Large Language Models as Evaluators for Scientific Synthesis

Julia Evans, Jennifer D'Souza, and **Sören Auer**
TIB - Leibniz Information Centre for Science and Technology,
Hannover, Germany
Correspondence: jennifer.dsouza@tib.eu

Abstract

Our study explores how well the state-of-the-art Large Language Models (LLMs), like GPT-4 and Mistral, can assess the quality of scientific summaries or, more fittingly, scientific syntheses, comparing their evaluations to those of human annotators. We used a dataset of 100 research questions and their syntheses made by GPT-4 from abstracts of five related papers, checked against human quality ratings. The study evaluates both the closed-source GPT-4 and the open-source Mistral model's ability to rate these summaries and provide reasons for their judgments. Preliminary results show that LLMs can offer logical explanations that somewhat match the quality ratings, yet a deeper statistical analysis shows a weak correlation between LLM and human ratings, suggesting the potential and current limitations of LLMs in scientific synthesis evaluation.

1 Introduction

Large Language Models (LLMs) have made a significant impact on natural language processing (NLP), demonstrating exceptional performance in tasks like text generation, sentiment analysis, machine translation, and question answering, with outputs that often rival human-created content (Huang et al., 2023). In addition to their direct applications, LLMs offer substantial benefits in streamlining machine learning model development, particularly in evaluation processes. They reduce the dependency on human-generated ground truth data and the necessity for human evaluators (Bai et al., 2023) in two key ways: by facilitating the generation of synthetic ground truth data and by serving as evaluators for model predictions themselves. This approach not only speeds up the evaluation process but also broadens the scope of evaluation criteria to include factors such as diversity and coverage, enhancing the efficiency and comprehensiveness of model assessments.

This study investigates the use of LLMs as evaluators to streamline the evaluation process, moving away from traditional reliance on human evaluators and human-generated ground truth data. It specifically examines the effectiveness of LLMs in synthesizing scientific abstracts seen generally as multi-document summarization tasks. The main focus of this research is to assess how two state-of-the-art LLMs—the proprietary GPT-4 Turbo (OpenAI, 2023) and the open-source Mistral-7B (Jiang et al., 2023)—perform in evaluating scientific syntheses. Furthermore, leveraging LLMs meant better versatility in evaluation considerations, which meant that the evaluations tested varied dimensions of syntheses quality, viz. comprehensiveness, trustworthiness, and utility.

This paper is structured as follows. First, section 2 presents a review of related work in the fields of text summarization and LLM evaluation. In section 3, we show our approach to using LLMs for scientific synthesis evaluation, wherein subsection 3.1 describes the LLM output, while subsection 3.2 presents a qualitative evaluation of this output. In subsection 3.3, we analyze the correlation between LLM ratings and human judgments. A discussion of our findings and final conclusions is described in section 4.

2 Related Work

Evaluation Metrics for Text Summarization. The most common automatic evaluation metric used within summarization research – both single-document and multi-document – is the ROUGE family of metrics (Ma et al., 2022; Akter et al., 2022; Cohan and Goharian, 2016; Kryscinski et al., 2019; Lloret et al., 2018). ROUGE metrics (Lin, 2004) calculate the lexical overlap between a human-written reference document and an automatically generated one, although variants incorporating semantic information also exist. Within text summarization

research, the most commonly used are ROUGE-N and ROUGE-L (Ma et al., 2022), both of which are purely lexical-matching metrics. ROUGE-N evaluates the recall of n-grams by comparing a reference text with a corresponding machine-generated text, whereas ROUGE-L calculates the longest common subsequence of tokens shared between reference and machine-generated texts (Lin, 2004).

Despite its predominance within the field, ROUGE nonetheless has some notable limitations. First, the most commonly used metrics lack semantic awareness (Akter et al., 2022; Ma et al., 2022). Studies have pointed out that ROUGE may not accurately estimate summary quality in cases of terminological variations, paraphrasing, and differences in sentence structure (Cohan and Goharian, 2016). Moreover, there exist 192 ROUGE variants (Graham, 2015), with meaningful differences in how well each performs on a given system or specialized task (Cohan and Goharian, 2016; Graham, 2015; Kryscinski et al., 2019) and how well they correlate with human judgements (Kryscinski et al., 2019; Graham, 2015). Finally, ROUGE evaluates only content selection but not linguistic quality aspects such as grammaticality and referential clarity (Pitler et al., 2010) or overall quality, including the ordering of information and structural clarity (Graham, 2015).

Although no other metrics have gained widespread adoption, other approaches exist. Additional lexical-matching metrics include BLEU (Papineni et al., 2002) and Pyramid (Nenkova et al., 2007). Semantically enriched metrics include METEOR (Banerjee and Lavie, 2005), an expansion of BLEU, and approaches utilizing word embeddings, such as BERTScore (Zhang et al., 2020), MoverScore (Zhao et al., 2019), and SUPERT (Gao et al., 2020). However, none of these metrics address all of ROUGE's weaknesses, and the limited use of such metrics within the research community means that ROUGE remains the "de facto" standard (Lloret et al., 2018).

LLMs for Text Evaluation. Using LLMs for text evaluation is still a nascent research topic. Several recent works have compared LLMs' text evaluations to human evaluations on multiple tasks, and report that LLMs produce results similar to human judgements (Chiang and Lee, 2023b; Liu et al., 2023; Wang et al., 2023). One work finds only minor variations in results depending on task instructions and hyperparameters, whereas they find

a high degree of variation in performance of different LLMs and the quality characteristics being assessed (Chiang and Lee, 2023b). In evaluating the quality of story fragments by *grammaticality*, *cohesiveness*, *likability*, and *relevance*, they find only a weak correlation between humans and LLMs on *grammaticality*, but a moderate correlation on *relevance*. Contrarily, another work finds that ChatGPT's performance is sensitive to prompt instructions (Wang et al., 2023). They also show that ChatGPT evaluations correlate especially well with human evaluations for creative tasks like story generation (Wang et al., 2023). Another work demonstrates that requiring LLMs to provide a justification for their ratings "significantly improves the correlation between the LLMs' ratings and human ratings" (Chiang and Lee, 2023a).

Only one work has investigated the task of text summarization evaluation (Liu et al., 2023). They evaluate single-document news article summaries on the aspects of *coherence*, *consistency*, *fluency*, and *relevance*; their results exceed the correlation with human judgements of most automatic approaches, including ROUGE. In another task, ChatGPT successfully identifies implicit hate speech in Tweets and generates explanations of why the texts are hateful, which human annotators judge equally informative to human-written explanations and of greater clarity (Huang et al., 2023).

3 LLMs for the Scientific Synthesis Evaluation Task

The accurate evaluation of scientific syntheses is a critical task in research, ensuring the integrity and reliability of the synthesized information. While recent advancements have demonstrated the efficacy of LLMs in generating such syntheses (Pride et al., 2023), their potential in evaluating them remains relatively unexplored. Motivated by the limitations of existing evaluation metrics, such as the ROUGE family, and the success of LLMs in other text evaluation tasks, our work seeks to investigate the suitability of LLMs for the task of assessing the quality of scientific syntheses.

To address this question, we employ the proprietary GPT-4 Turbo (OpenAI, 2023) and the open-source Mistral-7B models (Jiang et al., 2023) to evaluate the CORE-GPT dataset (Pride et al., 2023). This dataset comprises 100 research questions spanning 20 diverse domains, each accompanied by the titles and abstracts of five related works and an an-

swer to the research question generated by GPT-4 by synthesizing the provided abstracts. Additionally, human ratings from two annotators, on a scale of 0 to 10, are available on the quality of each synthesis in three dimensions, viz. *comprehensive*, *trust*, and *utility*.

For our task, we query the LLMs to evaluate the syntheses according to the same three aspects as the CORE-GPT human raters. Our prompt follows a similar structure to previous work (Chiang and Lee, 2023a). It contains two lines of task instruction, explanation of the quality aspects (as defined for the CORE-GPT dataset annotators) and the rating scale, response format instructions, and finally the answer to be evaluated with its question and abstracts. The response is requested in JSON format, with a numeric rating between 0 and 10 for each aspect as well as a rationale for each rating. The full text of the prompt is in Appendix A.

3.1 LLM Synthesis Evaluation Output

A representative example of the evaluation output from GPT-4 Turbo and Mistral is shown in Appendix B and Appendix C, respectively. The output from GPT-4 was exactly as requested, while Mistral had some variability. In one case, Mistral returned ratings of "excellent," "good," and "high" rather than numeric scores; this output was excluded from the analysis. In several other cases, Mistral included a paragraph after the JSON object which summarized the ratings and rationales provided within it. These paragraphs were discarded and only the JSON object content was evaluated.

An overview of LLM performance was obtained by reviewing one synthesis from each domain evaluated by both GPT-4 and Mistral. Qualitatively, both models demonstrated credible and logically consistent ratings and rationales. GPT-4 provided more detailed rationales compared to Mistral, with slightly lower ratings overall.

In their rationales for *comprehensive*, both LLMs would sometimes highlight relevant topics from the abstracts which were not included in the synthesis, with GPT-4 producing such rationales more often than Mistral. Occasionally, some rationales contained justifications relating to content more specific than just the topics, suggesting more information on the results or the methodology of the studies would improve it.

The LLMs seemed to show the greatest discrepancy between rating and rationale, and the greatest inconsistencies, in their evaluations of *trust*. In one

Mistral evaluation with a rating of 5, the rationale noted that the citations only improved trustworthiness "as long as the abstract accurately represents the study's findings." In the absence of any evidence the abstract is suspect, this rating is disproportionately low. GPT-4 was notably more conservative than human annotators, as it did not give a single 10. Especially for *trust*, it was often difficult to understand why a rating wasn't higher. For instance, the rationale for one rating of 8 praised the synthesis for accuracy and avoiding unsupported claims.

For the *utility* ratings, it appears that most rationales from GPT-4 suggested additional content which could make the synthesis more useful, such as actionable information, more detailed examples, technical details of methodologies and implementation, and so on. Mistral made such suggestions less frequently; its rationales tended to echo the rationale for *comprehensive*. However, Mistral did sometimes provide guidance on who would or would not find the synthesis useful.

3.2 Qualitative Evaluations

LLMs are known to sometimes generate content on topics that lack factual basis with a highly persuasive level of linguistic proficiency (Bang et al., 2023; Liu et al., 2023). For scientific syntheses which provide an answer to a question, it is especially important that the content is genuinely a synthesis of the provided abstracts, with appropriate citations, and not independently generated based on the LLM's training data. For this reason, we were particularly interested in how the LLMs evaluated quality, and most importantly *trust*, when there was reason to believe the abstracts were not the (primary) source of the generated content, as in the following three scenarios. The complete question and answer pairs, along with their GPT-4 and Mistral evaluation scores and *trust* rationales, can be found in Appendix D.

Response Explicitly States Absence of Relevant Abstracts. In six cases, the synthesis directly expressed limitations due to the relevancy of the provided abstracts, e.g. "[...] the provided search results do not offer specific information on the long-term health impacts of such medications on these organs." Human annotators responded very positively to this, with such responses "scored highly for trustworthiness" (Pride et al., 2023). Mistral rated four of these syntheses as 10 for trust, citing factual accuracy and abstract sourcing, while two

3

scored 7. GPT-4 ratings varied, at 5, 5, 5, 7, 7, and 8. Mistral rationales did not reference the stated limitation, while GPT-4 acknowledged it positively in three cases. However, as these syntheses were scored 8, 7, and 5, it is unclear to what extent this acknowledgement may have influenced the scores.

Response Contains No Citations. There were three responses which answered the question but contained no citations. GPT-4 gave *trust* scores of 0, 0, and 1, with rationales referring to the lack of citations. In contrast, Mistral scored 8, 10, and 10, with rationales stating the information was common knowledge or referenced from the abstracts.

Response Contains One Citation. Finally, there were five syntheses which cited only one of the abstracts, which does not align with the task of synthesizing multiple abstracts to provide an answer to the given question. For GPT-4, the *trust* scores were 5, 7, 8, 8, and 9, with most rationales stating that the synthesis relied on general knowledge without directly referencing the abstracts, despite one citation being present in each case. Meanwhile, the Mistral scores were 7, 9, 9, 10, and 10, with most rationales indistinguishable from those of syntheses with many more citations - three of them claimed that the synthesis accurately references the content in the provided abstracts.

	A1	A2	GPT-4	Mistral
A1				
ρ	-	**0.710**	0.248	0.015
p-value	-	**0.001**	0.305	0.951
A2				
ρ	**0.710**	-	0.058	-0.038
p-value	**0.001**	-	0.814	0.878
GPT-4				
ρ	0.248	0.058	-	**0.786**
p-value	0.305	0.814	-	**0.000**
Mistral				
ρ	0.015	-0.038	**0.786**	-
p-value	0.951	0.878	**0.000**	-

Table 1: Spearman's ρ calculated for the combined mean of *Comprehensive*, *Trust*, and *Utility* scores. Statistically significant results are in bold.

3.3 Correlation

Spearman's ρ was calculated to assess the relationship between the human annotators' scores and the LLM-generated scores. Using the publicly-available data from CORE-GPT (Pride et al., 2023)[1], separate vectors for each annotator were obtained. To calculate the correlations, we found the overall mean score for each domain; due to the format of the published data, it was not possible to match individual scores to their corresponding syntheses. Our results for the overall mean are presented in Table 1.

We find that only two results showed statistically significant p-values. Human annotators exhibited a strong positive correlation (0.710), as did GPT-4 Turbo and Mistral (0.786). However, correlations between annotators and LLMs were weak or very weak, with p-values indicating insufficient evidence for genuine association. These findings suggest LLMs cannot directly replicate human performance in evaluating scientific syntheses. Despite this, the strong positive correlation between GPT-4 Turbo and Mistral indicates consistency between the two LLMs.

4 Discussion and Conclusion

We explore the capacity of LLMs in assessing scientific syntheses. GPT-4 Turbo and Mistral are utilized to obtain quality ratings for 100 syntheses from the CORE-GPT dataset (Pride et al., 2023), accompanied by a rationale for each rating. Correlation analysis using Spearman's ρ indicates that the LLM performance does not align with the human annotators' judgements. However, a qualitative evaluation of the responses finds a more mixed result.

Both LLMs generally produce credible and logically consistent ratings and rationales, but GPT-4 appears more conservative in its ratings and provides more detail and specific recommendations in its rationales. GPT-4 also displays greater sensitivity to the presence or absence of citations compared to Mistral. However, both LLMs' rationales occasionally contained inaccuracies or flaws, raising concerns about the credibility of their scores. Moreover, the extent to which the responses are evaluated as *syntheses* and not simply as *answers*, without reliance on general knowledge, remains unclear, particularly in the case of Mistral.

Our findings highlight both promising developments and current limitations of leveraging LLMs for the task of evaluating scientific syntheses, illustrating the need for further research to validate and refine the methodology.

Limitations

We acknowledge several limitations that may influence the interpretation and generalizability of our findings. First, the reliance on a single, relatively small dataset presents limitations in terms of data representativeness. Moreover, the data format necessitated aggregating scores, which may have obscured potential nuances in individual annotations.

Second, the study focused exclusively on GPT-4 Turbo and Mistral, limiting the generalizability of our conclusions to other LLMs. While these models represent the state-of-the-art, future iterations or alternative architectures may exhibit different performance. Additionally, we were able to obtain only one set of ratings from each LLM. Given the variability of LLM output, taking the average of several runs is preferable, but due to financial limitations, this was not possible in our study.

We note that past work has found LLMs particularly adept at evaluating creative texts (Wang et al., 2023), so the narrow output scope of synthesis for scientific question answering may pose a greater challenge. We also note the difficulty of assessing the quality of syntheses from such a diverse assortment of domains. Judging how comprehensive a synthesis is requires some knowledge of the scope of potential information which might be appropriate to include. Highly specialized domain knowledge still presents a challenge to general use LLMs.

Ethical Considerations

In this work we have presented our study of the efficacy of two LLMs, one proprietary and one open-source, in evaluating the quality of scientific syntheses. There were no living subjects analyzed in this study. Overall, this study complies with the ACL Ethics Policy.

In querying the LLMs for synthesis quality evaluations, we declare that the instructions were intended to align the behavior of the language models towards producing responses that are both helpful (fulfilling our objective) and harmless (not causing any physical, psychological, or social harm to individuals or the environment). All of the intellectual property which was passed to the LLMs is open-access.

References

Mousumi Akter, Naman Bansal, and Shubhra Kanti Karmaker. 2022. Revisiting Automatic Evaluation of Extractive Summarization Task: Can We Do Better than ROUGE? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1547–1560, Dublin, Ireland. Association for Computational Linguistics.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023. Benchmarking Foundation Models with Language-Model-as-an-Examiner. In *Advances in Neural Information Processing Systems*, volume 36, pages 78142–78167. Curran Associates, Inc.

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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023a. A Closer Look into Using Large Language Models for Automatic Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023b. Can Large Language Models Be an Alternative to Human Evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Arman Cohan and Nazli Goharian. 2016. Revisiting Summarization Evaluation for Scientific Articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 806–813, Portorož, Slovenia. European Language Resources Association (ELRA).

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Yvette Graham. 2015. Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal. Association for Computational Linguistics.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 294–297, New York, NY, USA. Association for Computing Machinery.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural Text Summarization: A Critical Evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Elena Lloret, Laura Plaza, and Ahmet Aker. 2018. The challenging task of summary evaluation: an overview. *Language Resources and Evaluation*, 52:101–148.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2022. Multi-Document Summarization via Deep Learning Techniques: A Survey. *ACM Computing Surveys*, 55(5).

Ani Nenkova, Rebecca Passonneau, and Kathleen Mckeown. 2007. The Pyramid Method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, 4(2).

OpenAI. 2023. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic Evaluation of Linguistic Quality in Multi-Document Summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 544–554, Uppsala, Sweden. Association for Computational Linguistics.

David Pride, Matteo Cancellieri, and Petr Knoth. 2023. CORE-GPT: Combining Open Access Research and Large Language Models for Credible, Trustworthy Question Answering. In *Linking Theory and Practice of Digital Libraries*, pages 146–159. Springer Nature Switzerland.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations (ICLR'20)*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Prompt

A.1 Main Evaluation Prompt

Evaluate the quality of the following question and answer pair. The answer should succinctly address the question while referencing only content from the provided abstracts.

Consider the following characteristics while assessing the answer. For each characteristic rate the quality from 0 (poor) to 10 (excellent). Provide a short rationale for each rating.

1. Comprehensive: How comprehensively is the question answered?

2. Trust: How trustworthy is the answer?

3. Utility: How useful is the answer?

Return your response in JSON format:
{Comprehensive : {'rating' : '', 'rationale' : ''}, Trust : {'rating' : '', 'rationale' : ''}, Utility : {'rating' : '', 'rationale' : ''}}

Question: <content>

Answer: <content>

Abstracts: <content>

A.2 Additional Mistral System Prompt

Follow the user instructions to the letter. The output should be only what the user requests, do not respond with anything else. Provide the answer in JSON with no additional text.

B Example GPT-4 Turbo Output

COMPREHENSIVE: rating: 7, rationale: The answer provides a broad overview of how living in a low-income neighborhood can affect social networks and well-being, referencing various studies that discuss mental health, social inclusion, and subjective well-being. However, it does not fully elaborate on the specific mechanisms or provide detailed insights from each abstract, such as the role of men in community well-being from abstract 3 or the impact of natural capital from abstract 4.

TRUST: rating: 8, rationale: The answer is trustworthy as it directly references findings from the provided abstracts, indicating a reliance on research-based evidence. However, the answer could be more trustworthy if it included more specific data or findings from the abstracts to support the claims.

UTILITY: rating: 7, rationale: The answer is useful as it gives an overview of the potential impacts of living in a low-income neighborhood on residents' social networks and well-being. It could be more useful if it provided actionable insights or recommendations based on the findings of the studies.

Figure 1: Example output from GPT-4 Turbo.

C Example Mistral Output

COMPREHENSIVE: rating: 9, rationale: The answer provides specific examples and references to studies that directly address the question of how social media influences the formation and maintenance of relationships among young adults in the United States, specifically focusing on LGBTQ+ individuals with disabilities and young adults who have moved abroad.

TRUST: rating: 10, rationale: The answer is based on empirical research and provides clear citations to support the claims made.

UTILITY: rating: 10, rationale: The answer is directly relevant to the question and provides useful information for those interested in understanding how social media impacts relationship formation and maintenance among young adults in the United States.

The given question and answer pair is of high quality. The question is clear, specific, and comprehensive, while the answer is well-researched, trustworthy, and useful. It directly addresses the question by providing examples from empirical studies that explore how social media influences the formation and maintenance of relationships among young adults in the United States, specifically focusing on LGBTQ+ individuals with disabilities and young adults who have moved abroad. The answer is also trustworthy as it is based on reliable sources and provides clear citations to support the claims made. Finally, the answer is useful as it provides valuable insights into how social media impacts relationship formation and maintenance among a specific population in the United States.

Figure 2: Example output from Mistral. The text under the separator line was returned outside of the JSON response.

D *Trust* Evaluation

D.1 Response Explicitly States Absence of Relevant Abstracts

How do LLMs evaluate when the response states that a complete answer cannot be given, due to lack of relevant abstracts?

D.1.1 Chemistry

Question: What is the effect of pH on the efficiency of a common catalyst used in the production of biodiesel?

Answer: The effect of pH on the efficiency of a catalyst used in biodiesel production can be understood through the specific catalyst being used. For instance, mixed oxide catalysts like magnesium oxide (MgO) and calcium oxide (CaO) are used to promote the transesterification of vegetable oils into biodiesel. The basic property of these catalysts plays a crucial role in the biodiesel yield during the reaction [3]. However, the provided search results do not offer a direct relationship between pH and the efficiency of such catalysts in biodiesel production. Further research is required to establish a clear connection between pH levels and the performance of common catalysts used in biodiesel production.

	Comp.	Trust	Utility	Trust Rationale
GPT-4	7	8	6	The answer is trustworthy as it accurately reflects the information available in the provided abstracts. It does not make unsupported claims and is honest about the limitations of the information provided in the abstracts regarding the direct effect of pH on catalyst efficiency.
Mistral	5	7	6	The answer is based on accurate information from the abstracts, specifically the third abstract which discusses the effect of basicity (pH) on the efficiency of magnesium oxide and calcium oxide catalysts in biodiesel production.

Table 2: Ratings for the answer to the question "What is the effect of pH on the efficiency of a common catalyst used in the production of biodiesel?"

D.1.2 Chemistry

Question: What is the effect of varying concentrations of hydrogen peroxide on the rate of catalase enzyme activity?

Answer: Varying concentrations of hydrogen peroxide (H_2O_2) can affect the rate of catalase enzyme activity. Catalase is an enzyme that breaks down H_2O_2 into oxygen and water, neutralizing oxidative stress in cells. This process is crucial for maintaining appropriate levels of oxidative stress because reactive oxygen species, such as H_2O_2, can cause cellular damage by targeting DNA and proteins, which can lead to complications like cancer, diabetes, neurodegenerative diseases, and aging [2]. While the given search results do not provide exact information on how different concentrations of H_2O_2 impact the rate of catalase enzyme activity, it is evident that altering H_2O_2 levels will directly influence catalase's efficiency in breaking down H_2O_2 and neutralizing oxidative stress in cells [2].

	Comp.	Trust	Utility	Trust Rationale
GPT-4	3	5	3	The answer is somewhat trustworthy as it correctly describes the function of catalase and the consequences of oxidative stress, which is supported by the abstracts. However, it lacks direct evidence from the abstracts regarding the specific impact of different H_2O_2 concentrations on catalase activity, which reduces the trustworthiness of the response in the context of the question asked.
Mistral	5	10	8	The answer is trustworthy as it is directly derived from the provided abstracts and does not contain any incorrect or misleading information.

Table 3: Ratings for the answer to the question "What is the effect of varying concentrations of hydrogen peroxide on the rate of catalase enzyme activity?"

D.1.3 Economics

Question: What is the impact of minimum wage laws on employment rates and income inequality in the United States?

Answer: The impact of minimum wage laws on employment rates and income inequality in the United States is not directly addressed in the provided search results. However, one study suggests that an effective mechanism for managing labor productivity and wage issues is necessary for achieving greater economic equality and growth [1]. Moreover, the study highlights the lack of clear definitions of salary, salary payment, and minimum wage, which can contribute to challenges in addressing income inequality [1]. While the provided search results do not specifically discuss the impact of minimum wage laws on employment rates and income inequality in the United States, it is clear that issues related to wage management and labor productivity remain essential factors in ensuring fair compensation for workers and reducing inequality [1].

	Comp.	Trust	Utility	Trust Rationale
GPT-4	2	5	2	The answer is somewhat trustworthy as it references a study that discusses labor productivity and wage issues, but it does not provide information from a study that directly addresses the impact of minimum wage laws on employment rates and income inequality in the United States.
Mistral	0	10	2	The answer is based on content from one of the provided abstracts, which is cited and referenced appropriately.

Table 4: Ratings for the answer to the question "What is the impact of minimum wage laws on employment rates and income inequality in the United States?"

D.1.4 Mathematics

Question: What is the relationship between advanced mathematics coursework and problem-solving skills among undergraduate students in STEM majors?

Answer: The relationship between advanced mathematics coursework and problem-solving skills among undergraduate students in STEM majors is not explicitly discussed in the provided search results. However, the study on STEM in General Education [1] indicates that student mathematics skills were analyzed to determine their influence on course selection. Another study discusses the importance of engaging students in mathematical and scientific practices according to mathematics and science standards, such as the Common Core State Standards of Mathematics (CCSSM) and the Next Generation Science Standards (NGSS) [4]. This suggests that exposure to advanced mathematical coursework could potentially enhance problem-solving skills in undergraduate STEM students, though further research is needed to establish a concrete relationship.

	Comp.	**Trust**	**Utility**	**Trust Rationale**
GPT-4	5	7	4	The answer is trustworthy in the sense that it accurately reflects the lack of explicit discussion in the provided abstracts. It cautiously suggests a potential link between advanced mathematics coursework and problem-solving skills, which is a reasonable inference based on the emphasis on standards and practices in abstract [4].
Mistral	4	7	5	The answer is grounded in the content of the provided search results and cites relevant studies to support its suggestions. However, it also includes speculation about the potential relationship between advanced mathematics coursework and problem-solving skills based on indirect evidence.

Table 5: Ratings for the answer to the question "What is the relationship between advanced mathematics coursework and problem-solving skills among undergraduate students in STEM majors?"

D.1.5 Medicine

Question: What are the potential long-term health impacts of regular use of over-the-counter pain medications on the liver and kidney function in young adults?

Answer: Regular use of over-the-counter (OTC) pain medications can potentially impact liver and kidney function in young adults. However, the provided search results do not offer specific information on the long-term health impacts of such medications on these organs. To obtain a comprehensive answer, further research on this topic would be necessary.

	Comp.	Trust	Utility	Trust Rationale
GPT-4	1	5	2	The answer is trustworthy in the sense that it does not fabricate information and admits the absence of specific details in the provided abstracts. However, it does not demonstrate a thorough examination of the abstracts to determine if any related information could be inferred.
Mistral	0	10	2	The answer is truthful and does not contain any false or misleading information.

Table 6: Ratings for the answer to the question "What are the potential long-term health impacts of regular use of over-the-counter pain medications on the liver and kidney function in young adults?"

D.1.6 Physics

Question: How does the Doppler effect affect the observed frequency of sound waves emitted from a moving source and its relevance in real-world applications?

Answer: The Doppler effect affects the observed frequency of sound waves emitted from a moving source by causing a change in frequency and wavelength due to the relative motion between the source and the observer. When the source is moving towards the observer, the observed frequency increases, and the wavelength decreases; conversely, when the source is moving away from the observer, the observed frequency decreases, and the wavelength increases. Although the provided search results do not directly give real-world applications, the Doppler effect is relevant in various fields such as astronomy, radar technology, medical imaging, and seismology. For instance, it helps astronomers measure the speed of distant celestial objects and plays a critical role in Doppler radar systems for weather monitoring and forecasting.

	Comp.	Trust	Utility	Trust Rationale
GPT-4	5	7	6	The answer is generally trustworthy as it correctly describes the Doppler effect's impact on frequency and wavelength. However, it lacks specific references to the provided abstracts, which could have strengthened the trustworthiness by showing direct connections to the source material.
Mistral	5	10	7	The information provided in the answer is factual and accurate.

Table 7: Ratings for the answer to the question "How does the Doppler effect affect the observed frequency of sound waves emitted from a moving source and its relevance in real-world applications?"

D.2 Response Contains No Citations

How do the LLMs evaluate syntheses with no citations?

D.2.1 Art

Question: How has the evolution of technology impacted the production and consumption of art in the 21st century?

Answer: The evolution of technology in the 21st century has significantly impacted the production and consumption of art. The widespread access to digital tools and platforms has revolutionized artistic practices, enabling artists to experiment with new mediums and techniques beyond traditional forms. Furthermore, technology has democratized the consumption of art, providing individuals across the world with opportunities to view, share, and engage with creative works online. This shift has also amplified the global reach of art, exposing diverse perspectives and fostering cross-cultural exchanges. However, the rise of digital art has also raised questions about the authenticity and ownership of creative works, prompting a reevaluation of conventional definitions and practices in the art world.

	Comp.	Trust	Utility	Trust Rationale
GPT-4	1	1	2	The answer cannot be considered trustworthy in the context of the provided abstracts, as it does not cite any relevant information from them to support the claims made.
Mistral	5	8	7	The answer is based on common knowledge about the impact of technology on art and does not contain any false or misleading information.

Table 8: Ratings for the answer to the question "How has the evolution of technology impacted the production and consumption of art in the 21st century?"

D.2.2 History

Question: How did the American Civil War impact the social, economic, and political landscape of the United States?

Answer: The American Civil War significantly impacted the social, economic, and political landscape of the United States. On the social front, the war provided a catalyst for the abolition of slavery, which altered racial dynamics and paved the way for civil rights movements. Economically, the war devastated the South and forced a transition from an agrarian, slave-driven economy to a more diverse, industrial one. Reconstruction policies and investments led to the rapid growth of infrastructure and industry in the North, shaping the national economy for decades. Politically, the Civil War reaffirmed the primacy of the federal government over individual states and brought about transformative constitutional amendments such as the 13th, 14th, and 15th, which abolished slavery, granted citizenship to former slaves, and provided voting rights regardless of race, color, or previous servitude, respectively. The war reshaped the balance of power and the role of government in Americans' lives, laying the groundwork for a more unified and inclusive nation.

	Comp.	Trust	Utility	Trust Rationale
GPT-4	0	0	0	The answer cannot be trusted as it does not reference the provided abstracts, which do not relate to the American Civil War. The information seems to be general knowledge rather than sourced from the abstracts.
Mistral	10	10	9	The answer is trustworthy as it is derived directly from the provided abstracts and does not contain any additional or inaccurate information.

Table 9: Ratings for the answer to the question "How did the American Civil War impact the social, economic, and political landscape of the United States?"

D.2.3 Physics

Question: How does the shape of an object affect its resistance to motion through a fluid medium?

Answer: The shape of an object can significantly impact its resistance to motion through a fluid medium. This resistance, referred to as drag, depends on factors such as the size, shape, and surface texture of the object, along with the properties of the fluid in question. A streamlined shape minimizes drag, while a more irregular or bluff shape increases the force exerted by the fluid on the object, thus making it more resistant to motion. In fluid dynamics, this concept is commonly applied to improve the efficiency of objects such as vehicles and aircraft, where a streamlined design results in reduced fuel consumption and increased speed, due to the lowered drag forces acting against the object's movement through the fluid medium.

	Comp.	**Trust**	**Utility**	**Trust Rationale**
GPT-4	0	0	0	The answer cannot be trusted because it does not reference any of the provided abstracts, which are the only sources that should be used to inform the answer.
Mistral	0	10	3	The answer is trustworthy as it accurately references information from the provided text.

Table 10: Ratings for the answer to the question "How does the shape of an object affect its resistance to motion through a fluid medium?"

D.3 Response Contains One Citation

How do the LLMs evaluate syntheses with only one citation source?

D.3.1 Geology

Question: What geological factors contribute to the formation and sustainability of hydrothermal ecosystems in deep sea environments?

Answer: Hydrothermal ecosystems in deep sea environments are primarily formed and sustained by geological factors, such as tectonic activity and volcanic processes. These factors promote the circulation of seawater through the oceanic crust, which results in the formation of hydrothermal vents. The seawater is heated by the underlying magma and interacts with crustal rocks, enriching it with minerals and chemical elements. Once the heated and mineral-rich fluid meets the cold deep-sea water, it precipitates, creating chimney-like structures from which the hydrothermal fluid is expelled. These vents provide a unique habitat with diverse microbial and chemosynthetic communities, which serve as primary producers in the food chain. Furthermore, these deep-sea ecosystems are enriched by specialized fauna adapted to extreme conditions, such as high pressure, low oxygen, and varying temperatures, contributing to the overall sustainability of these environments [5].

	Comp.	Trust	Utility	Trust Rationale
GPT-4	7	8	7	The answer appears to be trustworthy as it aligns with scientific understanding of hydrothermal vent formation and the factors that sustain deep-sea ecosystems. It does not directly reference the provided abstracts, but the information given is consistent with general scientific knowledge on the topic.
Mistral	10	10	9	The answer is based on accurate information from the provided abstracts and uses reputable sources, such as scientific research, to support the statements made.

Table 11: Ratings for the answer to the question "What geological factors contribute to the formation and sustainability of hydrothermal ecosystems in deep sea environments?"

D.3.2 Mathematics

Question: What is the impact of using manipulatives in mathematics instruction on the achievement and attitudes of elementary school students?

Answer: The use of manipulatives in mathematics instruction has been shown to have a positive impact on the achievement and attitudes of elementary school students. In a study exploring the relationship between Affective Instructional Design (AID) and children's attitudes towards math learning, a significant positive correlation was found between attitude and math learning ($r = 0.936$, $p = 0.000$). Furthermore, significant differences were found between the baseline (pre-intervention) mean score and the final intervention lesson for both attitudes and math learning, indicating the potential effectiveness of manipulatives in enhancing students' attitudes towards mathematics and improving their achievement in the subject [1].

	Comp.	Trust	Utility	Trust Rationale
GPT-4	7	8	7	The answer is trustworthy as it accurately reflects the findings from the first abstract, including the significant correlation between attitude and math learning. The statistical evidence ($r = 0.936$, $p = 0.000$) adds credibility to the claim. However, the answer does not provide information on the sample size or the generalizability of the study, which could affect trust in the findings.
Mistral	10	9	9	The answer cites a peer-reviewed academic study, adding credibility to the information provided. However, it would be helpful to know more about the study's methodology and sample size to fully assess its trustworthiness.

Table 12: Ratings for the answer to the question "What is the impact of using manipulatives in mathematics instruction on the achievement and attitudes of elementary school students?"

D.3.3 Philosophy

Question: What is the relationship between free will and determinism in philosophy?

Answer: In philosophy, the relationship between free will and determinism revolves around whether an individual's actions are genuinely a result of their own choices or are predetermined by external factors. There are two main schools of thought: compatibilism and incompatibilism. Compatibilism argues that free will and determinism can coexist, meaning individuals can still be morally responsible for their actions even if determinism is true. Incompatibilism, on the other hand, claims that free will and determinism cannot coexist; if determinism is true, then individuals cannot have free will and cannot be held morally responsible for their actions. The debate between these two positions contributes to the ongoing philosophical discussions about responsibility, choice, and the implications of determinism [1].

	Comp.	**Trust**	**Utility**	**Trust Rationale**
GPT-4	7	9	8	The answer is trustworthy as it accurately reflects the general philosophical debate between compatibilism and incompatibilism, which are well-established positions in the literature on free will and determinism.
Mistral	8	9	7	The answer is based on accurate information from the provided abstracts, which discuss the relationship between free will, determinism, and moral responsibility. It does not contain any incorrect or misleading statements.

Table 13: Ratings for the answer to the question "What is the relationship between free will and determinism in philosophy?"

D.3.4 Physics

Question: How does the acceleration due to gravity affect the trajectory and motion of a projectile, and how can this be accurately measured and predicted using equations of motion?

Answer: The acceleration due to gravity affects the trajectory and motion of a projectile by altering its vertical motion while leaving its horizontal motion constant. As a result, the trajectory forms a parabolic path. To accurately measure and predict a projectile's motion, equations of motion can be used, which consider the initial velocity, launch angle, and time. The initial velocity can be split into horizontal and vertical components, and equations for both components can be applied independently. For the horizontal motion, an equation is used that doesn't consider the acceleration due to gravity, while for the vertical motion, the effect of gravity is integrated into the equation. These equations can be used to predict the projectile's motion, including its range and height [5].

	Comp.	Trust	Utility	Trust Rationale
GPT-4	8	7	9	The answer appears to be based on a general understanding of physics principles related to projectile motion and the effects of gravity. While it does not directly reference the provided abstracts, the information given is consistent with basic physics concepts. The trustworthiness would be higher if the answer included direct references to relevant data or studies from the abstracts.
Mistral	10	10	10	The answer is trustworthy as it directly references content from the provided abstracts regarding projectiles and equations of motion.

Table 14: Ratings for the answer to the question "How does the acceleration due to gravity affect the trajectory and motion of a projectile, and how can this be accurately measured and predicted using equations of motion?"

D.3.5 Psychology

Question: What are the underlying factors that contribute to the development of personality disorders and how can these disorders be effectively treated?

Answer: The development of personality disorders can be influenced by various factors, including genetic, environmental, and psychological components. While the specific causes for each individual's development of a personality disorder can vary, some contributing factors may include childhood trauma, family history of mental disorders, and negative early life experiences. Effective treatment for personality disorders typically involves a combination of psychotherapy, medication, and lifestyle changes. Therapy approaches such as cognitive-behavioral therapy, dialectical behavior therapy, and psychodynamic therapy can help individuals understand and manage their symptoms. In some cases, medications like antidepressants, mood stabilizers, and antipsychotics may be prescribed to help with specific symptoms. Additionally, self-help strategies, social support, and proper self-care can contribute to the overall well-being of individuals with personality disorders [mental health challenge - https://core.ac.uk/works/18689246].

	Comp.	Trust	Utility	Trust Rationale
GPT-4	5	5	6	The answer appears to be based on general knowledge about personality disorders and their treatment rather than the provided abstracts. While the information given is in line with common understanding in the field of mental health, the lack of direct references to the abstracts or specific studies reduces the trustworthiness of the answer.
Mistral	8	7	9	The answer is based on accurate information, but the sources are not explicitly cited, which may raise some questions about its trustworthiness.

Table 15: Ratings for the answer to the question "What are the underlying factors that contribute to the development of personality disorders and how can these disorders be effectively treated?"