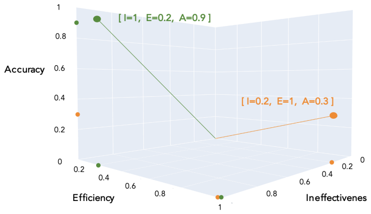

Figure 1. Visual representation of two performance vectors.

responses are highly inaccurate, compromising its usefulness in applications where accuracy is critical.

Then, for a CA i, we can combine the normalized positive mean contributions of A and E, subtracts the normalized negative mean contribution of I, and condense them into $P_i \in [-1;1]$. A higher value of P_i indicates better performance.

$$P_i = w_1\hat{A}_i + w_2\hat{E}_i - w_3\hat{I}_i$$

The weights w_1, w_2, and w_3 reflect the relative importance of each term. They can be adjusted depending on the specific goals of the study or application context but should add up to 1 to ensure a more valid and meaningful calculation.

EXPERIMENT: CA PERFORMANCES' COMPARISON

The goal is to apply the meticulously selected metrics combined in P to examine and compare ChatGPT and ChatSonic - currently two of the currently most advanced CAs - on specific tasks from the computer science domain.

Hypothesis

We expect that these metrics will be useful in constructing a credible index for evaluating HMT performance. Furthermore, we expect ChatGPT to outperform ChatSonic.

Participants

We recruit 100 computer science experts representative of the population. 50 are randomly assigned to ChatSonic, 50 to ChatGPT.

Procedure

The experimental procedure involves each participant engaging with the assigned CA to complete a set of 12 computer science domain-specific tasks. The tasks are carefully crafted to evaluate the performance of the CA on various aspects of their conversational abilities: the understanding of theoretical concepts, the accuracy in providing helpful information, and the proficiency in executing technical solutions. The tasks include 5 multiple-choice theoretical questions, 5 open-ended theoretical questions, and 2 function implementation.

To measure the **efficiency**, we record the time each participant requires to complete each task. For **accuracy** evaluation, multiple-choice tasks are provided with a ground-truth answer; open-ended tasks with a list of considerations that an acceptable answer must contain; coding-related tasks by comparing the computational result with the one of a ground-truth function. After the completion of the tasks, we ask the participants to fill out the subjective NASA-TLX to evaluate the CAs' **ineffectiveness** based on their perceived mental workload.

The results of the experiment provide insights into the strengths and weaknesses of each system. The metrics can now be visualized and used to compute $P_{chatGPT}$ and $P_{chatSonic}$.

EXPECTED RESULTS

Based on the hypothesis and the design of the experiment, we expect ChatGPT to outperform ChatSonic in terms of overall performance, as measured by the P formula.

We expect this result due to ChatGPT's advanced NLP capabilities and large training data, allowing it to generate more accurate and coherent responses than ChatSonic. We also expect ChatGPT to perform better in terms of efficiency and effectiveness. In fact, ChatSonic, despite being an advanced CA, has a much heavier user interface, which could notably cause an increase in the user's mental workload and - consequently - also in the time required to accomplish each task.

Regarding the P measure, we obtain a holistic evaluation of the CAs' performance, which considers both objective and subjective aspects, both crucial in evaluations of HMTs. We expect P to provide a more comprehensive evaluation compared to the individual metrics alone. Nevertheless, the 3d-plot still offers a useful representation of the 3 metrics individually.

Furthermore, by using min-max normalization and weighting each metric by importance, we ensure P reflects the goals and priorities of the evaluation. Additionally, by limiting the range of P to $[-1;1]$, we obtain a standardized and easily interpretable measure, which can facilitate the comparison between different CAs and experiments. Clearly, the more data collected for each CA with the same experimental settings, the more reliable the visualization and the P metric will be.

Overall, we expect that the results of the experiment will provide valuable insights into the performance of ChatGPT and ChatSonic on computer science domain-specific tasks and demonstrate the usefulness of the performance vector representation and P as a metric for evaluating the efficiency, accuracy, and ineffectiveness of CAs in the domain of HMTs.

CONCLUSION

The literature review presented here highlights a significant challenge in assessing the performance of HMTs and their corresponding CAs. There are various metrics available for evaluation but there is no clear consensus on a defined methodology. This emphasizes the need for a holistic approach that incorporates both general and - considering the specific domain of analysis - domain-specific metrics. Furthermore, the importance of human reviews regarding the interaction cannot be overstated. The subjective experiences of users play a crucial role in the performance assessment.

Our proposed performance vector and metric P offer a comprehensive solution that takes into account all relevant factors: subjective, objective, general, and domain-specific metrics, providing a complete representation of a CA's performance in the HMT domain.

REFERENCES

[1] Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. 2020. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer* 53, 8 (2020), 18–28. DOI:http://dx.doi.org/10.1109/MC.2020.2996587

[2] Abdo Al-Qadri and Salah Ahmed. 2023. Assessing the ChatGPT Accuracy Through Principles of Statistics Exam: A Performance and Implications. (03 2023). DOI: http://dx.doi.org/10.21203/rs.3.rs-2673838/v1

[3] Merav Allouch, Amos Azaria, and Rina Azoulay. 2021. Conversational Agents: Goals, Technologies, Vision and Challenges. *Sensors* 21, 24 (2021). DOI: http://dx.doi.org/10.3390/s21248448

[4] Fares Antaki, Samir Touma, Daniel Milad, Jonathan El-Khoury, and Renaud Duval. 2023. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of its Successes and Shortcomings. *medRxiv* (2023). DOI: http://dx.doi.org/10.1101/2023.01.22.23284882

[5] Jan Auernhammer. 2020. Human-centered AI: The role of Human-centered Design Research in the development of AI. (08 2020). DOI: http://dx.doi.org/10.21606/drs.2020.282

[6] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://aclanthology.org/W05-0909

[7] Guy Boy, Julien Dezemery, Benoit Haffreingue, Raymond Lu Cong Sang, and Chloe Morel. 2020. MOHICAN: human-machine performance monitoring through trust and collaboration analysis. Towards smarter design of a virtual assistant and real time optimization of machine behavior. (03 2020).

[8] Morel C. Boy GA. 2022. The machine as a partner: Human-machine teaming design using the PRODEC method. (2022). DOI: http://dx.doi.org/10.3233/WOR-220268

[9] Nicky Case. 2018. How To Become A Centaur. *Journal of Design and Science* (jan 8 2018).

[10] IBM Corporation. accessed March 24, 2023. IBM Watson for Oncology. https://www.ibm.com/common/ssi/cgi-bin/ssialias?appname=skmwww&htmlfid=897/ENUS5725-W51&infotype=DD&subtype=SM&mhsrc=ibmsearch_a&mhq=IBM. (accessed March 24, 2023).

[11] Praveen Damacharla, Ahmad Y. Javaid, Jennie J. Gallimore, and Vijay K. Devabhaktuni. 2018. Common Metrics to Benchmark Human-Machine Teams (HMT): A Review. *IEEE Access* 6 (2018), 38637–38655. DOI: http://dx.doi.org/10.1109/ACCESS.2018.2853560

[12] Dominik Dellermann, Adrian Calma, Nikolaus Lipusch, Thorsten Weber, Sascha Weigel, and Philipp Ebel. 2019. The Future of Human-AI Collaboration: A Taxonomy of Design Knowledge for Hybrid Intelligence Systems. DOI:http://dx.doi.org/10.24251/HICSS.2019.034

[13] Ian M. Goldstein, Julie Lawrence, and Adam S. Miner. 2017. Human-Machine Collaboration in Cancer and Beyond: The Centaur Care Model. *JAMA Oncology* 3, 10 (10 2017), 1303–1304. DOI: http://dx.doi.org/10.1001/jamaoncol.2016.6413

[14] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. DOI:http://dx.doi.org/https://doi.org/10.1016/S0166-4115(08)62386-9

[15] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[16] Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2018. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. (2018).

[17] John McCarthy. 1959. Programs with Common Sense.

[18] Engineering National Academies of Sciences and Medicine. 2022. *Human-AI Teaming: State-of-the-Art and Research Needs*. The National Academies Press. DOI:http://dx.doi.org/10.17226/26355

[19] Donald A. Norman. 2002. *The design of everyday things*. Basic Books.

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. DOI: http://dx.doi.org/10.3115/1073083.1073135

[21] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr 2018). DOI: http://dx.doi.org/10.1609/aaai.v32i1.11321

[22] Terry Winograd. 2006. Shifting viewpoints: Artificial intelligence and human–computer interaction. *Artif. Intell.* 170 (12 2006), 1256–1258. DOI: http://dx.doi.org/10.1016/j.artint.2006.10.011

[23] Na Zhou, Chuan-Tao Zhang, Hong-Ying Lv, Chen-Xing Hao, Tian-Jun Li, Jing-Juan Zhu, Hua Zhu, Man Jiang, Ke-Wei Liu, He-Lei Hou, Dong Liu, Ai-Qin Li, Guo-Qing Zhang, Zi-Bin Tian, and Xiao-Chun Zhang. 2018. Concordance Study Between IBM Watson for Oncology and Clinical Practice for Patients with Cancer in China. *The Oncologist* 24, 6 (09 2018), 812–819. DOI: http://dx.doi.org/10.1634/theoncologist.2018-0255