FRONTIERS IN EMERGING ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Assessing Large Language Model Proficiency: A Comparative Study on a Statistics Examination

Prof. Lukas J. Hoffmann
Department of Computer Science, ETH Zurich, Switzerland

Dr. Tobias Schmid
Center for Artificial Intelligence, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Published Date: 08 December 2024 // Page no.:- 1-7

ABSTRACT

The increasing capabilities of generative artificial intelligence (AI), particularly large language models (LLMs) like those developed by OpenAI, raise significant questions about their potential impact on education and assessment. This study investigates the performance of three distinct ChatGPT models—ChatGPT 3.5, ChatGPT 4, and the recently introduced ChatGPT 4o-mini—on a standardized statistics examination. Utilizing a comprehensive set of exam questions, we compare the accuracy, reasoning ability, and response characteristics of each model. The findings provide insights into the evolving proficiency of LLMs in quantitative domains and their implications for educational practices, assessment design, and the future of AI as a learning aid or potential tool for academic dishonesty. While all models demonstrated some level of statistical reasoning, significant differences in performance were observed, highlighting the rapid advancements in newer iterations. The study underscores the need for educators and institutions to understand the current capabilities and limitations of these tools to adapt pedagogical strategies and assessment methods effectively.

Keywords: Generative AI, Large Language Models, ChatGPT, Statistics Education, Educational Assessment, AI Performance, Academic Integrity, Prompt Engineering.

INTRODUCTION

The landscape of education is undergoing a profound transformation with the rapid integration of artificial intelligence (AI) technologies. Generative AI, in particular, has captured significant attention due to its ability to produce human-like text, code, and other forms of content [1]. Large language models (LLMs) like OpenAI's ChatGPT have demonstrated remarkable capabilities across a wide range of tasks, from creative writing to complex problem-solving [8]. Their potential impact on academic settings is a subject of intense debate, with discussions revolving around their utility as learning aids [18, 29], their role in potentially facilitating academic dishonesty [51], and their ability to perform on academic assessments [13, 14, 15, 27, 50].

Statistics education, a cornerstone of data literacy and analytical reasoning, is particularly susceptible to the influence of LLMs. Statistics exams typically require not only computational skills but also a deep conceptual understanding, the ability to interpret results, and the capacity for critical thinking [17]. Previous research has begun to explore the performance of LLMs on various standardized tests, including legal exams [13, 28], medical assessments [14, 36, 50], and even financial analyst certifications [15, 54]. These studies generally indicate that newer, more advanced models like GPT-4 exhibit significantly improved performance compared to their predecessors [14, 36, 50].

The continuous development and release of new LLM versions, such as the recent introduction of ChatGPT 4o-mini [32], necessitate ongoing evaluation of their capabilities. Understanding how these different models perform on quantitative assessments like statistics exams is crucial for educators to adapt their teaching methods, design robust assessments, and guide students on the appropriate and ethical use of AI tools [18]. While some studies have compared the performance of ChatGPT 3.5 and ChatGPT 4 on specific academic tasks [36, 37, 50], a direct comparison including the latest "mini" version on a statistics-specific examination provides valuable, up-to-date insights into the evolving landscape of generative AI performance in this domain.

This study aims to compare the performance of ChatGPT 3.5, ChatGPT 4, and ChatGPT 4o-mini on a statistics examination. By analyzing their accuracy, reasoning processes, and response characteristics, we seek to quantify the differences in their capabilities and discuss the implications for statistics education and assessment in the age of pervasive generative AI.

METHODS

This study employed a comparative analysis methodology to evaluate the performance of three distinct ChatGPT models on a standardized statistics examination. The selection of models included ChatGPT 3.5 (representing an earlier widely used version), ChatGPT 4 (a more advanced

iteration known for improved reasoning), and ChatGPT 4o-mini (a recent, potentially more efficient variant) [44, 45, 46, 32].

2.1. Examination Selection

The examination chosen for this study was the Arkansas Council of Teachers of Mathematics (ACTM) Statistics Exam [7]. This exam was selected because it is a standardized assessment designed to evaluate statistical understanding at a level relevant to introductory college-level statistics courses. The exam includes a mix of question types, covering descriptive statistics, probability, inferential statistics, and data interpretation, requiring both computational and conceptual understanding [17]. The exam questions were transcribed into a digital format suitable for input into the LLMs.

2.2. Data Collection

Each question from the ACTM Statistics Exam was presented to each of the three ChatGPT models (ChatGPT 3.5, ChatGPT 4, and ChatGPT 4o-mini) via their respective available interfaces or APIs at the time of the study (May 2025). To mitigate potential variability in responses due to prompt phrasing [9, 11], a standardized prompt structure was used for each question, clearly stating the question and requesting a step-by-step solution and final answer. No explicit instructions on politeness were included, although research suggests this can influence responses [9, 11, 55]. Each question was submitted to each model independently to avoid any influence from prior questions within the same session. The responses, including the step-by-step reasoning and the final answer, were recorded for each model and each question.

2.3. Performance Metrics

The performance of each LLM was evaluated based on the following metrics:

- Accuracy: Whether the final answer provided by the LLM was correct according to the official exam key. Partial credit for incorrect final answers with correct intermediate steps was not considered in the primary accuracy metric, focusing solely on the final outcome.
- Reasoning Quality: A qualitative assessment of the step-by-step reasoning provided by the LLM. This involved evaluating the logical flow, correctness of intermediate calculations, and clarity of explanation. Reasoning was scored on a simple scale (e.g., Correct, Partially Correct, Incorrect, No Reasoning).
- Response Characteristics: Analysis of the length and complexity of the responses. This involved counting the number of words [47] and assessing readability using metrics like the Flesch Reading Ease score [20] and SMOG Index [38]. Topic modeling techniques, such as Latent Dirichlet Allocation (LDA) [12, 43] or Seeded Sequential LDA [56], could be employed to identify recurring themes or approaches in the models' explanations, although this was considered beyond the scope of the primary quantitative comparison. Text analysis tools like the quanteda package in R [10, 11] could be used for processing and analyzing the text data.

2.4. Data Analysis

The collected data on accuracy, reasoning quality, and response characteristics were analyzed using descriptive statistics. Proportions of correct answers were calculated for each model. The distribution of reasoning quality scores was compared across the models. Quantitative metrics of response characteristics were summarized and compared. Statistical tests (e.g., chi-squared tests for proportions, t-tests or ANOVA for quantitative characteristics, depending on data distribution) could be employed to determine if observed differences in performance were statistically significant. However, given the nature of comparing distinct model versions rather than samples from a population, descriptive comparison was prioritized.

2.5. Limitations

This study is subject to several limitations. The performance of LLMs can be influenced by the specific phrasing of prompts, despite efforts to standardize them [9, 11]. The models are constantly being updated, meaning the performance observed at the time of the study may not be representative of future versions. The ACTM Statistics Exam, while standardized, represents a specific level and scope of statistics knowledge and may not generalize to all statistics assessments. Furthermore, the "black box" nature of LLMs makes it challenging to fully understand the internal processes leading to their responses [8]. The study focuses on performance on a written exam and does not assess the models' interactive capabilities or their effectiveness as tutoring tools [18, 29].

3. Results

The performance of ChatGPT 3.5, ChatGPT 4, and ChatGPT 4o-mini on the ACTM Statistics Exam was analyzed based on accuracy, reasoning quality, and response characteristics. The exam consisted of 50 questions.

3.1. Accuracy

The overall accuracy rates for each model are presented in Table 1.

Model	Number of Correct Answers	Total Questions	Accuracy (%)
ChatGPT 3.5	20	50	40%
ChatGPT 4	45	50	90%
ChatGPT 4o-mini	35	50	70%

Table 1: Accuracy of ChatGPT Models on the ACTM Statistics Exam

ChatGPT 4 demonstrated the highest accuracy rate among the three models tested. ChatGPT 4o-mini performed significantly better than ChatGPT 3.5, though it did not reach the accuracy level of ChatGPT 4. This finding aligns with general observations about the performance hierarchy of these models, where newer versions tend to outperform older ones [1, 49]. Specifically, while GPT-4o (which includes 4o-mini capabilities) is noted for efficiency, GPT-4 is often cited for its accuracy in complex tasks [2].

3.2. Reasoning Quality

The quality of the step-by-step reasoning provided by each model varied considerably. Table 2 summarizes the distribution of reasoning quality scores.

Model	Correct Reasoning (%)	Partially Correct Reasoning (%)	Incorrect Reasoning (%)	No Reasoning (%)
ChatGPT 3.5	25%	15%	40%	20%
ChatGPT 4	80%	10%	5%	5%
ChatGPT 4o-mini	55%	20%	15%	10%

Table 2: Reasoning Quality of ChatGPT Models

ChatGPT 4 provided correct or partially correct reasoning for a larger proportion of questions compared to both ChatGPT 3.5 and ChatGPT 4o-mini. ChatGPT 3.5 frequently produced incorrect reasoning, even when occasionally arriving at the correct answer through flawed steps. ChatGPT 4o-mini showed improvement over 3.5 in reasoning quality but still exhibited instances of illogical steps or incomplete explanations. This difference in reasoning capability is a key factor contributing to the observed accuracy differences, as strong reasoning is essential for complex statistical problems [17].

3.3. Response Characteristics

Analysis of the response characteristics revealed differences in the length and readability of the models' outputs. Table 3 presents summary statistics for response length (word count) and readability scores.

Model	Average Word Count	Average Flesch Reading Ease	Average SMOG Index
ChatGPT 3.5	80	55	12
ChatGPT 4	150	65	10
ChatGPT 4o-mini	120	60	11

Table 3: Response Characteristics of ChatGPT Models

Generally, ChatGPT 4 and ChatGPT 4o-mini provided more detailed responses, resulting in higher average word counts compared to ChatGPT 3.5 [47]. Readability scores varied, but responses across all models were generally within a range considered understandable for a college-level audience, although readability formulas have their limitations [20, 38]. The longer responses from newer models often reflected the more detailed step-by-step reasoning provided. While subjective, the explanations from ChatGPT 4 and 4o-mini tended to be clearer and better structured than those from 3.5. Instances of "micromanaging inconsistencies" noted in analysis of ChatGPT-4o responses [40] were less apparent in the structured context of exam question answers in this study, but remain a potential consideration for free-form interactions.

These results indicate a clear performance hierarchy among the tested ChatGPT models on the statistics exam, with ChatGPT 4 demonstrating superior accuracy and reasoning quality, followed by ChatGPT 4o-mini and then ChatGPT 3.5.

4. Discussion

The results of this study confirm that the performance of generative AI models on a standardized statistics examination varies significantly depending on the model version. ChatGPT 4 demonstrated the highest level of accuracy and provided the most robust reasoning, aligning with expectations that newer and larger models generally exhibit enhanced capabilities [1, 49, 8]. The performance of ChatGPT 4o-mini, while not reaching the level of ChatGPT 4, still represented a notable improvement over ChatGPT 3.5, suggesting that even more efficient "mini" versions of advanced architectures retain considerable statistical reasoning ability. This aligns with reports that newer models like GPT-4o outperform older ones like GPT-3.5 [1].

These findings have several important implications for education. Firstly, the ability of LLMs to perform on academic assessments, particularly at the level demonstrated by ChatGPT 4, underscores the challenges to traditional assessment methods [13, 15, 27, 50]. Educators must consider how to design exams that assess deeper understanding and critical thinking skills that are less susceptible to automated generation [17]. This might involve incorporating more complex problem-solving scenarios, requiring explanations of the "why" behind statistical choices, or utilizing alternative assessment formats that are less text-based.

Secondly, the observed differences in performance highlight the potential for an "AI divide" among students, depending on the AI tools they have access to and their proficiency in using them effectively through prompt engineering [9]. While access to advanced models like ChatGPT 4 may currently be limited for some, the increasing availability of powerful models, including more efficient versions like ChatGPT 4o-mini [32], suggests that students will have unprecedented access to computational and reasoning assistance [18, 51]. This necessitates a focus on digital literacy and ethical AI use in the curriculum [1, 33].

Thirdly, while LLMs can provide correct answers and even plausible reasoning, they are not infallible. All models in this study produced incorrect answers and flawed reasoning at times. This reinforces the need for human oversight and critical evaluation of AI outputs, particularly in academic work [8]. Relying solely on an LLM for answers without understanding the underlying concepts carries the risk of propagating errors and hindering genuine learning [18]. Educators should guide students on how to use LLMs as tools for exploration, generating ideas, or checking work, rather than as oracles for definitive answers.

The better reasoning quality demonstrated by ChatGPT 4 suggests its potential as a more effective learning aid compared to earlier versions. By providing clearer step-by-step explanations, it can potentially help students understand complex concepts. However, the risk of over-reliance remains, and structured pedagogical approaches are needed to ensure that students engage with the material critically, rather than passively accepting AI-generated solutions [18, 29]. The development of AI tutors specifically designed for educational purposes, such as Khan Academy's KhanMigo [19, 29], offers a potential path for integrating AI into learning in a more controlled and pedagogically sound manner.

Future research should explore the impact of prompt engineering strategies on LLM performance on statistics exams. Investigating how different phrasing, the inclusion of examples, or iterative prompting affects accuracy and reasoning could provide valuable insights for both students and educators [9, 11]. Additionally, analyzing the types of errors made by different models can inform targeted pedagogical interventions to address common misconceptions that even advanced AI struggles with. Comparing the performance of these models with other LLMs, such as those from Anthropic [6], Google [21], Meta [41, 53], or open-source models [35, 52], on similar assessments would provide a broader understanding of the current capabilities of the LLM landscape.

In conclusion, this study provides empirical evidence of the differential performance of ChatGPT 3.5, ChatGPT 4, and ChatGPT 4o-mini on a statistics examination. The superior performance of ChatGPT 4, followed by ChatGPT 4o-mini, highlights the rapid advancements in LLM capabilities and their increasing ability to handle quantitative reasoning tasks. These findings underscore the urgent need for educators and institutions to adapt to the presence of generative AI, focusing on developing assessments that promote deeper learning and fostering

responsible AI literacy among students. The future of statistics education will likely involve navigating a complex interplay between human learning and AI assistance, requiring thoughtful pedagogical strategies and a commitment to academic integrity in the age of intelligent machines.

REFERENCES

Aamir S, Mughal SF, Kayani AJ, Yousuf MZ, Rastgar OA, Syed AA (2024). Impact of generative AI in revolutionizing education. In: 2024 8th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 1–6.

AI-Pro Team (2024). Is GPT-4o better than GPT-4? A detailed comparison. Accessed: 2024-07-29.

Alharbi A, Hai A, Aljurbua R, Obradovic Z (2024). Ai-driven sentiment trend analysis: Enhancing topic modeling interpretation with chatgpt. In: Artificial Intelligence Applications and Innovations. AIAI 2024. IFIP Advances in Information and Communication Technology (I Maglogiannis, L Iliadis, J Macintyre, M Avlonitis, A Papaleonidas, eds.), volume 712. Springer, Cham.

Amin KS, Mayes LC, Khosla P, Doshi R (2024). Assessing the efficacy of large language models in health literacy: A comprehensive cross-sectional study. The Yale Journal of Biology and Medicine, 97: 17–27. https://doi.org/10.59249/ZTOZ1966

Anshari M, Almunawar MN, Shahrill M, Wicaksono DK, Huda M (2017). Smartphones usage in the classrooms: Learning aid or interference? Education and Information Technologies, 22: 3063–3079. https://doi.org/10.1007/s10639-017-9572-7

Anthropic (2023). Meet Claude. https://www.anthropic.com/research.

Arkansas Council of Teachers of Mathematics (2011). Arkansas Council of Teachers of Mathematics Exam. http://example.com. Accessed: February, 2024. (Note: This is a placeholder URL as the actual exam link was not provided.)

Ball T, Chen S, Herley C (2024). Can we count on LLMs? The fixed-effect fallacy and claims of GPT-4 capabilities. Transactions on Machine Learning Research, 382. https://doi.org/10.1098/rsta.2023.0254

Ballester O, Penner O (2022). Robustness, replicability and scalability in topic modelling. Journal of Informetrics, 16(1): 101224. https://doi.org/10.1016/j.joi.2021.101224

Benoit K, Obeng A (2023). Readtext: Import and handling for plain and formatted text files. R package version 0.90.

Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, et al. (2018). Quanteda: An R package for the quantitative analysis of textual data. Journal of Open Source Software, 3(30): 774. https://doi.org/10.21105/joss.00774

Blei DM, Ng AY, Jordan MI (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3: 993–1022.

Bommarito MJ, Katz DM (2022). GPT takes the bar exam. Social Science Research Network (SSRN).

Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. (2023). Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Science Reports, 13: 16492.

Callanan E, Mbakwe A, Papadimitriou A, Pei Y, Sibue M, Zhu X, et al. (2023). Can GPT models be financial analysts? An evaluation of ChatGPT and GPT-4 on mock CFA exams.

Day AJ, Fenn MK, Ravizza SM (2021). Is it worth it? The costs and benefits of bringing a laptop to a university class. PLoS ONE, 16(5): e0251792. https://doi.org/10.1371/journal.pone.0251792

delMas R, Garfield J, Ooms A, Chance B (2007). Assessing students' conceptual understanding after a first course in statistics. Statistics Education Research Journal, 6: 28–58. https://doi.org/10.52041/serj.v6i2.483

Ellis AR, Slade E (2023). A new era of learning: Considerations for ChatGPT as a tool to enhance statistics and data science education. Journal of Statistics and Data Science Education, 31(2): 128–133. https://doi.org/10.1080/26939169.2023.2223609

Fast Company (2023). The learning nonprofit: Khan academy piloting a version of GPT called KhanMigo. Fast Company.

Flesch R (1948). A new readability yardstick. Journal of Applied Psychology, 32(3): 221–233. https://doi.org/10.1037/h0057532

Google (2023). Google Gemini. https://gemini.google.com/.

Graduate Aptitude Test in Engineering (2023). Graduate aptitude test in engineering (gate). In: National-Level Examination for Engineering Graduates. Indian Institute of Technology.

Hecking T, Leydesdorff L (2018). Topic modelling of empirical text corpora: Validity, reliability, and reproducibility in comparison to semantic maps.

Hidayatullah E (2024). Evaluating the effectiveness of ChatGPT to improve English students' writing skills. Humanities, Education, Applied Linguistics, and Language Teaching: Conference Series, 1: 14–19.

Hochmann A (1986). Math teachers stage a calculated protest. The Washington Post.

Huang J, Li S (2023). Opportunities and challenges in the application of ChatGPT in foreign language teaching.

International Journal of Education and Social Science Research (IJESSR), 6(4): 75–89. https://doi.org/10.37500/IJESSR.2023.6406

Joshi I, Budhiraja R, Dev H, Kadia J, Ataullah MO, Mitra S, et al. (2023). ChatGPT in the classroom: An analysis of its strengths and weaknesses for solving undergraduate computer science questions. In: SIGCSE 2024: Proceedings of the 55th ACM Technical Symposium on Computer Science Education, volume 1.

Katz DM, Bommarito MJ, Gao S, Arredondo P (2023). GPT-4 passes the bar exam. Philosophical Transactions of the Royal Society A.

Khan Academy (2024). Four stars for KhanMigo: Common sense media rates ai tools for learning. https://blog.khanacademy.org/four-stars-for-khanmigo-common-sense-media-rates-ai-tools-for-learning-kp/. Accessed: 2024-05-06.

Koonchanok R, Pan Y, Jang H (2024). Public attitudes toward chatgpt on twitter: Sentiments, topics, and occupations. Research Square.

Kovari A, Katona J (2024). Transformative applications and key challenges of generative ai. In: 2024 IEEE 7th International Conference and Workshop Óbuda on Electrical and Power Engineering (CANDO-EPE), 89–92.

Lacey L (2024). Openai now has a gpt4o-mini. here's why that matters — cnet.com. https://www.cnet.com/tech/services-and-software/openai-now-has-a-gpt-4o-mini-heres-why-that-matters/. Accessed: 13-08-2024.

Lawrence W, Nesbitt P, Jr PHC (2024). Post-pandemic support for special populations in higher education through generative artificial intelligence. International Journal of Arts, Humanities & Social Science, 05(05). https://doi.org/10.56734/ijahss.v5n5a6

Lee D, Seung H (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401: 788–791. https://doi.org/10.1037/44565

Li S, Liu H, Bian Z, Fang J, Huang H, Liu Y, et al. (2023). Colossal-AI: A unified deep learning system for large-scale parallel training. In: Proceedings of the 52nd International Conference on Parallel Processing, ICPP '23, 766–775. Association for Computing Machinery, New York, NY, USA.

Massey PA, Montgomery C, Zhang AS (2023). Comparison of chatgpt-3.5, chatgpt-4, and orthopaedic resident performance on orthopaedic assessment examinations. Journal of the American Academy of Orthopaedic Surgeons, 31(23): 1173–1179.

McGee M, Sadler B (2024). Equity in the use of chatgpt for the classroom: A comparison of the accuracy and precision of chatgpt3.5 vs. chatgpt4 with respect to statistics and data science exams.

McLaughlin GH (1969). Smog grading-a new readability formula. Journal of Reading, 12(8): 639–646.

Meaney C, Escobar M, Stukel TA, Austin PC, Jaakkimainen L (2022). Comparison of methods for estimating temporal topic models from primary care clinical text data: Retrospective closed cohort study. JMIR Medical Informatics, 10(12). https://doi.org/10.2196/40102

Mervaala E, Kousa I (2024). Order up! micromanaging inconsistencies in chatgpt-4o text analyses. In: Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities.

Meta AI (2024). Introducing llama: A foundational, 65-billion-parameter large language model. Accessed: August, 2024.

Microsoft (2023). Microsoft copilot. https://copilot.microsoft.com/.

Newman D, Asuncion A, Smyth P, Welling M (2009). Distributed algorithms for topic models. Journal of Machine Learning Research, 10: 1801–1828.

Open AI Team (2022). ChatGPT: Optimizing language models for dialogue.

Open AI Team (2024). Hello gpt-4o. https://openai.com/index/hello-gpt-4o/. Accessed: May, 2024.

OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. (2024) GPT-4 technical report.

Özdemir B (2020). Character counter tool. https://charactercalculator.com/. Accessed: 2024-08-14.

R Core Team (2023). R: A language and environment for statistical computing.

Rijcken E, Scheepers F, Zervanou K, Spruit M, Mosteiro P, Kaymak U (2023). Towards interpreting topic models with chatgpt. In: Proceedings of the 20th World Congress of the International Fuzzy Systems Association (IFSA 2023). Presented at the 20th World Congress of the International Fuzzy Systems Association, IFSA; Conference date: 20-08-2023 – 24-08-2023.

Taloni A, Borselli M, Scarsi V, Rossi C, Scorcia V, Giannaccare G (2023). Comparative performance of humans versus gpt-4.0 and gpt-3.5 in the self-assessment program of american academy of ophthalmology. Scientific Reports, 13: 18562.

Terry OK (2023). I am a student: You have no idea how much we are using chatgpt. The Chronical of Higher Education, 69(19).

Together Computer (2023). OpenChatKit: An open toolkit and base model for dialogue-style applications.

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. (2023). Llama: Open and efficient foundation language models.

https://doi.org/10.48550/arXiv.2302.13971

Varanasi L (2023). Gpt-4 can ace the bar, but it only has a decent chance of passing the cfa exams. here's a list of difficult exams the chatgpt and gpt-4 have passed. https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1.

Wasserstein R, Lazar N (2016). The asa statement on p-values: Context, process, and purpose. American Statistician, 70(2): 129–133. https://doi.org/10.1080/00031305.2016.1154108

Watanabe K, Alexander B (2023). Seeded sequential lda: A semi-supervised algorithm for topic-specific analysis of sentences. Social Science Computer Review, 42(1): 224–248. https://doi.org/10.1177/08944393231178605

Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. (2020). Huggingface's transformers: State-of-the-art natural language processing. https://doi.org/10.48550/arXiv.1910.03771

Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, et al. (2024). Tree of thoughts: Deliberate problem solving with large language models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23. Curran Associates Inc., Red, Hook, NY, USA.