

◀◀◀◀ FINAL PRESENTATION

HUMAN-AI TEAMING FOR AN AUGMENTED INTELLIGENCE

Seminar Human Computer Interaction

Student Jacopo Caratti

Supervisor Simon Ruffieux

INTRODUCTION

Human Machine Teaming



2005 – 3 chess grandmasters lose against team Zacks of 2 average players and 3 computers



CONVERSATIONAL AGENT

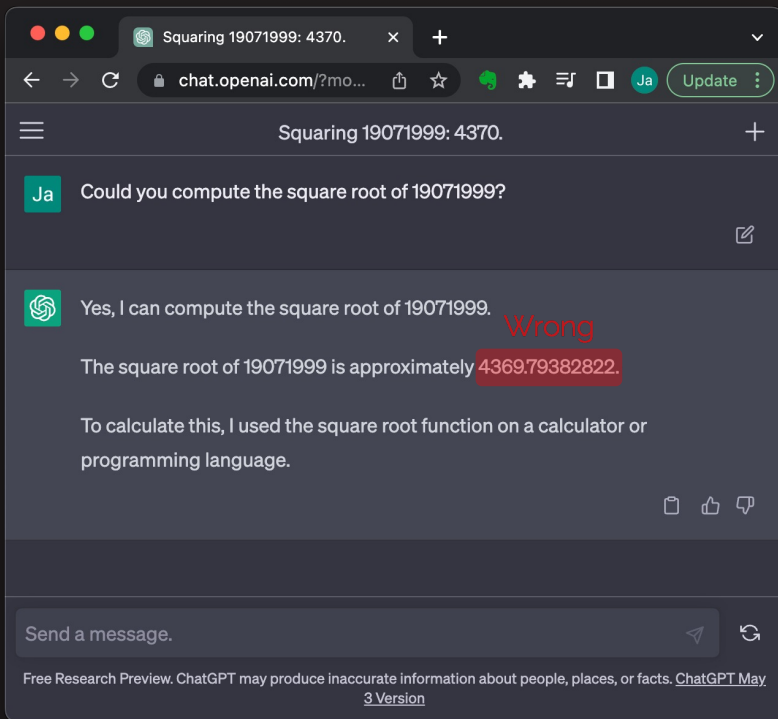
Dialogue system that conducts NLP and responds automatically using human language, deployed as chatbots or device assistants.

[DeepAI](#)



'Man talking to a robot, Van Gogh'

/[AI]

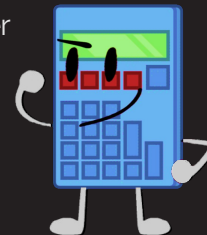


Results

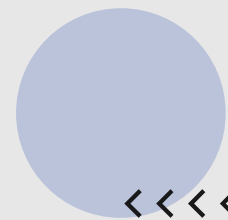
Better at answering
general domain questions

Biased on training data

Who's better
now?



REALLY GOOD...?

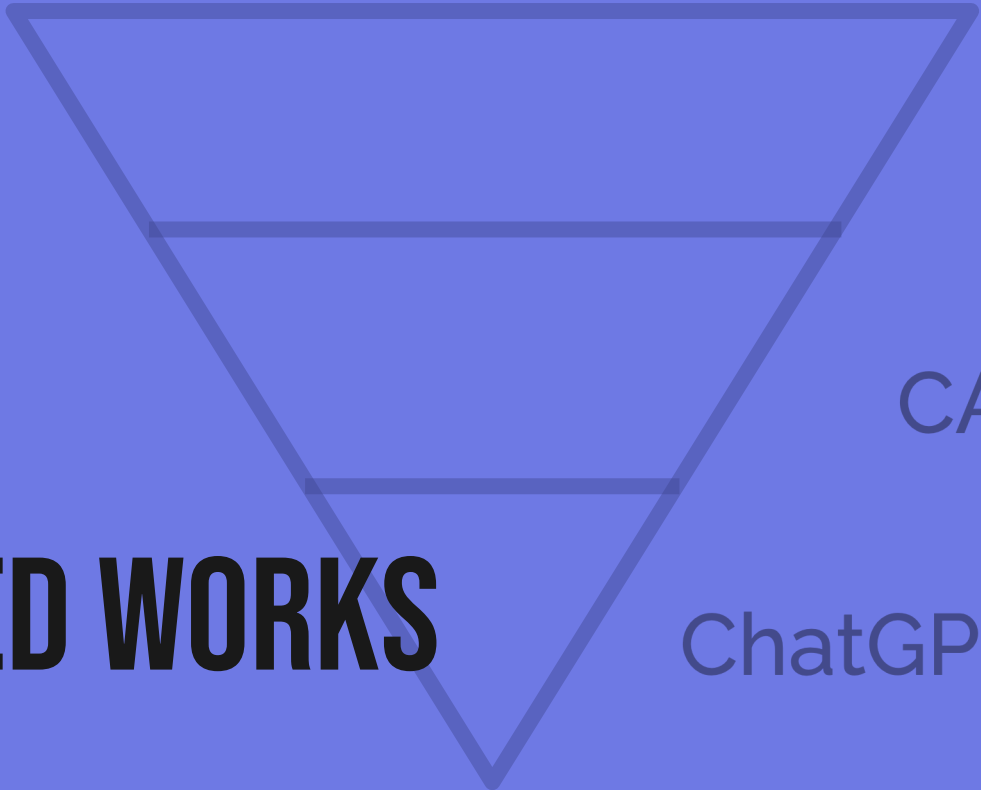


RESEARCH QUESTION

How to measure and compare the performances of two CAs?

A case study in the Computer science's domain





HMT

CA

ChatGPT

RELATED WORKS

Evaluation



HMT EVALUATION



TRUST & COLLABORATION

Important to assess HMT performance.

EFFECTIVENESS

EFFICIENCY

RELIABILITY

SITUATION AWARENESS

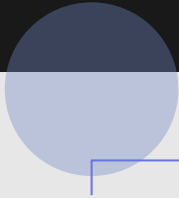
Boy GA, Morel C. The machine as a partner: Human-machine teaming design using the PRODEC method. *Work*. 2022;73(s1):S15-S30. doi: 10.3233/WOR-220268. PMID: 36214030.



AL
EN
AD

CA EVALUATION

Allouch, Merav and Azaria, Amos and Azoulay, Rina, "Conversational Agents: Goals, Technologies, Vision and Challenges", Sensors, vol. 21, no. 24, 2021, doi: 10.3390/s21248448



Domain-specific metrics

| Domain | CA | Description | Evaluation |
|----------------|------------|-------------------------------------|---|
| Open-Domain | KBot | Knowledge chatbot | F-score, precision, recall, intent classification |
| Social Support | ELIZA | The first CA: emulate psychology | People experience |
| Educational | Sara | Student's assistant | Pre/post test scores of learners, pre/post survey |
| Healthcare | CoachAI | Patient's support chatbot | User engagement, system acceptance and rating |
| Goal-Oriented | SUGILITE | Programming-by-demonstration system | Task completion time |
| Commercial | SuperAgent | Customer-service chatbot | 2 customer reviews |



AL
EN
AD



STATISTICS

5 open questions with ground-truth

Abdo Al-Qadri and Salah Ahmed, "Assessing the ChatGPT Accuracy Through Principles of Statistics Exam: A Performance and Implications", March 2023, doi: <http://dx.doi.org/10.21203/rs.3.rs-2673838/v1>



OPHTHALMOLOGY

2 rounds of 260 multiple-choice questions
varying in difficulty with ground-truth

Fares Antaki, Samir Touma, Daniel Milad, Jonathan El-Khoury, and Renaud Duval, "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>

CHATGPT EVALUATION

CONCEPTION

Metric Definition

/[AI]

Ineffectiveness

I

Efficiency

E

Accuracy

A

PERFORMANCE VECTOR

Domain-specific and
general metrics

Accuracy

Ineffectiveness

$$P = w_1A + w_2E - w_3I$$

Efficiency

$$w_1 + w_2 + w_3 = 1$$

Vectorial representation



Performance vector

REPRESENTATION

EXPERIMENT

ChatGPT vs ChatSonic



Group 1
CHATGPT



Group 2
CHATSONIC

50 computer scientists per group

/[AI]/[AI]/

ARTI

IGEN

[AI]



Correctness

ACCURACY

Ineffectiveness

I

Efficiency

E

Accuracy

A

MULTIPLE CHOICE

5

What is the time complexity of a binary search algorithm?

$O(1)$ $O(n)$ $O(\log n)$ $O(n^2)$

OPEN QUESTION

5

Explain the concept of object-oriented programming and give an example of real-world applications.

CODING

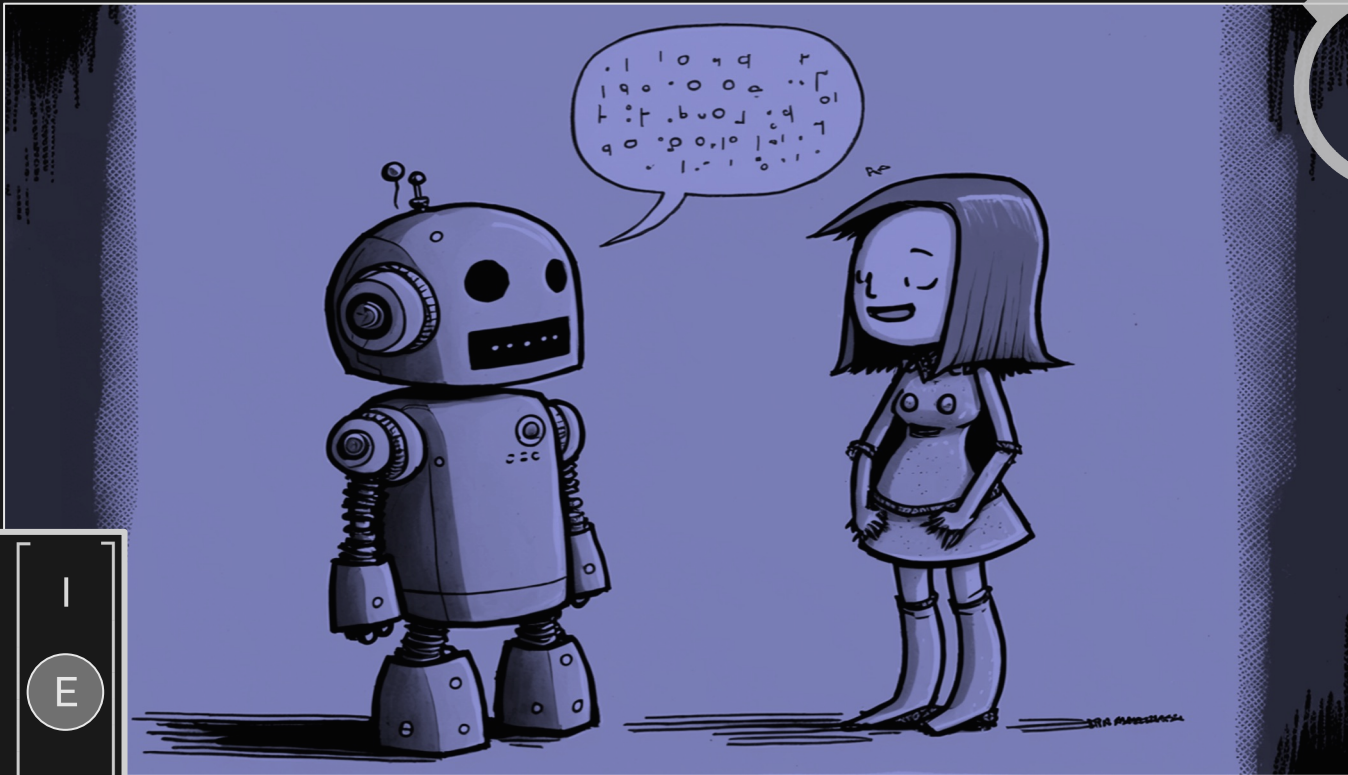
2

Implement a function in Python that takes a list of integers and returns the sum of all its even numbers.



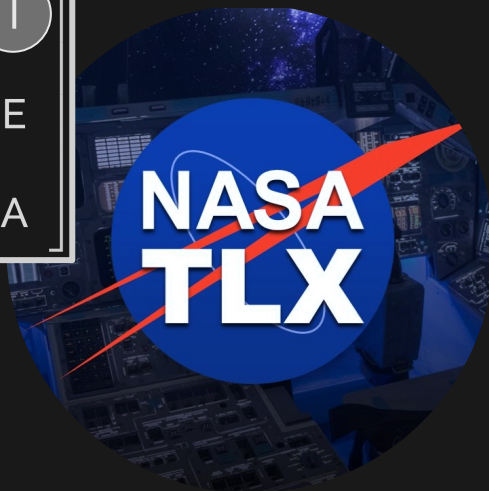
>>>> EFFICIENCY

Time for each task



| | |
|-----------------|---|
| Ineffectiveness | I |
| Efficiency | E |
| Accuracy | A |

Sandra G. Hart and Lowell E. Staveland. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in *Advances in Psychology*, vol. 52, 1988, doi: [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9).



Mental Workload

>>>> **INEFFECTIVENESS**

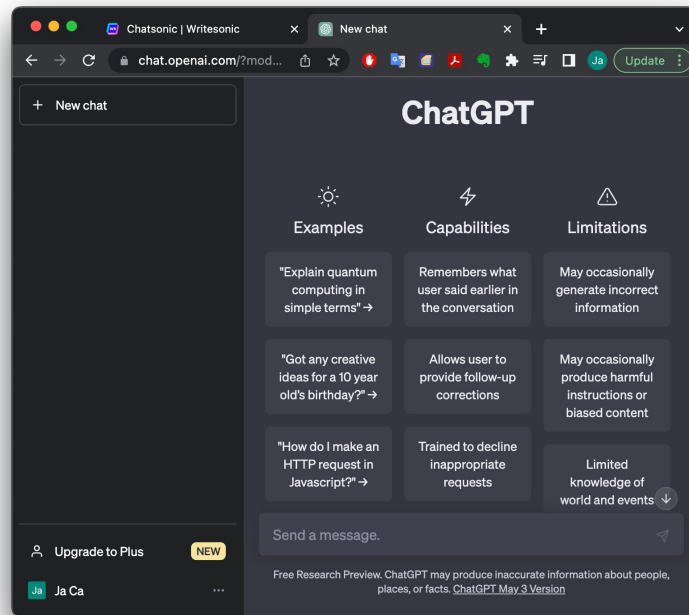
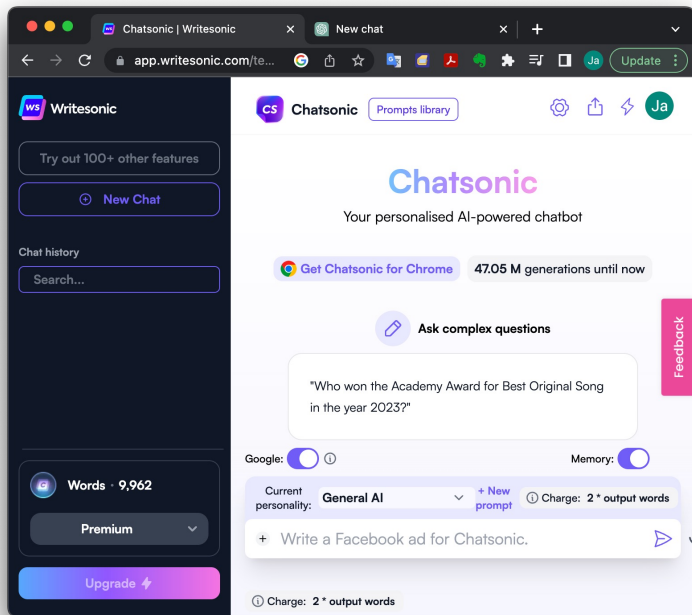
| Indicator | Question | Score |
|-----------------------------|--|-------|
| Mental Demand (MD) | In your opinion, how much mental effort is needed for your work? | 0-100 |
| Physical Demand (PD) | In your opinion, how much physical effort is needed for your work? | 0-100 |
| Temporal Demand (TD) | In your opinion, how much pressure do you feel is related to the time to do your job? | 0-100 |
| Performance (P) | In your opinion, what is your level of success in doing your work? | 0-100 |
| Frustration (F) | In your opinion, how much anxiety, pressure, and stress you feel is related to the time to do your work? | 0-100 |
| Effort (E) | In your opinion, how much mental and physical work does it take to complete your work? | 0-100 |

WHAT DO I EXPECT?

ChatGPT
wins...



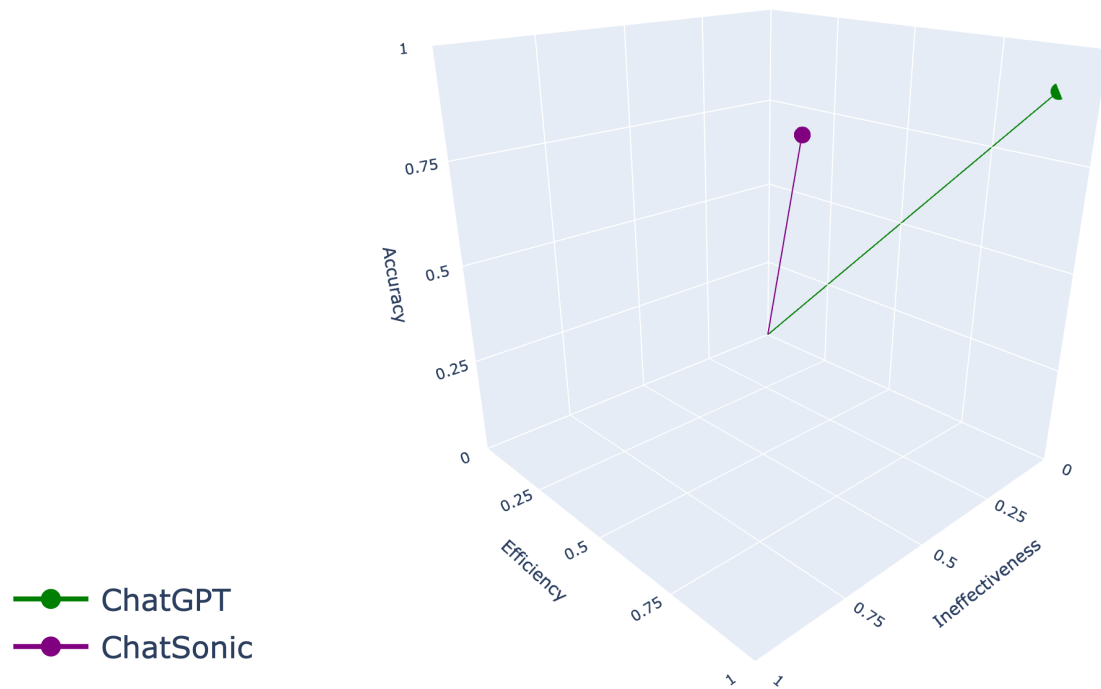
$W_1 = W_2 = W_3$ **PERFORMANCE P**



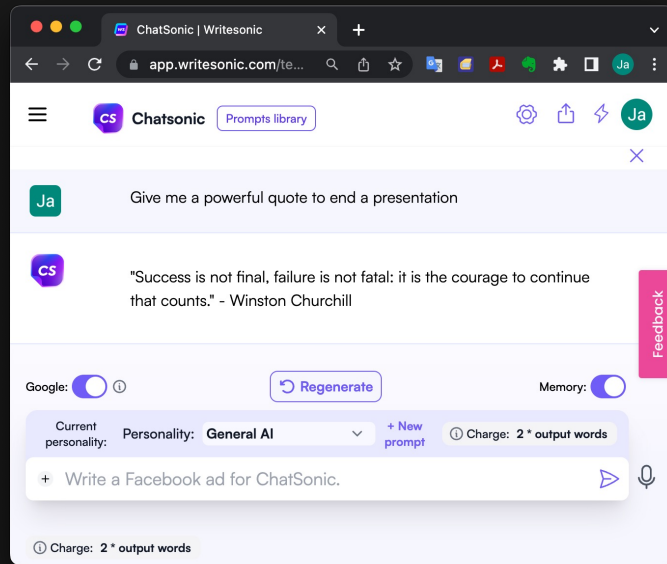
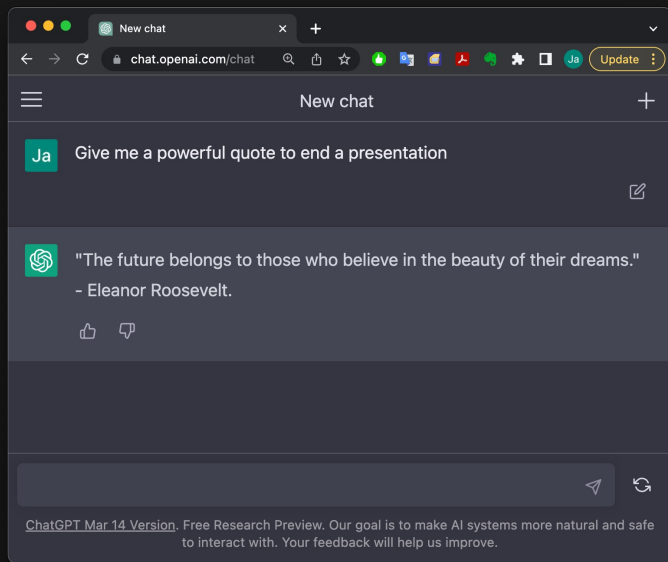
| | |
|---------------------|-------------------------------|
| 0.9 Accuracy | $P_{\text{chatSonic}} = 0.33$ |
| 0.8 Efficiency | |
| 0.7 Ineffectiveness | |

| | |
|---------------------|-----------------------------|
| 0.9 Accuracy | $P_{\text{chatGPT}} = 0.60$ |
| 0.9 Efficiency | |
| 0.0 Ineffectiveness | |

VECTOR P



AI/AI



Thank you | Do you have any questions?
jacopo.caratti@students.unibe.ch