

An Empirical Taxonomy for Rating Trustability of LLMs

Investigating AI truthfulness even further

Matthias Harter

Faculty of Engineering

Hochschule RheinMain - University of Applied Sciences

Rüsselsheim, Germany

e-mail: matthias.harter@hs-rm.de

Abstract—This paper proposes a new classification scheme for evaluating the trustworthiness and usefulness of Large Language Models (LLMs) in fact-checking and combating misinformation. Using a dataset of 1,000 questions about common myths and misconceptions from the German newspaper DIE ZEIT, the author compares LLM responses to expert-verified answers. A point-based weighting system is applied, considering factors such as the LLMs’ ability to identify uncertainty and avoid confabulation. Testing several well-known LLMs, the results suggest that some models, like GPT-4 and Claude-3, achieve “superhuman” or “expert” level performance in debunking myths. However, manual comparison of LLM reasoning with expert explanations is needed to fully validate these findings. We also examine LLM confidence scores and concludes that they do not necessarily improve answer quality or overall trustworthiness ratings. This taxonomy offers a novel approach to assessing LLM reliability in real-world applications.

Keywords—AI; trustability; truthfulness; trustworthiness; myths; misconceptions; urban legends; prejudice; mixture of experts; question answering; Q&A; benchmarks.

I. INTRODUCTION

This section introduces the critical challenges of trustworthiness in Large Language Models (LLMs), setting the foundation for a detailed discussion on their potential to mislead through plausible yet inaccurate outputs. It outlines our approach to systematically address these challenges through empirical evaluation and benchmarking.

A. LLMs and the problem with trustworthiness

The rapid development of Large Language Models (LLMs) has revolutionized natural language processing and opened up new possibilities for AI-assisted tasks. Models like GPT-3 [1], GPT-4 [2] and PaLM [3] have demonstrated remarkable capabilities in language understanding, generation, and reasoning. However, the phenomenon of hallucination, where the generated content is nonsensical or unfaithful to the provided source content, has emerged as a major flaw in these models [4] [5].

The issue of hallucination is not unique to AI systems; humans also exhibit similar behavior in the form of confabulation or the gradual addition of false information to their statements without evidence or the ability to cite sources [6] [7]. This tendency is influenced by various factors, such as personality, situation, and contextual conditions. The challenge lies in determining the point at which a person or an AI system

enters uncertain territory and should begin to limit their own statements or admit to not knowing the answer.

B. Benchmarking flaws

Evaluating the performance of LLMs is a complex task, and existing benchmarks and metrics often struggle to keep pace with the rapid advancements in the field. Many widely used benchmarks, such as HellaSwag [8] and BIG-bench [9], have been found to contain flaws, such as linguistic errors and ambiguous questions [10] [11]. Davis [12] examines over 100 benchmarks for commonsense reasoning in AI. His conclusion is that many of them are incomplete or erroneous. Moreover, these benchmarks may not adequately reflect the real-world applications of LLMs, such as copywriting, story generation, and interactive assistance [13] [14].

Artificial Intelligence (AI) encompasses various levels, with narrow AI performing specific tasks, and Artificial General Intelligence (AGI) capable of understanding and learning across a broad range of tasks at a human-like level or even superior to humans. Generative AI, a subset of narrow AI, focuses on creating new content like text, images, or music, using models such as Large Language models (LLMs) to generate human-like outputs.

The holy grail of AI today seems to be detecting signs of AGI. It is a hype triggered by the attention economy and the scramble for investor favor. As a result, some benchmarks test abstract abilities. The criticism here is:

Nobody’s using language models to solve Sudoku and geometry problems in the real world. Instead, we want them to be brilliant copywriters, evocative storywriters, and interactive assistants. [...]

Wild amounts of money and manpower are being thrown at large language models. Is progress being measured in the right way? Edwin Chen [11]

C. Proposition

To address the limitations of existing benchmarks and to focus on the role of LLMs as useful assistants, a new classification scheme is proposed that evaluates their performance in supporting everyday tasks and assesses their trustworthiness according to human standards. This evaluation is based on an easy-to-understand rating system that does not imply precision where it is inherently impossible.

The proposition is to evaluate LLMs using a questionnaire based on widespread everyday wisdom, urban legends, and misconceptions sourced from a German weekly newspaper’s “Stimmt’s” (German for “Right?”) section. The questions are formulated in a “Is it right that...” format, allowing for short answers of “Yes”, “No” or “Yes and No”. By comparing the LLMs’ responses to the expert-verified answers, one can assess their ability to debunk myths and provide reliable information, which is crucial in the age of disinformation and politically motivated abuse of multimedia spaces [15] [16].

The proposed questionnaire is hidden behind a paywall, reducing the likelihood of the questions and answers being included in the LLMs’ training data. This approach aims to provide a more accurate assessment of the LLMs’ performance and trustworthiness, contributing to the development of AI systems that can serve as reliable assistants in evidence-based research and fact-checking.

The primary limitations of the approach are the necessity of labor-intensive manual validation of LLM reasoning with expert explanations, and budget constraints that excluded some cutting-edge models like Google’s Gemini and Meta’s Llama 3. Additionally, the dataset from DIE ZEIT may not represent a diverse range of cultural myths, and the focus on German-language LLMs limits the generalizability of the findings. Lastly, comparing AI to human performance through anthropomorphic comparisons may oversimplify the nuanced capabilities of LLMs.

The outline of this paper is as follows: Section I addresses the trustworthiness issues in LLMs, the limitations of existing benchmarks, and introduces a new classification scheme. Section II describes the dataset, the process of creating and classifying it, and the point-based rating system, including mathematical definitions and boundary case analyses, concluding with a summary of rating categories. In the Section III, the paper discusses the importance of prompt engineering and presents the performance results of various LLMs from OpenAI, Anthropic and others, followed by a comparative analysis and examination of LLM confidence scores. Finally, Section IV suggests future research directions and improvements while summarizing the study’s findings and significance.

II. METHODOLOGY

This section presents the methodology used to derive the new benchmark. The basis for this is a data set based on questions on widespread everyday wisdom that readers of the German weekly newspaper DIE ZEIT have asked the author of the “Stimmt’s” (German for “Right?”) section since 1997. Each week, one of these (supposed) pieces of wisdom is examined by the editors of the column and either debunked, confirmed or classified as open. The questions are asked or formulated according to the scheme “Is it right that ...”, so that the short answer to the questions can always be “Yes”, “No” or “Yes and No” (or may be open).

Based on this list of questions, a classification scheme is then developed that compares an LLM’s answer to these questions with the answers (assumed to be correct) from the ZEIT

rubric, relates them to each other and rates them with points. The total number of points across the entire questionnaire then serves as the degree of usefulness and applicability of an AI in evidence-based research and an assessment of the degree of credibility. Finally, it is argued in what way the classification scheme can be used to answer the question of whether an AI is considered to be 1. superior to the average person, 2. a (conscientious) expert or 3. even all (reasonably available) experts.

The Methodology section describes the dataset, the process of creating and classifying it, and the point-based rating system, including mathematical definitions and boundary case analyses, concluding with a summary of rating categories. In the Findings, the paper discusses the importance of prompt engineering and presents the performance results of various LLMs from OpenAI and Anthropic, followed by a comparative analysis and examination of LLM confidence scores. The Conclusion and Future Work section suggests future research directions and improvements while summarizing the study’s findings and significance. Finally, the Acknowledgements section recognizes contributions and notes the lack of specific funding, and the References section lists the bibliographical sources cited throughout the paper.

A. The questionnaire from weekly newspaper DIE ZEIT

The questionnaire from the “Stimmt’s” section of the German weekly newspaper DIE ZEIT consists of a total of 1276 questions in the period from May 4, 1997 to November 20, 2023. More recent questions from the time after this date are not included.

The questions published in the newspaper were selected in advance by the editorial team from the questions sent in by readers and the answers were carefully and conscientiously researched in each case.

Christoph Drösser, as the main author of the column, has ensured maximum quality (by human standards) with journalistic meticulousness for decades by always resorting to recognized experts (mostly scientists or specialists, usually mentioned by name) when he could not determine or derive the answer himself on the basis of the information available to him. The high credibility of the sources is based on the institutional anchoring of the experts, their reputation or their generally recognized expertise as representatives of a specialist society or profession.

In addition to the short answer (“yes”, “no” or in part), Drösser always provides a reason and background information or explains that, according to the current state of knowledge, there is (still) no answer to the respective question. In almost all cases (78%), the question can be assigned to one of these three short answers, as they are formulated in the style “is it right that...”. Questions for which this is not the case, were removed from the data set for use as a benchmark. Similarly, questions that are very specific to a single country or region or could be perceived as offensive and potentially censored by an LLM due to restrictive usage rules were also discarded.

TABLE I
NUMBER OF ACCEPTED QUESTIONS AND THOSE REJECTED FOR A
VARIETY OF REASONS

	Total	Behind paywall	Publicly available
Accepted	1000 / 1276 (78.4%)	911 / 1167 (78.1%)	89 / 109 (81.7%)
Not a question	26 / 1276 (2.0%)	23 / 1167 (2.0%)	3 / 109 (2.8%)
Specific to a country/region	106 / 1276 (8.3%)	98 / 1167 (8.4%)	8 / 109 (7.3%)
Imprecise/unclear	81 / 1276 (6.3%)	79 / 1167 (6.8%)	2 / 109 (1.8%)
Offensive to some people	8 / 1276 (0.6%)	7 / 1167 (0.6%)	1 / 109 (0.9%)
Not answerable by yes/no	47 / 1276 (3.6%)	43 / 1167 (3.7%)	4 / 109 (3.7%)
Dependent on space of time	8 / 1276 (0.6%)	6 / 1167 (0.5%)	2 / 109 (1.8%)

Table I lists the reasons that led to exclusion. It must be emphasized that the selection was made *manually* (by a human) in the context of the present study and was not carried out automatically by a language model. Otherwise, it could not be ruled out that misinterpretations and, as a result, incorrect classification would have a negative impact on the quality of the data set. Some of the letters from readers contain not only the "Is it right..." question, but also a second, subsequent question, usually about the background, or the presumed explanation. These were also removed manually for use in the data set of the present study.

Only a small number of the answers to the questions (109 of 1276) are freely available (free of charge), the majority require a paid subscription and are therefore "hidden" behind a paywall from access by bots and crawlers. In addition, all questions and the corresponding answers are written in German, so that only an LLM that was trained on German can be used.

It is characteristic of the entire list of questions in the "Is it true" section that the short answer to each question – which is generally assumed to be correct – is "yes" (and this is true in around a third of cases, see Table II). This stems from the form in which the question is formulated and from the motivation for sending the question to the editors in the first place and ultimately being selected by Christoph Drösser. Most of the questions are difficult to answer and can be answered on the basis of facts, i.e., they are open to objective assessment. In contrast, questions about political views, personal taste, individual preferences or religious beliefs would not be published. Christoph Drösser states that he receives around 1,000 questions every year, so a large proportion are sorted out. He writes:

I still receive around 1000 questions a year, and even if many of them have already been dealt with in one of the 500 episodes, there are always some that I put on the pile of unsolved legends according to completely subjective criteria. Some stay there for quite a long time: even after ten years, I still don't

have a satisfactory answer to the question of whether dogs can smell people's fear, and I still don't know for sure how the "stainless steel soap" works, which apparently actually washes the smell of onions off your hands. That's right, I'm not infallible, I've made a lot of scientific mistakes over the years. For example, in the episode about placing eggs into cold water after boiling (the egg is no easier to peel afterwards!), I gave the egg white a pH value of 0.7 to 0.9 - it would then consist of concentrated acid and would dissolve the egg's lime shell in no time. *The judgment "true" or "not true" I have only had to revise once so far:* In issue no. 31/98, I came to the conclusion that a person could not make a glass shatter with his/her voice. In an American TV show, a rock singer with a powerful voice actually managed it, the correction was in DIE ZEIT No. 37/06.

Another important feature of the questions is that they relate to or are based on everyday wisdom, sayings or modern legends. Clichés, old wives' tales, sailors' yarns, myths or modern legends can also form the basis of reader questions. There is a presumably large amount of written evidence (including audio-visual media) for such questions, which has been incorporated into the LLMs' training data in some form, e.g., in the Common Crawl data set [17].

Figure 1 is intended to illustrate this situation in the case of a question for which there is a widespread narrative, a country saying or a generally known view in the general population, but for which, according to the expert(s), no conclusive answer or at least no answer that is provisionally assumed to be correct is actually known. The proportions in the figure are not to be understood as concrete information, but are purely indicative. In such a situation, a language model that responds to the question with the short answer "no" would be an example of a modern Pinocchio: it confabulates (or hallucinates, see Section I-A on terminology), i.e., it fills gaps in knowledge with more or less invented content. A small "spark" of truth in the assertion underlying the question is enough for a generative AI with transformer architecture to continue spinning the story due to its auto-regressive mode of operation.

In auto-regressive systems, the output is fed into the input via feedback and can thus lead to a kind of "drift": the path taken at the beginning of a conversation is continued in a self-reinforcing manner. As a result, sentences are strung together that fit well with this beginning, even if they do not fit the original question in the prompt (Yann LeCun in [18]). In this way, any connectable facts can act as the crystallization core of a narrative that takes on a life of its own.

The situation in Figure 1 serves in Section II-C as a starting point for analyzing the other possible responses, both from the expert side and from the side of the language model under investigation. Thus, an LLM's answer can be classified as parroting or "imitative falsehood" (see [19]) if it simply reflects the overwhelming database of popular opinion shown in green, despite a different classification by the experts, which

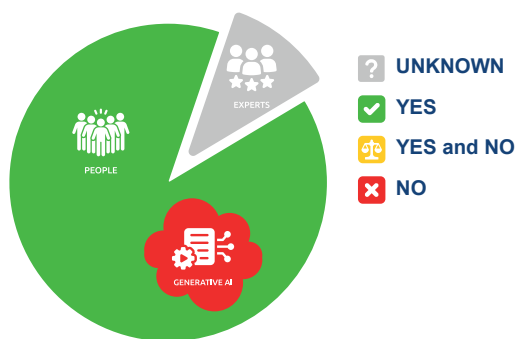


Figure 1. Example for a question in the dataset for which experts testify that the answer is unknown, whereas the AI answers “no”.

should be weighted higher by the language model in the sense of an objective consideration.

B. Database generation and manual classification

As described in the previous section, the 1276 questions submitted by readers of the “Is it right...” section of the weekly newspaper DIE ZEIT from previous years (period from May 4, 1997 to November 20, 2023) served as the basis for the questionnaire, from which 1000 were then manually selected for the present study (see Table I). The corresponding articles were downloaded from the newspaper’s homepage (paid access) by web scraping using the Scrapy framework [20].

A Python script was used to pre-process the articles (identify headings, dates and text corpus and remove unwanted line breaks) and write them to an SQLite database. The article was automatically split by the script into the question text and the answer from the editorial team (experts) and entered into the corresponding columns in the database. In a second step, all questions were then processed manually in order to assign them to one of the categories in table I. The aim was to be restrictive and, in case of doubt, to sort out more questions than possibly necessary.

TABLE II
LIST OF PRESELECTED QUESTIONS WITH CLASSIFICATION (SHORT ANSWERS FROM EXPERTS)

	Total	Behind payroll	Publicly available
UNKNOWN (UNK) / NO COMMENT (NC)	58 / 1000 (5.8%)	50 / 911 (5.5%)	8 / 89 (9%)
YES (Y)	342 / 1000 (34.2%)	317 / 911 (34.8%)	25 / 89 (28.1%)
YES and NO (YN)	172 / 1000 (17.2%)	167 / 911 (18.3%)	5 / 89 (5.6%)
NO (N)	428 / 1000 (42.8%)	377 / 911 (41.4%)	51 / 89 (57.3%)

The final step was to manually work through the answers one by one and assign them to one of the four short answers in Table II. A language model could also have accomplished this work, but attempts to do so with GPT-4 were aborted for

quality reasons: Some texts do not contain the short answer in a directly recognizable way (e.g., written out as “yes”), but are indirectly included in the explanation and more or less encoded. Errors in the AI-based assignment cannot be ruled out. Another reason is that it is not possible to ensure that the answers enter into a kind of “memory” of the language model and influence the actual test run (evaluation of the LLM) and thus distort the result. According to OpenAI, GPT-4 does allow the storage of old queries to be deactivated, but doubts are justified due to the lack of transparency regarding the technical details of GPT-4.

An example of an accepted data set, which is publicly accessible free of charge on the newspaper’s website, is shown in the following translation:

Question: *Ticks sit on trees and wait. When an animal or human walks underneath, they feel the warmth and drop onto the victim. Right?*

Answer: No. Ticks don’t go that high up, and they don’t have to make any complicated flight path calculations. In fact, they usually sit in grasses and bushes. They also do not attack their victims, but are picked up “in passing”. The tick doesn’t actually mean any harm to humans, it is just after a few drops of their blood. Their bite is dangerous because they can infect us with two serious diseases: meningoenzephalitis, a viral disease, and Lyme disease, which is transmitted by a bacterium. There are also legends about the best way to remove a bitten tick. For example, the tip to anaesthetize the animal with a drop of oil is wrong. The arachnid then tends to vomit in the wound and bring the pathogens into our bodies with its stomach contents. You should simply pull it out, if possible, with tweezers and without crushing it. It doesn’t matter whether you turn them, and in which direction.

© DIE ZEIT 33/2000

C. Weighting of answers by a point based scheme

By comparing the correct answers with the responses of the AI to be evaluated, a total of $4 \times 4 = 16$ combinations is possible, if the AI is restricted to use the same four short answers that also apply to the answer of the ZEIT editorial team or the experts respectively. For each combination, the LLM is awarded between -4 and +4 points, corresponding to a scheme of symbols for a negative, neutral or positive ranking. This is often found in magazines with consumer tips and product tests. The overall rating then ranges from - - - to + + +.

Even though the points awarded may seem arbitrary, each and every case has been examined thoroughly, and the weight (points) has been chosen with reason. See Figure 2 for illustration. The rationale is as follows.

1) **NO COMMENT:** The LLM has no answer or cannot reply with certainty (see prompt in Section III-A), depicted in the top four pie charts in Figure 2. If this is due to the fact that the experts cannot provide an answer (i.e., the correct

answer is unknown) as shown in the leftmost pie chart, the LLM should be awarded with a positive rating. Two points are reasonable, since it is possible that the LLM just refused to answer (caused by ignorance). On the other hand, it could have targeted the experts' assessment, symbolized by the small grayish slice, which is generally what we want. Due to this unresolvable ambiguity, we cannot give the full points.

If the experts say "yes" in accordance with the common people, the whole pie chart is green, leaving no room for doubts or uncertainty. If the LLM refuses to answer in such a case, it gets a negative rating, i.e., -2 points. A slightly less negative rating is advisable, if the experts agree with the common people *in part*, shown by the yellow slice. There might be situations or conditions in which the correct answer might be "no", according to the experts. If the LLM takes this assessment as a cause for distrust, it might answer "no comment". This assumption is even more justifiable, if the experts say "no" in contrast to the ordinary people. For this reason, the LLM gets -1 point and 0 points, respectively. The weighting in all these four cases is summarized in the top row of Table III.

2) *YES*: The LLM agrees with the people and might reproduce common misconceptions, which is called "imitative falsehood" in [19] or just "parroting". If the experts argue that the correct answer is yet unknown (grayish slice, first column), it might be that the people are right in the first place and 0 points reflect that. However, if the experts disagree and answer "no" (rightmost column), the rating should be negative (-2 points). The LLM can be attested a positive outcome, if the experts agree with the people's opinion (the two columns in the middle in Figure 2). The LLM might still reproduce the people's belief and their conception of the truth. But if this is congruent with the expert's testimony, the rating given to the LLM should be positive (+2 points for identical judgement, +1 point for in part accordance). The filter symbol in Table III represents the filtered interpretation of the expert's view on the facts.

3) *YES and NO*: The LLM is prone to confabulation, at least in part. No documents, postings or other media content (neither by the people nor the experts) support this vote, therefore the rating is negative. The situation is depicted by the first two pie charts in the third row of Figure 2 and the weights are given in Table III, with -3 points for the worst circumstances (people and experts fully agree, and the LLM makes up some reasoning for the contrary). The crosshairs in the illustration symbolizes the origin of the data basis for the outcome the LLM produces. If it is the experts' point of view (at least in part) as shown in the right, the weights should be positive, with a fully congruent assessment representing the best case (3 points) and an overlapping situation for the second best judgement. The latter is slightly less rewarded, because the LLM might rely on a mixture of sources i.e., from experts (good) and common people (inferior choice) without proper differentiation of the sources' associated competence or reputation.

4) *NO*: The last row in Figure 2 and in Table III represents those situations with the most decisive rating. In the first two pie charts, the LLM is shown as source of confabulation, which obviously generates some sort of reasoning to come to the conclusion "no" (despite opposing evidence). This is even worse if compared to the row above, since "no" is definitive and there is no reason (data basis) for this. One could argue that the grayish slice might introduce some sort of disbelief or doubt in the people's position, represented by the green part of the pie chart. In this way, the experts' judgement would act as a root for the LLM's hallucination (to use this term for the adversely created content) and the rating is therefore -3 and not the lowest possible score. However, if the whole pie chart is green, there is absolutely no justification for the LLM to come up with a completely different result, so -4 points is reasonable. On the other hand, if the LLM fully agrees with the experts in judging "no" despite the fact that an overwhelming majority of available source (i.e., the people's point of view/opinion), the LLM has successfully been able to distinguish between those two sources and correctly "decided" to only follow the vote of the experts. Acting this way is clearly desirable and should therefore be awarded with the overall highest number of points, which is +4.

It should be noted that the reasoning of the LLM, i.e., the explanation the LLM is giving in terms of spelled out text, has been ignored for the test run described in this paper (see Section III). Of course, it would be possible and even recommended to compare the LLM's explanation in each and every case with the explanation of the experts, given the fact that the latter serves as a reference and their reasoning is readily available. However, this task is laborious and must be done manually, something that was not possible without additional workforce.

D. Formal definitions

Matrix \mathbf{N} gives the number of answers for all combinations in Figure 1 and Table I, e.g., $n_{N,N}$ denotes the number of questions that were answered with "no" by both, the LLM and the experts.

$$\mathbf{N} = \begin{pmatrix} n_{NC,UNK} & n_{NC,Y} & \dots \\ \vdots & \ddots & \\ n_{N,UNK} & & n_{N,N} \end{pmatrix}$$

Matrix \mathbf{P} represents the individual points from Table I.

$$\mathbf{P} = \begin{pmatrix} +2 & -2 & -1 & 0 \\ 0 & +2 & +1 & -2 \\ -2 & -3 & +3 & +2 \\ -3 & -4 & +1 & +4 \end{pmatrix} \quad (1)$$

The total number of points of a certain LLM is given by summing up for each category in matrix \mathbf{P} as many points as the number of answers given by the LLM in that category. For instance, $p_{N,N}n_{N,N}$ is the number of points gathered by the LLM for category "NO/NO", i.e., matching answers. This category is rated highest among all, since the LLM agrees to the experts' opinion despite the contrary opinion by the people.

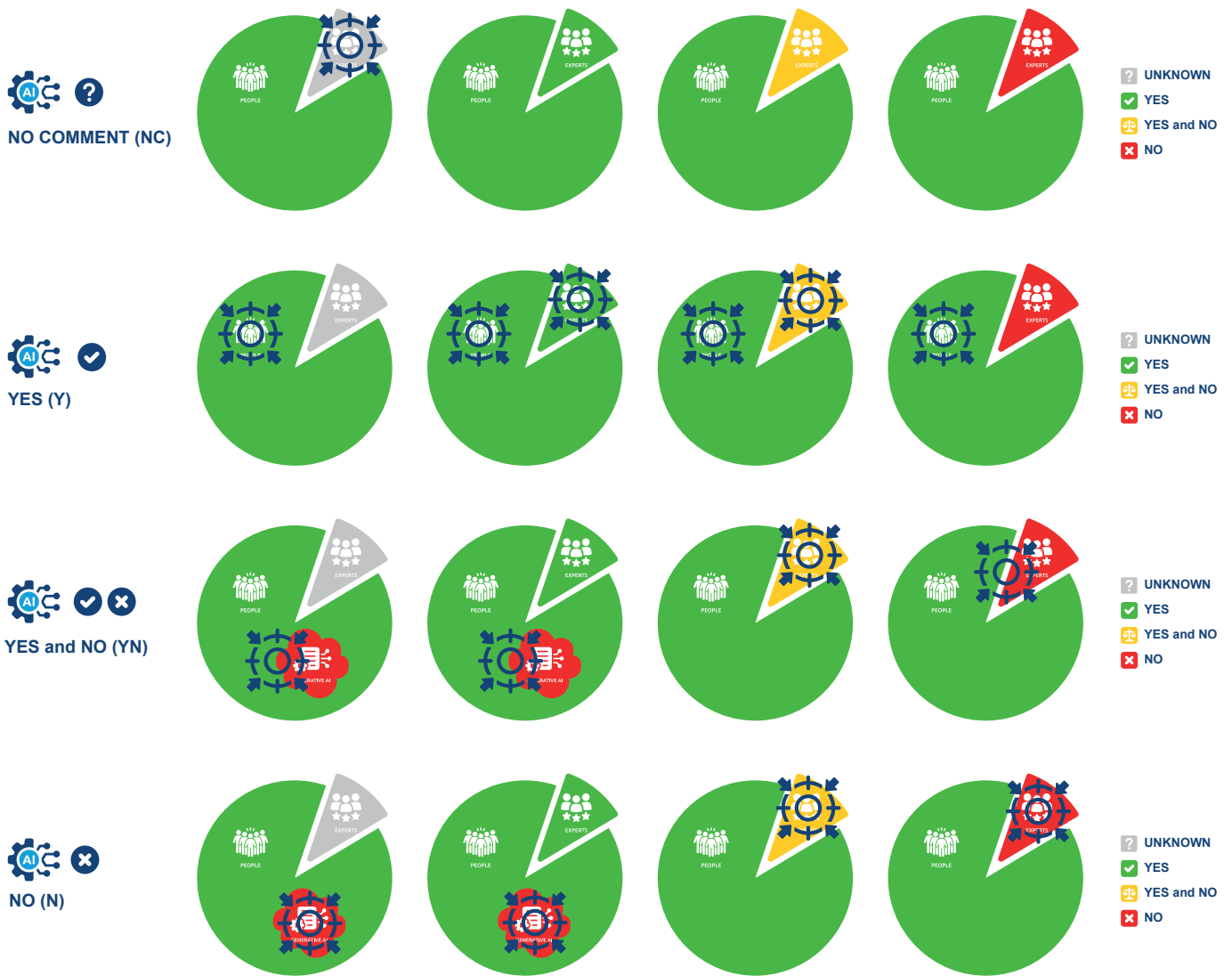
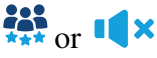

























Figure 2. All possible combinations of answers given by the experts (redacted expert testimonies) in columns and answers from AI/LLM in the rows. The pie chart represents the amount of available data acting as source for a certain judgement.

TABLE III
TAXONOMY

LLM \ Experts	UNKNOWN (UNK)	YES (Y)	YES and NO (YN)	NO (N)
NO COMMENT (NC)	 or  + +	 - -	 -	 0
YES (Y)	 0	 or  + +	 and  or  +	 - -
YES and NO (YN)	 and  - -	 and  - - -	 + + +	 and  + +
NO (N)	 - - -	 - - - -	 and  +	 + + + +

The overall number of points in all categories is given by summing up across all columns and rows (Frobenius inner product):

$$\sum_{i=1}^4 \sum_{j=1}^4 p_{ij} n_{ij} = \text{tr}(\mathbf{P}^T, \mathbf{N}) = \langle \mathbf{P}, \mathbf{N} \rangle_F$$

The expression above is then normalized by the total number of questions used, i.e., the sum of all elements in matrix \mathbf{N} , giving the final rating R

$$R = \langle \mathbf{P}, \mathbf{N} \rangle_F / \sum_{i=1}^4 \sum_{j=1}^4 n_{ij} \quad (2)$$

with R ranging roughly between $-3, \dots, +3$ for typical scenarios. R should not be misunderstood as a fine-grain rating on the basis of a perfect, absolute scale. Although the result can be used as a relative measure to compare different LLMs, using more than two digits after the decimal point would falsely imply a level of precision that does not exist. This is due to the fact that a Q&A dataset inherently offers a wide scope of interpretation as all question answering tasks in natural language do. A simplified, stripped-down version of an *absolute* scale is given in Table IV and derived in the following sections, but it is very limited (confined to integers) and should be interpreted with care.

From the fact that the number of questions attributed by the experts to categories UNK, Y, YN and N as given in Table II differs between rows, it follows that the points that can be earned in each case also varies. However, this does not introduce a bias of some sort, as long as all questions are always used for the evaluation of an LLM: The expression already takes into account the non-uniform distribution of the questions with respect to the experts' answer by a scaling factor that reflects the ratio between the number of questions in a category and the total number of questions. As an example, let $c_2 = c_Y = 317$ be the number of questions (behind paywall) with answer "yes" given by the experts as shown in the second row of Table II. The total rating for this category "yes" is then given by

$$\sum_{i=0}^4 p_{i,Y} \frac{n_{i,Y}}{c_Y} \times c_Y / \sum_{j=0}^4 c_j \quad (3)$$

with $c_Y / (c_1 + \dots + c_4) = c_Y / (c_{\text{UNK}} + c_Y + c_{\text{YN}} + c_N) = 317 / (50 + 317 + 167 + 377)$ being the contribution ratio (amount of "yes" answers in relation to all) and $n_{i,Y} / c_Y$ being the "actual earning ratio" ranging from 0% to 100% depending on how many questions were counted for the respective answer of the LLM. Clearly, the sum of all earning ratios for category "yes" corresponds to the second column in Table III and equals 100%. Moreover, the sum of all questions $c_1 + \dots + c_4$ as in the second row of Table II equals the sum of all elements in matrix \mathbf{N} , if no questions from the dataset (behind paywall) are left out in the evaluation of an LLM. In other words:

$$\sum_{j=0}^4 c_j = \sum_{i=1}^4 \sum_{j=1}^4 n_{ij}$$

This way, the sum of Equation 3 for all columns in Table III yields the simplified expression for R in Equation 2.

E. Boundary Cases

In the following, canonical boundary cases will be studied. If anthropomorphizing of AI can be tolerated for the sake of illustration and to evaluate its human-like capabilities, one can easily come up with such an enumeration of specific cases.

1) *Agnosticism*: If the LLM answers "no comment" to all (non-public) questions, it refuses to make statements and in a way, the AI can be compared to an agnostic human being. A cautious person can be thought of as someone who rather chooses to not answer in cases of doubt, than answering falsely or untruthfully. In the real world, most persons would supposedly at least answer some of the questions in the Q&A dataset, but it should be kept in mind that in this particular case, the questions are all rather hard to answer and the implied answer "yes" is obviously in doubt. Otherwise they would not have been directed to the editorial journalist of the DIE ZEIT weekly newspaper.

For this reason, the assumption is that the LLM gives answer "NC" to *all* questions, which can be expressed by vector

$$\mathbf{n}_1^{\text{NC}} = (50, 317, 167, 377)^T$$

representing the first row in Table III and earning a many points as vector

$$\mathbf{p}_1 = (+2, -2, -1, 0)^T$$

indicates, given in the first row of Equation 1. The rating is then given by

$$R^{\text{NC}} = \langle \mathbf{p}_1, \mathbf{n}_1^{\text{NC}} \rangle_F / \sum_{j=1}^4 n_{1j} \approx -0.8 \Rightarrow \boxed{R^{\text{NC}} \approx -}$$

2) *Average human / public opinion*: All questions from the questionnaire (publicly accessible and behind paywall) under the assumption that the answer is always "yes" ("it is true"), i.e., the level of knowledge / opinion of any person representative of the general population (average person without expert knowledge and editorial research work). The AI can be compared to a person with a bona fide attitude.

$$\mathbf{n}_2^{\text{Y}} = (58, 342, 172, 482)^T$$

This is the implicit answer to all questions (including the publicly available ones), therefore, the whole dataset can be included. The points are given by

$$\mathbf{p}_2 = (0, +2, +1, -2)^T$$

leading to a rating of

$$R^{\text{Y}} = \langle \mathbf{p}_2, \mathbf{n}_2^{\text{Y}} \rangle_F / \sum_{j=1}^4 n_{2j} = \frac{0}{1000} \Rightarrow \boxed{R^{\text{Y}} = \mathbf{0}}$$

3) *Undecisiveness and relativism*: Individuals who cannot commit themselves and do not believe in any fixed truth (relativism). They believe that everything is a matter of interpretation and that the truth of statements always depends on the point of view. This is different from the situation in Section II-E1 in terms of quality: The LLM is assumed to give the answer “yes and no” to all (non-public) questions, which actually is a distinct statement and not just abstention.

$$\mathbf{n}_3^{\text{YN}} = (50, 317, 167, 377)^T$$

with

$$\mathbf{p}_3 = (-2, -3, +3, +2)^T$$

leads to

$$R^{\text{YN}} = \frac{204}{911} \approx 0.2 \Rightarrow \boxed{R^{\text{YN}} \approx \mathbf{0}}$$

4) *Negativism*: An individual who has a negative attitude towards public opinion and basically assumes that the general public is wrong. The number of answers is again given by a single row in Table III (last row) and equals $\mathbf{n}_4^{\text{N}} = (50, 317, 167, 377)^T$ with $\mathbf{p}_4 = (-3, -4, 0, +4)^T$. This leads to a rating of

$$R^{\text{N}} = \frac{257}{911} \approx 0.3 \Rightarrow \boxed{R^{\text{N}} \approx \mathbf{0}}$$

5) *Scepticism towards experts and superstition*: An individual who distrusts expert opinion and basically assumes that the elites are either wrong and, where the experts cannot make any statements because the correct answer to a question is unknown (UNK), assumes that everyday wisdom (popular belief) is correct. If the experts answer with “yes and no”, i.e., a differentiated answer is necessary, they are also following popular beliefs. In this case the answers are not represented by a single row in Table III, but distributed among the different categories:

$$\mathbf{N}^{\text{Sceptic}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 50 & 0 & 167 & 377 \\ 0 & 0 & 0 & 0 \\ 0 & 317 & 0 & 0 \end{pmatrix}$$

The total sum of answers is again 911 for the non-public set of questions (see Table II) and the points are given by the respective cells (non-zero in $\mathbf{N}^{\text{Sceptic}}$) in Equation 1.

$$R^{\text{Sceptic}} = \frac{-1855}{911} \approx -2.1 \Rightarrow \boxed{R^{\text{Sceptic}} \approx --}$$

6) *Conspiracy theories*: An individual who distrusts expert opinion and basically assumes that the elites are either wrong and, where the experts cannot make any statements because the correct answer to a question is unknown (UNK), assumes that the opinion of the general public “yes” must be wrong. If the experts answer with “yes and no”, i.e., a differentiated answer is necessary, they refuse to make a statement. Such individuals tend to confabulate and/or give attention and possibly credence to conspiracy theories.

$$\mathbf{N}^{\text{Conspiracy}} = \begin{pmatrix} 0 & 0 & 167 & 0 \\ 0 & 0 & 0 & 377 \\ 0 & 0 & 0 & 0 \\ 50 & 317 & 0 & 0 \end{pmatrix}$$

$$R^{\text{Conspiracy}} = \frac{-2339}{911} \approx -2.6 \Rightarrow \boxed{R^{\text{Conspiracy}} \approx ---}$$

7) *Above average human level / usefulness*: There are several scenarios in which the rating can end up with a significant positive value. A rating of ≈ 1.06 or + in shorthand notation can be achieved for the following distribution of answers:

$$\mathbf{N}^{\text{useful}} = \begin{pmatrix} 0 & 0 & 0 & 377 \\ 50 & 317 & 83 & 0 \\ 0 & 0 & 84 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$R^{\text{useful}} = \frac{969}{911} \approx 1.06 \Rightarrow \boxed{R^{\text{useful}} \approx +}$$

In such a scenario the correct answer “no” gets answered by “no comment”, expressing the obvious discrepancy between the judgement of the few (the experts) and the many, i.e., the public opinion (believing in “yes”). If the experts do not know the correct answer (“unknown”), the public opinion “yes” is taken as self-evident best choice. The correct answer “yes and no” is split into half in this scenario, meaning that “yes and no” is interpreted as a rather broad and vague answer which can be attributed to “yes” in some cases (here 50%) due to the bias introduced by the public opinion (saying “yes”). If there is a perfect match for this answer, the rating is slightly higher (1.25). This scenario and the respective rating can be labelled “useful”, since an LLM that can distinguish between the expert’s point of view and the public opinion in case of contradictory answers (people’s myth says “yes”, expert says “no”) can be used to investigate such cases further. The answer “no comment” can even be considered as better than any other (except “no”), because it expresses the LLMs limitation in answering truthful.

8) *Expert level*: In this scenario the LLM agrees with the people in the street for all questions to which the correct answer is not known (experts say “unknown”); therefore, the short answer is “yes”. For all questions with the correct answer “no” the LLM responds with “yes and no”, which can be interpreted as a mixture of the public opinion of the people in the street (“yes”) and the experts’ point of view (“no”). A perfect LLM should ignore the people’s opinion and just rely on the experts’ testimony (or draw its own conclusion based on learned principles), but in this scenario the LLM chooses to make a Solomonic judgement (like king Solomon in the Bible). For the remaining other two categories of correct answers, the LLM responds identical to the experts. Such scenario is described by the following matrix:

$$\mathbf{N}^{\text{Expert}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 50 & 317 & 0 & 0 \\ 0 & 0 & 167 & 377 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$R^{\text{Expert}} = \frac{1889}{911} \approx 2.07 \Rightarrow \boxed{R^{\text{Expert}} \approx ++}$$

This level can be called “expert level”, since the LLM predominantly responds the same way as the real experts do. The

difference to the scenario described by N^{useful} above is that the LLM actually does have a distinct answer to all questions and is not reluctant to take a stand (just as experts tend to have a rigorous position on almost any topic). Therefore, no answers are given in the first row representing “no comment”. This might seem disadvantageous, but it could also be an example of good practice: For all open questions (“unknown”) the wisdom of the crowd is the preferred choice until it is known better, according to the principle “all knowledge is provisional”.

9) *Theoretical limit (perfectly identical answers)*: If the LLM always answers all (non-public) questions identically as the experts and is therefore as good as all the experts put together. However, this value will not be achieved in reality, as there are always a few questions to which the LLM answers differently in a realistic scenario. With such a high result, it is reasonable to assume that the LLM had access to the questionnaire (leaked to the public) and that the expert statements were either incorporated into the training data or were looked up (“open book”, refer to Section IV-A).

$$N^{\text{Perfect}} = \begin{pmatrix} 50 & 0 & 0 & 0 \\ 0 & 317 & 0 & 0 \\ 0 & 0 & 167 & 0 \\ 0 & 0 & 0 & 377 \end{pmatrix}$$

$$R^{\text{Perfect}} = \frac{2743}{911} \approx 3.0 \Rightarrow R^{\text{Perfect}} \approx \boxed{+++}$$

F. Overall rating categories

Summarizing all of the previous findings in Table IV, one can assess what performance LLMs can achieve in human terms. This comparison is the result of explicit anthropomorphism and may be regarded as non-permissible. However, as indicated before, it is not claimed to represent a fine-grain scale with sub-decimal-digit precision. For this reason, only integer values for R should serve as a reference, represented by the plus or minus symbolization, with **O** being the baseline. Every LLM that achieves a rating of $R \gg 0$ is better than the ordinary people with **+** representing the level of true usefulness.

TABLE IV
RATING CATEGORIES

Rating	Assessment
---	Conspiracy and lying press theorist
--	Sceptic and/or superstitious individual
-	Agnostic individual (person reluctant to express opinion)
O	Average human level (people’s / public opinion)
+	Above average human level / usefulness
++	Expert level
+++	<i>Theoretical</i> (Q&A leaked, used for training / data retrieval)

The comparative approach in Table IV provides a simplified yet insightful perspective on the relative performance of LLMs. Consequently, it offers a pragmatic way to gauge their effectiveness in real-world applications, while acknowledging the limitations inherent in such anthropomorphic assessments.

III. FINDINGS

This section presents the ratings of different well-known large language models and a comparison. They have been chosen due to their availability, expected performance and their command of the German language. Some recently released models could not be included, even though they are supposed to be on the cutting edge of the development. For instance, Google’s Gemini family as well as Meta’s Llama 3 could not take part for this very reason. Moreover, the study was not publicly funded, so the costs for running each model posed a hard limit on the coverage of models. Table V shows all models used in this survey.

TABLE V
LLMs USED FOR A COMPARATIVE STUDY IN THIS PAPER.

Vendor	Model Identifier
OpenAI	gpt-3.5-turbo-0125
OpenAI	gpt-4-turbo-2024-04-09
OpenAI	gpt-4o-2024-05-13
Anthropic	claude-3-haiku-20240307
Anthropic	claude-3-sonnet-20240229
Anthropic	claude-3-opus-20240229
Mistral AI	Mixtral-8x7B-Instruct-v0.1
Aleph Alpha	luminous-base-control-20240215
Aleph Alpha	luminous-extended-control-20240215
Aleph Alpha	luminous-supreme-control-20240215

The selected models provide a comprehensive overview of the current capabilities in the field. This analysis highlights both the strengths and the limitations of the included models, offering valuable insights into their comparative performance.

A. Prompt engineering

Prompt engineering has become some sort of discipline of its own [21]. Proper formulation of the task the LLM should perform is important to yield the desired results.

The prompt was used to describe the scenario and the role the LLM was expected to take. The prompt was therefore separated into a first part with the instruction on how to answer and a second part with the actual question. Such structuring is considered to be good practice and generally improves the outcome. Improved results could be expected, if instead of this “zero-shot learning” approach, at least a single example of the answer to be generated would be presented to the LLM (“few-shot learning”) as part of the input prompt [1].

The following paragraph represents the input prompt used:

You are a helpful assistant and answer questions about the real world. Interpret each question literally and research each answer carefully without falling for common myths. Only give the short answer "yes" or "no" if you are sure that this is the correct answer. If you are sure that the answer to the question is both "yes" and "no", give the short answer "yes and no". If you cannot answer the question

with certainty, the short answer will be "no comment". Give a confidence score and justify your answer (rationale). Always answer in JSON format (without using Markdown code blocks or other formatting). The JSON schema should contain:

```
{"short_answer": string,
  "confidence_score": number (0-1),
  "rationale": string}
```

Question: ...

The prompt demands the LLM to output JSON data in order to easily evaluate its answers and to compare them with the expert's answer in the SQLite database. However, the "weaker" models did not follow this instruction: Aleph Alpha's "base" model only responded in plain text ("yes" or "no"), omitting the rationale in most cases and the confidence score for all queries. The "extended" and "supreme" models did output JSON in the majority of cases, but with erroneous string formatting (missing quotation marks). For a number of queries, the answer was plain text in case of the "extended" model. The Mixtral-8x7B-Instruct model was given the prompt above without JSON part ("Always answer in JSON format..." omitted), since it ignored this part anyway. Moreover, the model left out the rationale in many cases or it was not useful (e.g., containing only repetitions of the short answer) and the confidence score was always 1.0.

After all, the three OpenAI models and the three models of Anthropic did in fact respond accordingly, using the JSON format perfectly in case of OpenAI. Their models are advertised to be able to output JSON compatible responses, if an additional parameter is used in the query (`response_format={"type": "json_object"}`), so this behavior was expected. The Claude 3 family does not provide such a parameter, but the output was indeed in JSON format. The only flaw was the missing escape sequence `\` for quotation marks inside of the strings representing the rationale. They had to be escaped afterwards to yield proper JSON.

As pointed out before (see Section II-C), the explanation of the LLM as demanded in step 2 of the prompt was not used in the context of the present paper. However, instead of discarding it, it could be incorporated into the weighting scheme (points) in Table III, serving as justification for awarding the respective points in each and every actual case and to differentiate in the scheme even further.

B. OpenAI's GPT-Series

OpenAI is generally regarded as one of the leading companies in the field of generative AI and is known for its GPT series of LLMs. Figure 3 shows the results for two runs each with GPT-3.5-Turbo, GPT-4-Turbo and the newest GPT-4o model. The difference between the two runs serves as an indicator for the variability in the rating achieved, although a multitude of runs should be performed to get real statistics.

This was not possible due to budgetary limitations. However, as can be seen from the two runs, the rating varies slightly. It should be noted that the input to the models was exactly the same for the two runs, including the parameters used in the query. OpenAI introduced a seed parameter that can be used to produce reproducible output in the future. According to the documentation, this feature cannot be used reliably as of now.

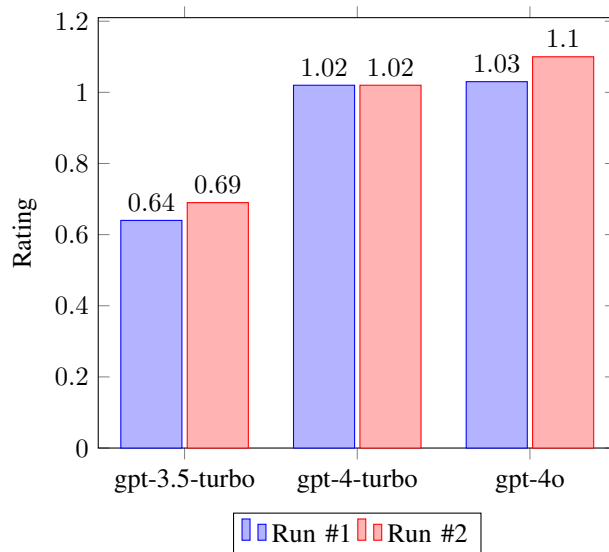


Figure 3. Results for OpenAI GPT-Series.

The results are impressive in terms of the model's capability to debunk common myths and can be classified as "superhuman level" or "expert level" in case of GPT-4-Turbo and GPT-4o. Both achieve a rating of +, provided that each rationale can be accepted for all the correct short answers given. This can only be certified eventually in a time consuming manual process by comparing each rationale with the corresponding explanation of the experts in the DIE ZEIT database. For all divergent reasoning, the short answer should be downgraded to a certain degree, which is yet to be determined.

C. Anthropic's Claude 3

Anthropic AI announced the "Claude 3" model family in March 2024 [22]. The rating results in Figure 4 for two different runs suggest that the reproducibility is quite good, with the best model Claude-3-Opus beating OpenAI's "frontier model" GPT-4o. The improvements from the cheapest (in terms of costs per query) model to the most expensive are significant and coincide with the advertised curve in performance.

D. Comparison

In this section, we present a comparative analysis of the ratings for LLMs from various vendors, expanding upon the vendor-specific results discussed previously. Figure 5 shows the best case results (for those with two runs) of all LLMs tested in this survey. For each vendor except Mistral's three sizes of models have been studied, with "base model" being the smallest (and cheapest) and "frontier model" being the

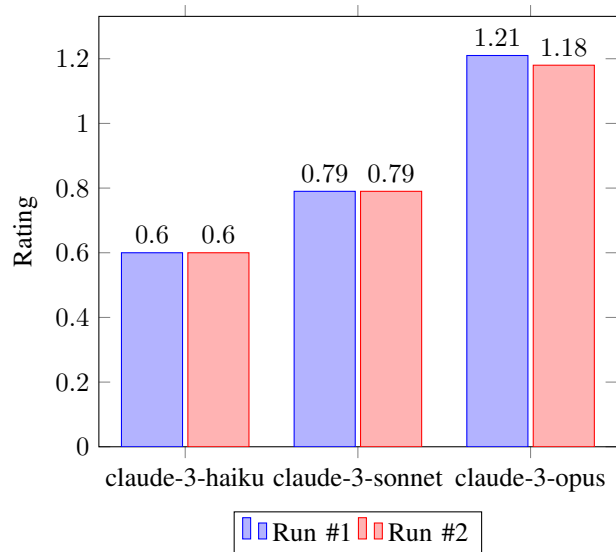


Figure 4. Results for Anthropic’s Claude 3.

most advanced (most expensive). “Standard model” denotes the established model. This categorization is not based on a consensus between vendors, but serves as a descriptive means in the context of this paper. For all ratings above the red line indicated by +, one can attest better than average human performance, with “human” representing the ordinary people in the street. Such LLMs can be classified as useful in the sense that they in part reach an expert’s level, surpassing normal persons on average. The expert in this context is not all knowing, but better in certain fields of expertise than a layperson who tends to fall for common myths or believes in the public opinion in lack of better knowledge. The red lines may imply a sharp threshold, but it should rather be interpreted as a threshold range.

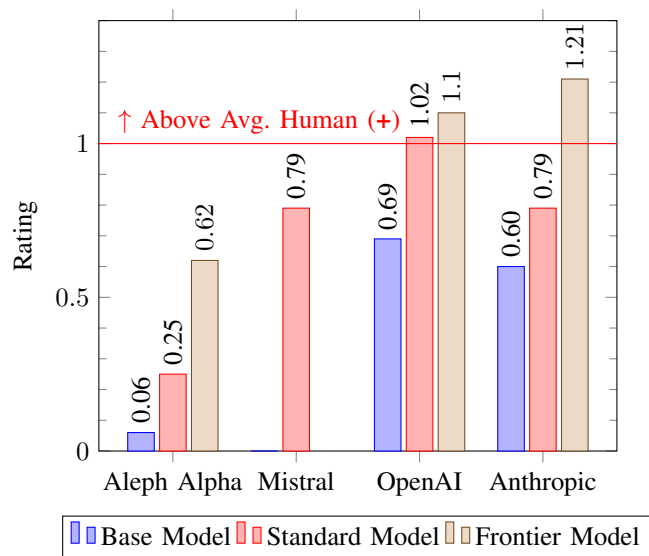


Figure 5. Comparison of the best case rating for all tested models.

The analysis in Figure 5 in underscores the potential of advanced LLMs to perform at or above human expert levels in specific domains, highlighting their practical applications and limitations.

E. Confidence Score

With the exception of the “Mixtral-8x7B-Instruct” and the “luminous-base-control” models, the LLMs responded with a confidence score, besides answering the question itself. This was demanded in the prompt, even though it can be considered redundant with respect to the phrase “...if you are sure...” as a prerequisite for giving one of the three distinct answers “yes”, “yes and no” or “no”. If unsure, the LLMs were instructed to output “no comment”. For this reason one would expect the LLMs to only return a confidence score of 1.0 (for 100%) in case of a distinct answer and a lower confidence score if the answer is “no comment”. However, the interpretation of the confidence score must be different: Analysis shows that the LLMs also gave short answers other than “no comment” for much lower confidence scores. Most of the distinct short answers were associated with a confidence score well above the 70% level, but a few were between 50% and 70% and a single one below 50%: When the model GPT-4o was run with a “temperature” higher than the obvious value of 0 (the most focused and deterministic setting), the model was more confident about its truthfulness, in spite of a low confidence score. In this run the parameter “temperature” was set to 1.0 leading to more randomness in the output as OpenAI’s documentation puts it. GPT-4o answered “yes” in this single case, with a confidence score as low as 30%, which clearly contradicts the instruction in the prompt. This may be regarded as singular fault or runaway value, owing to the higher temperature setting.

Figure 6 gives an impression of the distribution of the confidence scores for the best case runs of all models which returned a confidence score. The granularity of the score was always constricted by the LLMs to the values given in the legend of the figure, i.e., steps of 5% to differentiate. Scores of 98% and 99% were only given by the two leading edge LLMs GPT-4-Turbo and Claude-3-Opus. The other models responded with the coarser graduation of 5%.

The plot shows no clear pattern, except for increasing confidence for larger models within a family of models: Claude-3 associates a higher number of answers with a confidence score of 90% and 95%, when moving from the base model “Haiku” to the next higher model “Sonnet”, and then to the most advanced model “Opus”. For the GPT family this is not true, since GPT-3.5-Turbo outputs most answers with a confidence score of 80% and above, whereas GPT-4-Turbo and GPT-4o have a significant amount of answers with confidence score of 70% and 75% (some even below).

When taking into account the varying levels of confidence in Figure 6 and their associated answers, the question arises: is the LLM capable of correctly distinguishing between “sure” and “not sure” as demanded by the prompt (see Section III-A)?

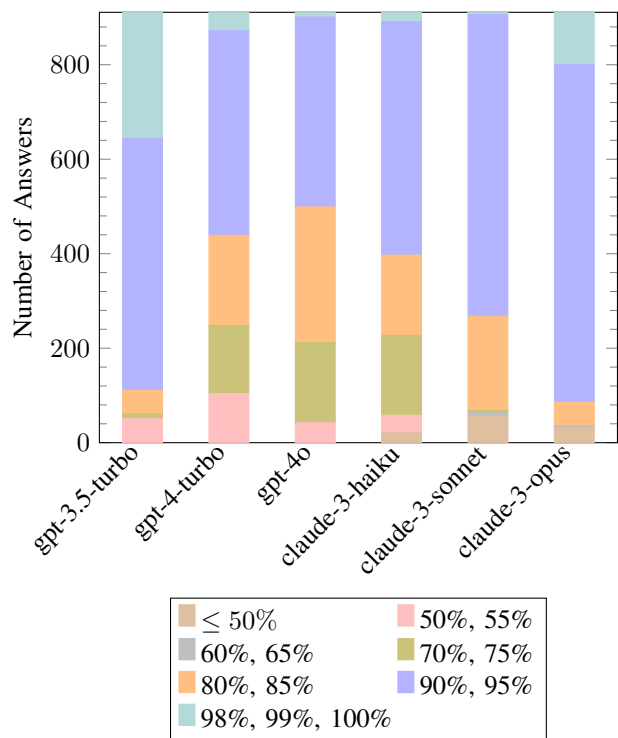


Figure 6. Distribution of confidence scores.

In this context “correctly” means truthful and based on facts and figures underlying the LLM’s training data.

Figure 7 shows the declining rating of the LLMs when plotting the rating against the confidence score as reported by the LLM. When the LLM gave an answer “yes”, “yes and no” or “no” with a confidence score below the given value on the horizontal axis, the answer was interpreted as “no comment”. This way the bar is raised step by step and the scores on the rightmost side of the plot represent the most rigorous situation. With such a high expectation regarding confidence, the score drops significantly for all models, reaching a negative level for the second best model of Aleph Alpha (“luminous-extended-control”). The overall conclusion to be drawn from this plot is that taking the confidence score into account does not improve the quality of the answers and thus the rating or vice versa.

IV. CONCLUSION AND FUTURE WORK

The following section examines the steps that need to be taken to advance the concept presented and summarizes the findings of this study.

A. Next Steps

One of the most obvious steps to be taken next is a comprehensive evaluation of all the other major LLMs like Meta’s Llama 3, Google’s Gemini or Grok of xAI on the basis of the rating scheme presented in this paper (provided they have a command of German). Currently exist 28 publicly available and just as many closed source models, having a size larger than 10B [23]. Besides these well-known models, specialized and optimized versions also seem worthwhile,

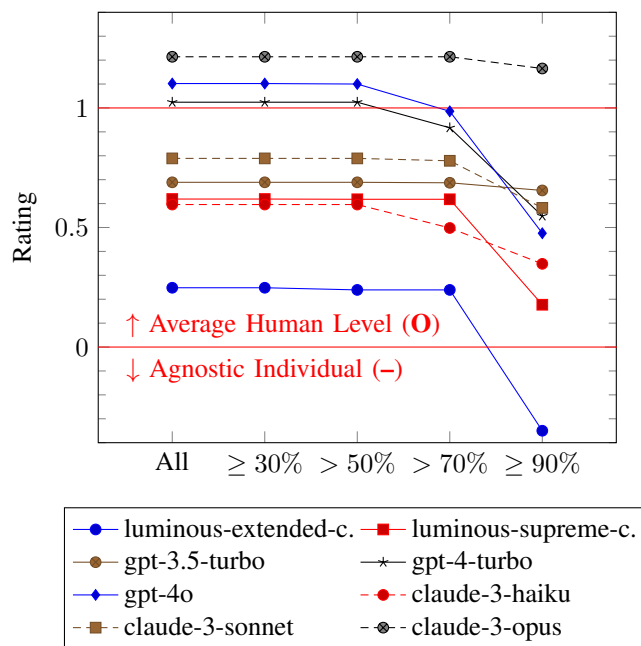


Figure 7. Rating versus confidence score.

especially the ones with a Mixture-of-Experts (MoE) architecture. This approach might yield better results if the “experts” are mixed in such a way that it resembles the combination of those experts that were consulted by Christoph Drösser, the author of the ZEIT rubric. His approach also represents a “mixture of experts” in a very literal sense.

Another field of study is the influence of the prompting on the results. The LLMs were instructed to take the role of an assistant for the present survey. Do the results get better if the LLMs are given the role of an expert instead? Or, on the other hand, do they even get worse, because in media, experts are always self-confident and mostly have a distinct opinion, whereas the answer “no comment” is very seldom. Real experts are usually asked for their opinion if it is assumed that they actually have something valuable to say and this might introduce a bias in the training data of the LLMs.

The results might also benefit from techniques like Chain-of-Thought (COT) prompting. One attempt in this way could be to ask for the reasoning first, and then afterwards in a second step to ask for the short answer. A modification of the COT-technique has been published in [24] and was titled “Chain-of-Verification Reduces Hallucination in Large Language Models”. This approach would be worthwhile to investigate.

The concept of “open-book” questioning means that the AI does not only generate answers from its training dataset in the primordial manner of LLMs, but is also capable of looking up answers on the internet or from various other publicly available sources [25] [26] [27]. How and where this is done can either be left to the model or be directed by a human instructor. If it is the model solely, a beneficial strategy in doing this can be interpreted as another type of intelligent task, broadening

our understanding of today’s AI capabilities significantly. The taxonomy presented in this paper can help to evaluate the chances of success of such an undertaking.

B. Summary

This paper proposes a new classification scheme for evaluating the trustworthiness and usefulness of Large Language Models (LLMs) in supporting everyday tasks, particularly in the context of fact-checking and combating misinformation. We argue that existing benchmarks and metrics are insufficient and often flawed, failing to keep pace with the rapid development of LLMs.

The proposed methodology involves using a questionnaire based on a dataset of questions about widespread everyday wisdom, urban legends, and misconceptions, sourced from the German weekly newspaper DIE ZEIT “Stimmt’s” section. The questions are formulated in a “Is it right that...” format, allowing for short answers of “Yes”, “No” or “Yes and No.” We manually selected 1,000 questions from a total of 1,276, excluding those that were country-specific, potentially offensive, or not suitable for the proposed format. The LLMs’ responses to these questions are then compared to the expert-verified answers from the ZEIT dataset, and a point-based weighting scheme is applied to rate the LLMs’ performance. The scheme assigns points ranging from -4 to +4 based on the agreement or disagreement between the LLMs’ answers and the expert-verified answers, considering factors such as the LLMs’ ability to identify unknown or uncertain answers and their tendency to confabulate or reproduce common misconceptions.

We tested several well-known LLMs, including OpenAI’s GPT series, Anthropic’s Claude 3, and others, comparing their performance using the proposed rating system. The results suggest that some LLMs, such as GPT-4-Turbo, GPT-4o, and Claude-3-Opus, achieve “superhuman” or “expert” level performance in debunking common myths. However, the author notes that a more thorough manual comparison of the LLMs’ reasoning with the experts’ explanations is necessary to fully validate these findings. The paper also examines the confidence scores provided by the LLMs and concludes that these scores do not necessarily improve the quality of the answers or the overall rating of the LLMs’ trustworthiness.

ACKNOWLEDGEMENTS

We would like to thank the referees for very useful comments on the original submission. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

- [1] T. B. Brown *et al.*, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf
- [2] OpenAI *et al.*, “Gpt-4 technical report,” 2024.
- [3] A. Chowdhery *et al.*, “Palm: scaling language modeling with pathways,” *J. Mach. Learn. Res.*, vol. 24, no. 1, mar 2024.
- [4] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On faithfulness and factuality in abstractive summarization,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 1906–1919. [Online]. Available: <https://aclanthology.org/2020.acl-main.173>
- [5] Z. Ji *et al.*, “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, p. 1–38, Mar. 2023. [Online]. Available: <http://dx.doi.org/10.1145/3571730>
- [6] M. Moscovitch, “Confabulation and the frontal systems: Strategic versus associative retrieval in neuropsychological theories of memory,” in *Varieties of memory and consciousness: Essays in honour of Endel Tulving*, H. L. I. Roediger and F. I. M. Craik, Eds. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1989, pp. 133–160.
- [7] G. D. Barba, “Confabulation: Knowledge and recollective experience,” *Cognitive Neuropsychology*, vol. 10, no. 1, pp. 1–20, 1993. [Online]. Available: <https://doi.org/10.1080/02643299308253454>
- [8] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, “HellaSwag: Can a machine really finish your sentence?” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Márquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4791–4800. [Online]. Available: <https://aclanthology.org/P19-1472>
- [9] A. Srivastava *et al.*, “Beyond the imitation game: Quantifying and extrapolating the capabilities of language models,” *Transactions on Machine Learning Research*, 2023. [Online]. Available: <https://openreview.net/forum?id=uyTL5Bvosj>
- [10] L. Ruis, J. Andreas, M. Baroni, D. Bouchacourt, and B. M. Lake, “A benchmark for systematic generalization in grounded language understanding,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 19861–19872. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/e5a90182cc81e12ab5e72d66e0b46fe3-Paper.pdf
- [11] Edwin Chen, “Hellaswag or hellabad? 36% of this popular llm benchmark contains errors,” 2022, [retrieved: May 2024]. [Online]. Available: <https://www.surgehq.ai/blog/hellaswag-or-hellabad-36-of-this-popular-llm-benchmark-contains-errors>
- [12] E. Davis, “Benchmarks for automated commonsense reasoning: A survey,” *ACM Comput. Surv.*, vol. 56, no. 4, oct 2023. [Online]. Available: <https://doi.org/10.1145/3615355>
- [13] S. Gehrmann *et al.*, “The GEM benchmark: Natural language generation, its evaluation and metrics,” in *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 96–120. [Online]. Available: <https://aclanthology.org/2021.gem-1.10>
- [14] Z.-Y. Dou, P. Liu, H. Hayashi, Z. Jiang, and G. Neubig, “GSum: A general framework for guided neural abstractive summarization,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 4830–4842. [Online]. Available: <https://aclanthology.org/2021.naacl-main.384>
- [15] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” *Science*, vol. 359, pp. 1146–1151, 03 2018.
- [16] D. Lazer *et al.*, “The science of fake news,” *Science*, vol. 359, pp. 1094–1096, 03 2018.
- [17] Gil Elbaz and Peter Norvig and Nova Spivack and Carl Malamud and Kurt Bollacker and Joi Ito, “Common crawl — open repository of web crawl data,” 2024, [retrieved: May 2024]. [Online]. Available: <https://commoncrawl.org/>
- [18] L. Fridman, “#416 – yann lecun: Meta ai, open source, limits of llms, agi & the future of ai,” Podcast, 2024, retrieved: May 2024. [Online]. Available: <https://lexfridman.com/yann-lecun-3/>
- [19] S. Lin, J. Hilton, and O. Evans, “TruthfulQA: Measuring how models mimic human falsehoods,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3214–3252. [Online]. Available: <https://aclanthology.org/2022.acl-long.229>

- [20] Zyte and contributors, “Scrapy — a fast and powerful scraping and web crawling framework,” 2024, [retrieved: May 2024]. [Online]. Available: <https://scrapy.org/>
- [21] S. Diao, P. Wang, Y. Lin, and T. Zhang, “Active prompting with chain-of-thought for large language models,” 2023. [Online]. Available: <https://arxiv.org/abs/2302.12246>
- [22] Anthropic, “The claude 3 model family: Opus, sonnet, haiku,” 2024, [retrieved: May 2024]. [Online]. Available: <https://www.anthropic.com/news/claude-3-family>
- [23] W. X. Zhao *et al.*, “A survey of large language models,” *arXiv preprint arXiv:2303.18223*, 2023. [Online]. Available: <http://arxiv.org/abs/2303.18223>
- [24] S. Dhuliawala *et al.*, “Chain-of-verification reduces hallucination in large language models,” 2023. [Online]. Available: <https://openreview.net/forum?id=VP20ZB6DHL>
- [25] D. Chen, A. Fisch, J. Weston, and A. Bordes, “Reading Wikipedia to answer open-domain questions,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, R. Barzilay and M.-Y. Kan, Eds. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1870–1879. [Online]. Available: <https://aclanthology.org/P17-1171>
- [26] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, “Can a suit of armor conduct electricity? a new dataset for open book question answering,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2381–2391. [Online]. Available: <https://aclanthology.org/D18-1260>
- [27] G. Kokaia, P. Sinha, Y. Jiang, and N. Boujemaa, “Writing your own book: A method for going from closed to open book qa to improve robustness and performance of smaller llms,” 2023.