Science Innovation
2025, Vol. 13, No. 6, pp. 162–169
https://doi.org/10.11648/j.si.20251306.15

Research Article

Demonstrating the Usage of Content Reliability Metrics on Comparing the Contentual Performance of Two LLMs

Johannes Miklos[1], **Zsolt Saffer**[1, 2, *]

[1]Department of Information Systems Engineering and Management, The Distance-Learning University of Applied Sciences (FERNFH), Wiener Neustadt, Austria

[2]Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria

Abstract

Using generative AI in education requires content reliability of the generated answers when the same question or its slightly changed version is asked again. Therefore, assessing the content reliability of the generated answers is an important issue. This requires appropriate contentual reliability metrics. In this paper we propose two new contentual reliability metrics for evaluating LLMs: content consistency and contentual robustness. They enable us to assess content reliability related performance, which cannot be evaluated by regular metrics like accuracy and faithfulness, but necessary due to the intrinsic random nature of generative AI. We demonstrate the usage of these content reliability metrics in assessing the contentual performance of two LLMs: one without and the other with reasoning model. The experiments run under identical software and hardware settings, and the source of information is restricted to a single PDF that serves as ground truth. The experiments are performed on two locally executed models, which ensures reproducibility of the experiments. The experiments use two question types: "easy questions" and "complicated questions" requiring multi-step reasoning. The results show that the MS phi-4-reasoning-plus model produces answers to complex questions not only with higher accuracy, but also with improved content consistency and contentual robustness. The mean accuracy changes from 0.36 to 0.84 when MS phi-4-reasoning-plus model is used instead of the MS phi-4 model, which corresponds to 133% improvement. Interestingly the experimental results do not show any difference in standard deviation of accuracy among the two models. The mean content consistency and mean contentual robustness changes from 0.55 to 0.92 and from 0.74 to 0.93 when the MS phi-4-reasoning-plus model is used instead of the MS phi-4 model, corresponding to 67% and 26 % improvements, respectively. The standard deviation of both content reliability metrics drops significantly when the MS phi-4 model is replaced by the reasoning model. Evaluating the effect of incorporating a reasoning model into an LLM to the contentual performance gives an insight into the operational reliability of the LLMs on contentual level. These results justified the hypothesis that not only the accuracy but also the operational reliability of the LLMs on contentual level has been significantly improved due to incorporating the reasoning model. The experiment design, the results and their evaluation demonstrated successfully the usage of the newly proposed content reliability metrics for assessing and comparing the contentual performance of LLMs.

Keywords

Performance Evaluation, LLM, Contentual Reliability Metrics, Contentual Performance

*Corresponding author: zsolt.saffer@tuwien.ac.at (Zsolt Saffer)

Received: 28 October 2025; **Accepted:** 19 November 2025; **Published:** 11 December 2025

1. Introduction

The paper "Attention Is All You Need " [1] presenting the Transformer architecture was a milestone on the way of developing generative Artificial Intelligences (AIs) imitating human-like textual answering capability and ultimately reaching the breakthrough with the release of ChatGPT ([2, 3]) in 2022. From that time many applications scenario and tools based on generative AI have been realized and using generative AI has become a part of the nowadays life. Introducing the use of reasoning model in generative AI caused the next breakthrough due to enhancing the capabilities of generative AI by logical reasoning.

The intrinsic nature of generative AI implies, especially due to the probabilistic frameworks on which it relies, and the random mechanisms applied in it, that the content of the generated answer is not necessarily the same when the same question or its slightly changed version is asked again. However, using generative AI in education requires content reliability of the generated answers. Therefore, assessing the content reliability of the generated answers becomes an important issue.

However, the regular metrics for assessing the performance of a Large Language Model (LLM), like e.g. accuracy and faithfulness ([4-7]), do not address the contentual reliability of the generated answer. To the best of our knowledge there are no formally defined metrics available for evaluating content reliability. It makes it necessary to introduce metrics addressing the content reliability of the generated answers.

In this paper two content reliability metrics are proposed. Content consistency measures roughly speaking the consistency of the content of the answers for repeating the same questions several times. Contentual robustness measures the ratio of the correctly retained content-related parts in the answers for the light changed, but contentually the same version of a question.

We demonstrate the usage of these content reliability metrics in assessing the contentual performance of two LLMS: one without and the other with reasoning model. The experiments run under identical software and hardware settings, and the source of information is restricted to a single document in Portable Document Format (PDF) that serves as ground truth. The experiments are performed on two locally executed models, which ensures the reproducibility of the experiments.

The experiments use two question types to reflect different cognitive demands. The first set comprises 5 "easy questions" that can be answered by locating information directly in the PDF. The second set comprises 5 "complex questions" that require multi step reasoning.

One regular metric and the two newly introduced content reliability metrics are used to judge and evaluate the effect of incorporating a reasoning model into an LLM to the contentual performance of the answers. This gives an insight into the operational reliability of the LLMs on contentual level.

The contributions of the work are

1. proposing two content reliability metrics,
2. demonstrating the usage of the new metrics by comparing the contentual performance of two selected LLMs by reproducible experiments using locally executed models and
3. evaluating the effect of incorporating a reasoning model into an LLM to the contentual performance of the answers.

The rest of this paper is organized as follows. In section 2 we introduce the new contentual performance metrics for LLM and define them formally. The experiment design including both the description of the environment and the composition of the test questions are discussed in section 3. In section 4 the experimental workflow and the architecture are provided. The test results and their evaluation are presented in section 5. Final remarks close the paper in section 6.

2. Contentual Performance Metrics for LLM

In this section we give a brief description of several regular contentual metrics for LLM and introduce the new contentual performance metrics by explaining their meaning and defining them formally.

2.1. Regular Contentual Metrics

We consider two regular contentual metrics from information retrieval point of view: Accuracy (Acc) and Faithfulness (Fai).

Accuracy gives the proportion of questions correctly answered by LLM to the total number of asked questions. In other words

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N} I_{\{answer\ to\ i-th\ question\ is\ right\}},$$

where N is the number of questions asked to LLM.

Faithfulness [8] is an evaluation metric for LLMs with provided information source determining the contentual context, like e.g. in case of restricting the information source to a given file or for Retrieval-Augmented Generation (RAG) system ([9, 10]). Faithfulness is the proportion of the number of claims in the answer of LLM being consistent to the context provided to the LLM as information source, among total number of claims in the answer. In other words:

$$\text{Fai} = \frac{the\ number\ of\ claims\ in\ the\ answer\ consistent\ to\ the\ provided\ context}{total\ number\ of\ claims\ in\ the\ answer}$$

2.2. New Contentual Reliability Metrics

We introduce two new contentual reliability metrics: content consistency (CoC) and contentual robustness (CoR). In contrast to the regular contentual metrics assessing the LLMs from information retrieval point of view, these new metrics evaluate the reliability of the content by focusing on the answering behavior of the LLM.

Content consistency is the average proportion of the correct content-related parts in the answers by repeatedly asking the same question to LLM. Formally content consistency is defined as

$$\mathrm{CoC} = \frac{1}{N}\sum_{i=1}^{N}\frac{k_i}{K},$$

where

1) K is the number of content-related parts in the right answer,
2) k_i is the number of right content-related parts in the i-th answer, i=1,…,K and
3) N is the number of asking the same question.

Contentual robustness is the average proportion of the correct content-related parts in the answers to light changed version of a question to LLM with the same content. Formally content consistency is defined as

$$\mathrm{CoR} = \frac{1}{NV}\sum_{i=1}^{NV}\frac{k_i}{K},$$

where NV is the number of lightly changed versions of the base question with the same content.

3. Experiment Design

In this section we describe the environment for controlled document-grounded assessment of two LLMs, the requirements again the test questions and the evaluation metrics used.

3.1. Environment for a Controlled, Document-grounded Assessment of Two LLMs

The experimental environment consists of two locally executed reasoning models, microsoft/phi-4 (MS phi-4) [11] and microsoft/phi-4-reasoning-plus (MS phi-4-reasoning-plus) [12-14], evaluated under identical software and hardware settings. The information source of the answers is restricted to a single PDF that serves as ground truth ([15]). Model lineage is made explicit for transparency and replication: MS phi-4 (Basic LLM = phi 14B) and MS phi-4-reasoning-plus (Basic LLM = phi 14.7B). The models were tested under the environment of LM Studio 0.3.30. All these together ensure the reproducibility of the experiments, i.e. to provide reproduceable results.

3.2. Requirements Against the Test Questions

The evaluation uses two question types to reflect different cognitive demands. The first set comprises 5 "easy questions" that can be answered by locating information directly in the PDF. The second set comprises 5 "complex questions" that require multi step reasoning because of processing demand.

The question should be constructed in that way that the right answer to them should have at least 3 good distinguishable content-related parts. This ensures a necessary level of contentual granulation which is needed to get the results on contentual reliability metrics being enough meaningful due to their formal definition.

Both the number of asking the same question and the number of lightly changed versions of the base question should be 5.

3.3. Test Questions

The test questions used for the experiments are constructed according to the above requirements. The test questions and their distinguishable content-related parts are summarized in Appendix.

3.4. Evaluation Metrics

The answers for the test questions are evaluated in terms of one of the above-mentioned regular contentual metrics and the newly introduced contentual reliability metrics. Thus, the used metrics are given as

- accuracy,
- content consistency and
- contentual robustness.

Note, that faithfulness was not used as evaluation metric for the considered experiments.

4. Experimental Workflow and Architecture

This section is devoted to the details of the experimental workflow and the architectural overview of the experiments.

4.1. Objective and Design Rationale

The experimental workflow operationalizes a controlled, document-grounded comparison of two locally executed models, one non-reasoning model MS phi-4 and one reasoning model MS phi-4-reasoning-plus, under a single execution boundary. The goal is to assess contentual performance while document fidelity is kept and environmental variance is minimized. To this end, all inference is performed on the same workstation within LM Studio 0.3.30, and all prompts are

answered without external retrieval. A single PDF, designated as Truth, provides the only authoritative reference for scoring. The design follows a paired, within-item paradigm so that each question functions as its own control across models.

4.2. Architectural Overview

4.2.1. Executional Constraints

LM Studio 0.3.30 encapsulates both models with preset standard runtime constraints (token limits, decoding parameters, seeds). Network access is disabled to enforce a closed-world setting. Caching across models is prevented to avoid cross-contamination.

4.2.1. Architecture

The architectural overview of the experiments is shown in Figure 1.

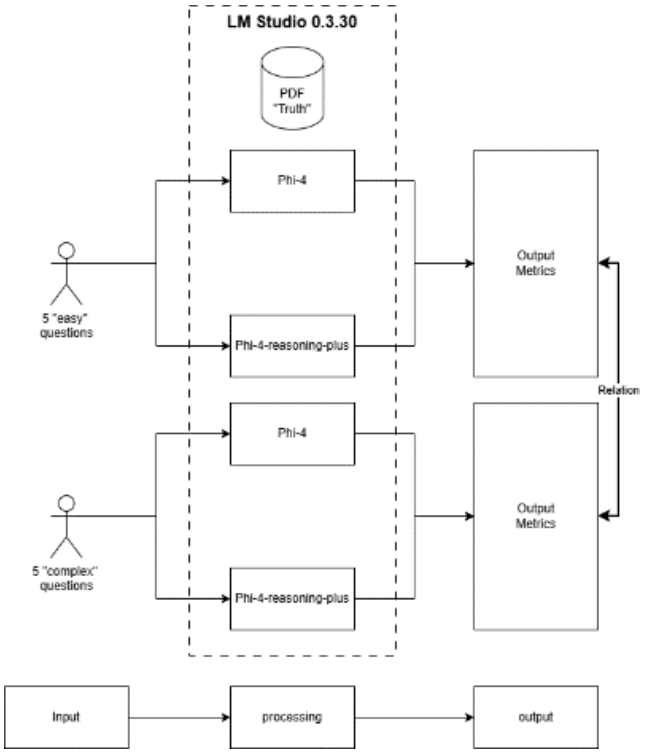

Figure 1. *Architectural overview of the experiments.*

The experiment is executed entirely within LM Studio 0.3.30, which serves as a single, controlled execution boundary for both models. Inside this sandbox, a document-grounded constraint is imposed: a single PDF ("Truth") is the unique reference against which answers are judged. Network access is disabled, and decoding parameters (seed, temperature, top-p, max tokens) are held fixed to ensure run-to-run comparability.

Two parallel streams of prompts enter the sandbox: a set of five "easy" questions that require direct lookup in the PDF and a set of five "complex" questions that require composition (e.g., combining table entries or applying study-defined transformations). Each question is decomposed into three content parts (a–c) and is posed to both models, MS Phi-4 (non-reasoning) and MS Phi-4-reasoning-plus, in a paired design so that each item acts as its own control across models.

For each stream, both models run under identical runtime constraints. Their raw responses are captured alongside minimal metadata (prompt, seed, timestamp) to enable exact replication. Because the PDF is the only allowed knowledge source, every claim in an answer must be traceable to a specific passage or table cell in that document.

5. Results and Evaluation

In this section the results of the experiments are provided. The results for MS phi-4 and MS phi-4-reasoning-plus models are compared and evaluated individually in terms of accuracy, content consistency and contentual robustness.

5.1. Accuracy

For the easy questions (SQ1.–SQ5.), accuracy was perfect, i.e. 100%, both with MS phi-4 and MS phi-4-reasoning-plus models. Hence all content parts were correct in all the five repetitions.

For the complex questions (CQ1.–CQ5.) the accuracy became lower and shows strong model dependent values.

For the MS phi-4 model the accuracy decreased stronger. A per-question view shows the accuracy of the individual questions as CQ1. = 2/5 (40.0%), CQ2. = 2/5 (40.0%), CQ3. = 1/5 (20.0%), CQ4. = 1/5 (20%), being the most difficult item, and CQ5. = 3/5 (60%), which was the most tractable among the complex questions.

For the MS phi-4-reasoning-plus model the accuracy remained relatively high for the complex questions. A per-question breakdown confirms in general strong performance within the complex set: CQ1. = 4/5 (80.0%), CQ2. = 5/5 (100%), CQ3. = 3/5 (60.0%), CQ4. = 4/5 (80.0%), and CQ5. = 5/5 (100%).

Figure 2. shows the comparison of the accuracy values for the individual questions among the MS phi-4 and MS phi-4-reasoning-plus models.

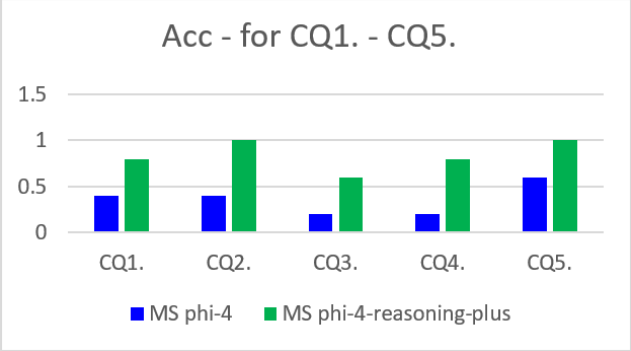

Figure 2. *Comparison of Acc for complex questions.*

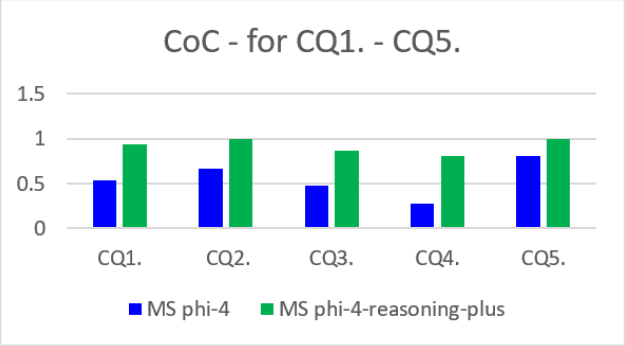

Figure 3. *Comparison of CoC for complex questions.*

Table 1 summarizes the accuracy values for the individual questions together with their mean and standard deviation (SD) both for the MS phi-4 and MS phi-4-reasoning-plus models.

Table 1. *Acc for complex questions.*

Acc CQs/LLM model	MS phi-4	MS phi-4-reasoning-plus
CQ1.	0.4	0.8
CQ2.	0.4	1
CQ3.	0.2	0.6
CQ4.	0.2	0.8
CQ5.	0.6	1
Mean	0.36	0.84
SD	0.15	0.15

It can be seen in Table 1 that the performance of the reasoning model is much better for each of the individual CQs. The results show that the mean accuracy changes from 0.36 to 0.84 when MS phi-4-reasoning-plus model is used instead of the MS phi-4 model. This corresponds to 133% improvement in mean accuracy. However, the standard deviation of the accuracy does not show any difference among the two models.

5.2. Content Consistency

Both MS phi-4 and MS phi-4-reasoning-plus models show perfect content consistency for the simple questions, since all the three content-related parts have arisen in the answers of all the five repetition of SQ1. – SQ5.

However, the content consistency shows degradation for complex questions in both models. The corresponding values for the individual complex questions are compared for the models in Figure 3.

Additionally, the content consistency values for the individual questions together with their mean and standard deviation both for the MS phi-4 and MS phi-4-reasoning-plus models are summarized in Table 2.

Table 2. *CoC for complex questions.*

CoC CQs/LLM model	MS phi-4	MS phi-4-reasoning-plus
CQ1.	0.53	0.93
CQ2.	0.67	1
CQ3.	0.47	0.87
CQ4.	0.27	0.8
CQ5.	0.8	1
Mean	0.55	0.92
SD	0.18	0.08

It can be seen in Table 2 that the performance of the reasoning model is much better for each of the individual CQs.

The results show that the mean content consistency changes from 0.55 to 0.92 when the MS phi-4-reasoning-plus model is used instead of the MS phi-4 model. This corresponds to 67% improvement. Additionally, the standard deviation of the content consistency drops from 0.18 to 0.08.

5.3. Contentual Robustness

Contentual robustness has been evaluated for a set of selected questions: SQ2., CQ1., CQ2., CQ4. and CQ5. The comparison of the corresponding values for these questions is shown for both models in Figure 4.

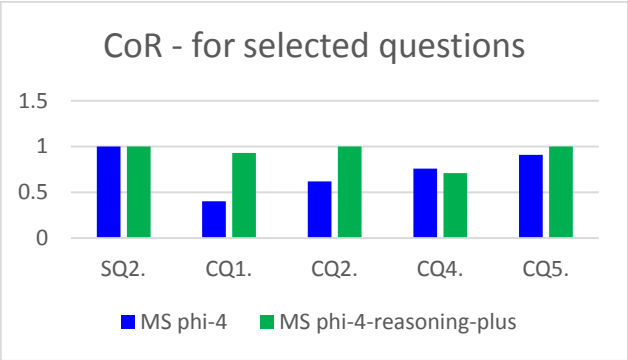

Figure 4. *Comparison of CoR for selected questions.*

Table 3 summarizes the contentual robustness values together with their mean and standard deviation both for the MS phi-4 and MS phi-4-reasoning-plus models.

Table 3. *CoR for selected questions.*

CoR CQs/LLM model	MS phi-4	MS - phi-4-reasoning-plus
SQ2.	1	1
CQ1.	0.4	0.93
CQ2.	0.62	1
CQ4.	0.76	0.71
CQ5.	0.91	1
Mean	0.74	0.93
SD	0.21	0.11

It can be seen in Table 3 that the performance of the reasoning model is better for each of the selected questions, except for CQ4.

The results show that the mean contentual robustness changes from 0.74 to 0.93 when the MS phi-4 model is replaced by the MS phi-4-reasoning-plus model. This corresponds to 26% improvement. Additionally, the standard deviation of contentual robustness with the reasoning model is approximately only half of that with the MS phi-4 model.

6. Summary

In this work we introduced two new content reliability metrics (content consistency and contentual robustness) by giving their meaning and also defining them formally.

Their usage for comparing the contentual performance of two selected LLMs were demonstrated. Locally executed models were applied in the experiments to ensure their reproducibility.

The experiments show that the MS phi-4-reasoning-plus model produces answers for complex questions with higher accuracy, achieving a mean accuracy of 0.84 on the investigated complex question set. The mean accuracy changes from 0.36 to 0.84 when MS phi-4-reasoning-plus model is used instead of the MS phi-4 model, which corresponds to 133% improvement. Interestingly the experiments do not show any difference in standard deviation of accuracy among the two models. However, this standard deviation of accuracy with the value 0.15 is small compared to the above mean accuracy.

Furthermore, both the newly proposed content reliability metrics show improvement in case of using the reasoning model, achieving a mean content consistency of 0.92 and mean contentual robustness of 0.93. The mean content consistency changes from 0.55 to 0.92 when the MS phi-4-reasoning-plus model is used instead of the MS phi-4 model, which corresponds to 67% improvement. The mean contentual robustness changes from 0.74 to 0.93 when the MS phi-4 model is replaced by the MS phi-4-reasoning-plus model, which means a 26% improvement. Moreover, the standard deviation of both content reliability metrics drops significantly when the MS phi-4 model is replaced by the reasoning model.

These results justified the hypothesis that not only the accuracy but also the operational reliability of the LLMs on contentual level has been significantly improved due to incorporating the reasoning model.

The experiment design, the results and their evaluation demonstrated successfully the usage of the newly proposed content reliability metrics for assessing and comparing the contentual performance of LLMs.

Potential future research work includes

- comparison of two LLMs both having reasoning models incorporated and
- constructing test questions with requirements for another tests.

Abbreviations

AI	Artificial Intelligence
LLM	Large Language Model
PDF	Portable Document Format
RAG	Retrieval-Augmented Generation
Acc	Accuracy
Fai	Faithfulness
CoC	Content Consistency
CoR	Contentual Robustness
SQx	Easy Question x, x=1,…,5
CQx	Complex Question x, x=1,…,5
SD	Standard Deviation

Author Contributions

Johanes Miklos: Experiment design, Running experiments

Zsolt Saffer: Conceptualization, Methodology

Funding

This work is not supported by any external funding.

Data Availability Statement

The data supporting the outcome of this research work has been reported in this manuscript (see Appendix).

Conflicts of Interest

The authors declare no conflicts of interest.

Appendix: Test Questions

Appendix I: Simple Questions SQ1. - SQ5.

SQ1. Model fit (day level three-factor CFA) — report (a) χ^2 and df, (b) CFI and TLI, (c) RMSEA. Please give a short and precise answer. Choose the wording and sentences as near to the document as p2.1.1. Sample Heading possible. Please answer in the same (letter) structure.

SQ2. Alternative models (direct comparison) — state (a) best two-factor model fit indices (χ^2, CFI, TLI, RMSEA), (b) $\Delta\chi^2$ vs. three-factor with df and p, (c) one-factor model χ^2 and RMSEA. Please give a short and precise answer. Choose the wording and sentences as near to the document as possible. Please answer in the same (letter) structure.

SQ3. Descriptives of the final sample — give (a) age range and mean, (b) mean weekly work hours, (c) percent women. Please give a short and precise answer. Choose the wording and sentences as near to the document as possible. Please answer in the same (letter) structure.

SQ4. Procedure highlights (diary protocol) - state (a) number/timing of diary days, (b) reminder time and channels, (c) screening rule. Please give a short and precise answer. Choose the wording and sentences as near to the document as possible. Please answer in the same (letter) structure.

SQ5. Measures snapshot — report for Coordination Activities (a) item/source wording, (b) Likert range, (c) measurement frequency. Please give a short and precise answer. Choose the wording and sentences as near to the document as possible. Please answer in the same (letter) structure.

Appendix II: Complex Questions CQ1. - CQ5.

CQ1. Participant flow (percentages of registrants) — compute (a) no basic questionnaire, (b) excluded for no diary, (c) final inclusion rate. Please give a short and precise answer. Choose the wording and sentences as near to the document as possible. Please answer in the same (letter) structure.

CQ2. Retention and largest exclusion — (a) final inclusion rate (final N ÷ registrants), (b) share excluded for no basic questionnaire, (c) share excluded for no pre-pandemic remote option (and whether it's the largest screen). Please give a short and precise answer. Choose the wording and sentences as near to the document as possible. Please answer in the same (letter) structure.

CQ3. Diary completion — (a) maximum possible day-level observations (5×N), (b) realized completion rate (actual ÷ maximum, %), (c) average completed days per participant (actual ÷ N, 2 decimals). Please give a short and precise answer. Choose the wording and sentences as near to the document as possible. Please answer in the same (letter) structure.

CQ4. Implied vs. analyzed diaries — (a) implied diaries (mean days × N), (b) absolute difference to analyzed days, (c) that difference as % of analyzed days. Please give a short and precise answer. Choose the wording and sentences as near to the document as possible. Please answer in the same (letter) structure.

CQ5. Gender ratios — (a) % men (100 − % women), (b) female:male ratio (% women ÷ % men), (c) express as "1 : X" (men per woman, 2 decimals). Please give a short and precise answer. Choose the wording and sentences as near to the document as possible. Please answer in the same (letter) structure.

References

[1] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Łukasz; Polosukhin Illia. "Attention is All you Need". 2017. In I. Guyon and U. Von Luxburg and S. Bengio and H. Wallach and R. Fergus and S. Vishwanathan and R. Garnett (ed.). *31st Conference on Neural Information Processing Systems (NIPS)*. Advances in Neural Information Processing Systems. Vol. 30. Curran Associates.

[2] Roumeliotis, Konstantinos I.; Tselikas, Nikolaos D. "ChatGPT and Open-AI Models: A Preliminary Review". 2023. Future Internet. 15 (6): 192. https://doi.org/10.3390/fi15060192

[3] Partha Pratim Ray. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. 2023. Internet of Things and Cyber-Physical Systems, Volume 3. Pages 121-154. https://doi.org/10.1016/j.iotcps.2023.04.003

[4] Taojun Hu, Xiao-Hua Zhou. Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions. 2024. ArXiv.abs/2404.09135. https://api.semanticscholar.org/CorpusID:269149430

[5] Yinuo Dong, Zili Zhang, Yang Zhi, Xiaoyun Li, Tengyu Guo, Lanlan He, Shuran Zhao, Xueli Yang, Jieting Tang, Wei Zhong, Qinghui Niu, Mingyang Ma, Zuxiong Huang, Yimin Mao. Evaluating Large Language Models' Performance in Answering Common Questions on Drug-Induced Liver Injury. 2025. JHEP Reports, 101579, ISSN 2589-5559, https://doi.org/10.1016/j.jhepr.2025.101579

[6] A. R ós-Hoyo, N. L. Shan, A. Li, A. T. Pearson, L. Pusztai, and F. M. Howard. "Evaluation of large language models as a diagnostic aid for complex medical cases,". 2024. Front Med (Lausanne), vol. 11, p. 1380148.

[7] L. Huang *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Trans Inf Syst*, 2023.

[8] Malin, Ben and Kalganova, Tatiana and Boulgouris, Nikolaos. A review of faithfulness metrics for hallucination assessment in Large Language Models.2025. IEEE Journal of Selected Topics in Signal Processing, p. 1–13, http://dx.doi.org/10.1109/JSTSP.2025.3579203

[9] Shailja Gupta and Rajesh Ranjan and Surya Narayan Singh.A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. 2024. arXiv 2410.12837. https://arxiv.org/abs/2410.12837

[10] Zongxi Li, Zijian Wang, Weiming Wang, Kevin Hung, Haoran Xie, Fu Lee Wang. Retrieval-augmented generation for educational application: A systematic survey.2025.

[11] Abdin, M., Aneja, J., Behl, H. S., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., Lee, J. R., Lee, Y. T., Li, Y., Liu, W., Mendes, C. C., Nguyen, A., Price, E., Rosa, G. D., Saarikivi, O., Salim, A., Shah, S., Wang, X., Ward, R., Wu, Y., Yu, D., Zhang, C., & Zhang, Y. Phi-4 Technical Report. 2024. *ArXiv*, *abs/2412.08905*

[12] Marah Abdin and Jyoti Aneja and Harkirat Behl and S ébastien Bubeck and Ronen Eldan and Suriya Gunasekar and Michael Harrison and Russell J. Hewett and Mojan Javaheripi and Piero Kauffmann and James R. Lee and Yin Tat Lee and Yuanzhi Li and Weishung Liu and Caio C. T. Mendes and Anh Nguyen and Eric Price and Gustavo de Rosa and Olli Saarikivi and Adil Salim and Shital Shah and Xin Wang and Rachel Ward and Yue Wu and Dingli Yu and Cyril Zhang and Yi Zhang.Phi-4 Technical Report.2024. arXiv 2412.08905. https://arxiv.org/abs/2412.08905

[13] Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, Tushar Khot.Chain-of-Thought Hub: A Continuous Effort to Measure Large Language Models' Reasoning Performance. 2023. arXiv2305.17306. https://arxiv.org/abs/2305.17306

[14] Siwei Wu and Zhongyuan Peng and Xinrun Du and Tuney Zheng and Minghao Liu and Jialong Wu and Jiachen Ma and Yizhi Li and Jian Yang and Wangchunshu Zhou and Qunshu Lin and Junbo Zhao and Zhaoxiang Zhang and Wenhao Huang and Ge Zhang and Chenghua Lin and J. H. Liu.A Comparative Study on Reasoning Patterns of OpenAI's o1 Model.2024. arXiv 2410.13639. https://arxiv.org/abs/2410.13639

[15] M ühl, A., Sch öllbauer, J., Straus, E., & Korunka, C. Threatening relatedness while boosting social interactions: the inconsistent effect of daily task ambiguity on daily relatedness satisfaction among remote workers. 2024. *The International Journal of Human Resource Management*, *36*(1), 56-79. https://doi.org/10.1080/09585192.2024.2439264