FACULTY OF ENGINEERING AND COMPUTER SCIENCE
Department of Automotive and Aeronautical Engineering

HAW HAMBURG

Assessing ChatGPT for Aeronautical Applications

Evaluating answers given by OpenAI's ChatGPT 4.0 to aeronautical questions using a numerical system.

ChatGPT and other Large Language Models (LLMs) have taken the world by storm. Yet they should be employed only after careful consideration of their inherent risks and capabilities. If we wish to use LLMs, we need to find the tasks they excel at, while prohibiting or contextualizing their use in domains where they produce lackluster results. In order to do this, these domains need to be recognized and categorized, as well as a collection of best practices to be developed.

INTRODUCTION

Large Language Models have caused both anxiety as well as excitement in almost all domains of working life and education. An assessment is needed for each particular application as to which tasks can be adequately performed by Language Models, and for which tasks they are insufficient. As different tasks require different assessment methods, a granular approach must be chosen.

The use of new technology, especially non-deterministic technology, should not be implemented without careful consideration. The opportunities for streamlining and outsourcing of basic research and calculatory tasks is as large as the risk of misuse and root-fault propagation.

METHOD

The original method used here is comprised of assigning question-answers pairs several characteristics which allow a gradual assessment of their quality. To this end, question-answers pairs are assigned a Tier, ranging from Tier 1 to Tier 3, indicating their complexity. In addition, question-answer pairs can be distinguished with regards to the scope they are posed in. Tier I: Explanatory. Questions whose answer is an explanation, rather than a design decision or a calculation. Tier II: Evaluation. Questions whose answer is a design decision. Tier III: Calculation. Questions that are solved by calculation. While ChatGPT does not itself do calculations beyond a certain complexity, it does give formulas which then can be used to arrive at a value. Each Tier comes with its own qualitative aspects (Table 1) to be assessed, based on which the answers given by the model are graded on a scale from 0 to 1.0, from which a weighted average is calculated, indicating the overall quality of the answer.

Table 1: Aspect qualitative values by tier with characteristic questions.

Aspect Qualitative Values for Tier I and Tier II	Assessed Aspect	Characteristic Question
Q_t	Truthfulness	Is the answer given factually correct? Does it contain mistakes or misconceptions?
Q_l	Legibility	Is the answer worded in an legible, understandable manner? Are technical terms used appropriately?
Q_c	Completeness	Is the given answer complete? Does is provide all necessary information for the question to be considered as fully answered?
Q_a	Accuracy	Is the question answered accurately? Has the task been performed in a sensible way?
Aspect Qualitative Values for Tier III	**Assessed Aspect**	**Characteristic Question**
Q_e	Explanation	Is the method used by the model to arrive at the sought value understandable? Could the user reproduce the results from the model by manual calculation?
Q_v	Value, as a Deviation from a Referential Value	How significantly does the value given by the model deviate from the reference?

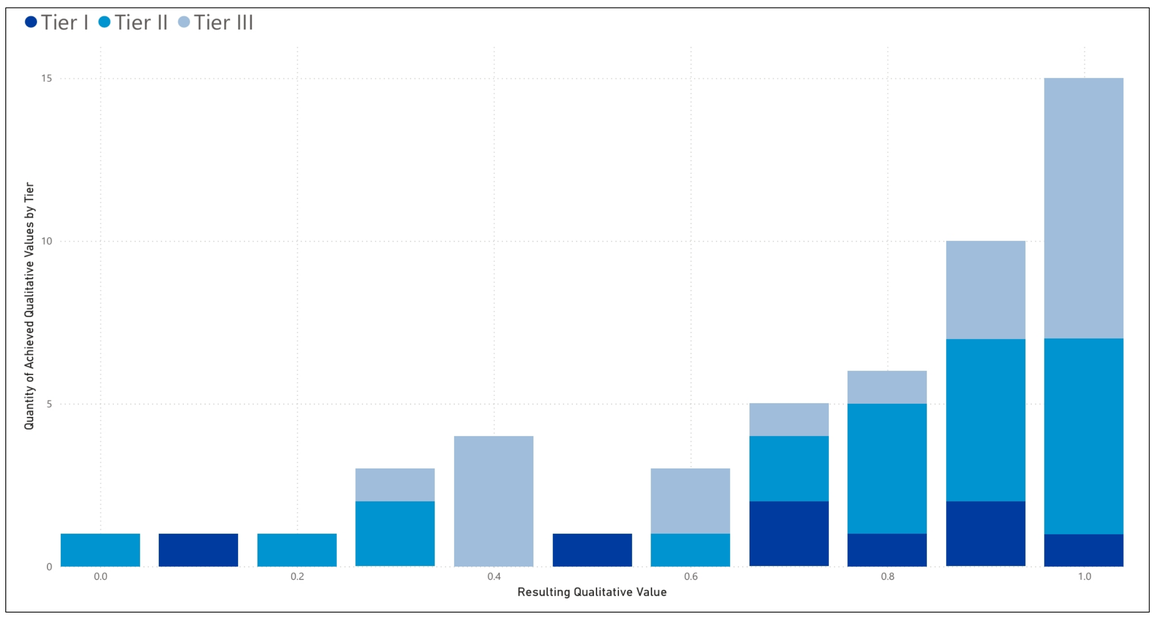

Figure 1: Distribution of resulting overall qualitative values.

RESULTS

Results indicate an overall positive outcome, but also underline the necessity to check and compare answers given by ChatGPT with literature and expert opinion. Figures 1 and 2 showcase that, while 60% of the answers provided can be considered good, answers with middling or faulty content make up 20% respectively. If these were taken at face value, slightly under half of the outcomes would have been non-satisfactory. In addition, Figure 1 indicates that the quality of an answer is generally independent of its complexity.

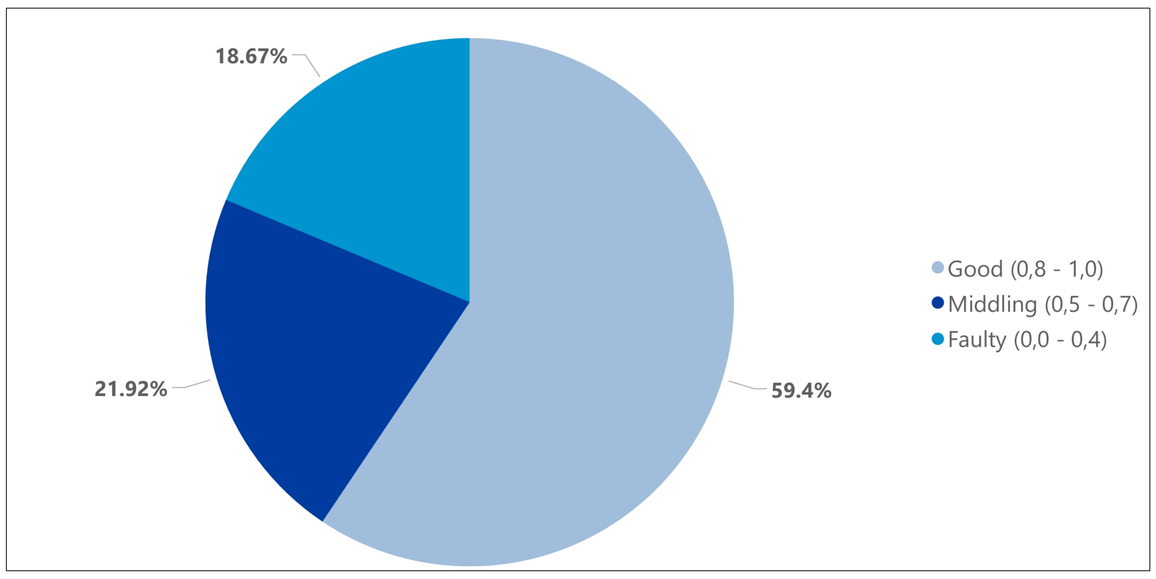

Figure 2: Results of qualitative value of all answers.

DISCUSSION

49 questions and answers have been evaluated. This allows only a rough assessment, due to the relatively low number. Further ways to assess LLMs could include more questions and more specialized assessments for specific tasks with Statistical Analysis. LLMs are overall a very agile field of development, and as such this study only represents a momentary assessment in a rapidly changing field. It is expected that the results will improve significantly, given further development of general and specialized models.

Bogdan Atanasoaie

Prof. Dr. Dieter Scholz, MSME

DGLR DLRK 2024 DEUTSCHER LUFT- UND RAUMFAHRTKONGRESS 30.09. – 02.10.2024 | HAMBURG

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences