Can GPT-4 Chat Pass a Polish Stockbroker Exam?

Tomasz Wyłuda[1]

[1]Wydział Zarządzania, Uniwersytet Warszawski
Warszawa, Poland

***Abstract*— This research investigates the performance of OpenAI's GPT-4, a sophisticated language model, in passing the Polish Stockbroker Exam conducted by the Polish Financial Supervisory Authority (KNF). The exam, covering a broad range of topics, including legal issues, finance theory, finance mathematics, and setting prices, requires theoretical and practical skills pertinent to the financial markets. The study is set against various evaluations where GPT-4 and its predecessors have been tested in numerous academic and professional settings, demonstrating strengths and weaknesses in different domains. The study aimed to determine whether GPT-4 can pass the Polish Stockbroker Exam and analyze its performance across different question types. Results indicated that GPT-4 consistently failed to meet the passing score. However, it performed better when given more time per question, suggesting a trade-off between accuracy and completeness. Analysis by question type revealed higher proficiency in legal and finance theoretical questions but significant struggles with specific questions related to the stockbroker job. Notably, GPT-4 showed improvement in finance calculation questions with more response time.**

Keywords— Finance, Law, Stockbroker, Artificial Intelligence, Investment.

I. INTRODUCTION

Since OpenAI released ChatGPT and its subsequent upgrade to GPT-4, these models have been extensively tested in academic settings to evaluate their capabilities. GPT-4, a deep learning model, has shown a notable improvement in understanding and text generation compared to its predecessor, ChatGPT.

SciBench (Wang et al., 2023) tested the GPT-4 chat on a range of college-level scientific problems such as mathematics, chemistry, and physics textbooks, and problems from undergraduate-level exams in computer science and mathematics. The study reveals that current LLMs fall short of delivering satisfactory performance.

GPT-4's performance in legal examinations has been remarkable. According to the GPT-3.5 technical report by OpenAI (Achiam et al., 2023), it would pass a bar exam with a score around the top 10% of test takers. This indicates GPT-4's advanced comprehension skills in legal matters. In standardized tests, GPT-4 has shown exceptional results (Achiam et al., 2023). In the SAT Reading & Writing section, it scored 710 out of 800 (93rd percentile). In the math section, it scored 700 (89th percentile). The GPT-4 chat achieved relatively average test results (54th percentile) in the Graduate Record Examinations (GRE) writing part. However, recent papers suggest slightly lower scores (Martinez, 2023). It is worth noting GPT-4 chat would also perform decently in law schools (Blair-Stanek, 2023).

In medical profession exams, GPT chat also performed well. GPT-4 could receive a passing grade from the Japanese Medical Licensing Examination (Takagi et al., 2023). Similarly, GPT-4 was able to pass the Polish Medical Final Examination (Rosol et al., 2023), the Indian pre-medical test (Farhat et al., 2023), the German State Examination in Medicine (Jung, 2023), the Korean National Licensing Examination for Medicine Doctors (Jang et al., 2023), and Turkish Medical Specialization Exam (Kilic, 2023).

In the engineering field, GPT-4 performs decently. For instance, the AI model passed the Fundamentals of Engineering (FE) Environmental Exam (Pursnani et al., 2023). GPT-3.5 also achieved excellent results in software engineering exams prepared for students (Loubier, 2023). The AI model has demonstrated impressive capabilities in physics; the GPT chat achieved First-Class grades on an essay writing assessment from a university physics module (Yeadon, 2023). In a different

ASEJ - Scientific Journal of Bielsko-Biala School of Finance and Law
Volume 28, No 1 (2024), pages 6
https://doi.org/10.19192/wsfip.sj1.2024.10
Received: January 2024 Accepted: March 2024
Published: March 2024

research (Yeadon & Douglas, 2023), GPT-4 and GPT-3.5 answered a set of 42 exam papers derived from 10 distinct physics courses (administered at Durham University from 2018 to 2022) and scored an average of 49.4% and 38.6%, respectively. This is not a passing score; however, it suggests that GPT chat is better in a writing assessment than multiple-choice tests.

In finance and business, the GPT and AI solutions were tested as helpful tools in human capital management (Bashynska et al., 2023), auditing (Karmańska, 2022), accounting (Beerbaum, 2023), banking (Fares et al., 2023), actuary (Balona, 2023), investment (Nametala et al., 2023).

GPT was tested against college test examinations in economics, finance, and management. Chat was able to pass the Test of Understanding in College Economics (TUCE) with excellent results - the 91st percentile for Microeconomics and the 99th percentile for Macroeconomics when compared to students who take the TUCE exam at the end of their principles course (Geerling et al., 2023). In a study named "Would Chat GPT3 Get a Wharton MBA?" Christian Terwiesch (2023) stated that even the GPT-3.0 version chat would be able to receive a B to B- grade on the graded exams. However, even GPT-4 performed poorly on Quantitative Finance Examinations (Malladi, 2023).

ChatGPT model can pass major accounting certification exams, including the Certified Public Accountant (CPA), Certified Management Accountant (CMA), Certified Internal Auditor (CIA), and Enrolled Agent (EA) certification exams (Eulerich, 2023). However, GPT-4 would probably fail the Chartered Financial Analyst (CFA) Level I and II exams (Callanan et al., 2023). The study conducted by a collaboration of researchers from Queens University, Virginia Tech, and J.P. Morgan's AI research division highlighted GPT-4's enhanced understanding of complex financial concepts, although it demonstrated more difficulty with Level II content.

Overall, GPT chat was tested against a wide range of college-level tests and standardized certification tests. Research shows that AI solutions can be valuable tools in education. However, the performance in the financial education and college-level finance examinations still needs improvement. Knowledge of financial topics requires a combination of reasoning, logic, and advanced mathematics skills.

The research on applying the tool in business education and Polish financial education remains limited. These diverse assessments of GPT models in academic and professional settings highlight their strengths in processing and generating complex information across various domains. While many papers examined GPT chat's performance, the subject is new and evolving. Thus, there is a research gap, especially in testing Polish examinations such as the Polish Stockbroker Exam administered by the Polish Financial Supervisory Authority (KNF). The study aims to answer the questions:

1) Can the GPT-4 chat pass the Polish Stockbroker Exam?
2) How does the model perform with different types of questions in the exam?
3) How does GPT-4 perform in different circumstances?

II. METHODOLOGY

The stockbroker's exam includes detailed thematic areas covering legal issues in civil law, commercial law, tax, foreign exchange law, and aspects related to securities and other financial instruments. It also addresses public offerings, trading in financial instruments, financial accounting, and ethical standards in the profession. This structured approach ensures that prospective stockbrokers are well-versed in theoretical knowledge and practical skills essential for their role in the financial markets.

To assess whether GPT chat could pass the securities stockbroker exam, a methodology akin to that employed in previously referenced studies was utilized. For this purpose, ten exams from previous years published on the KNF website were selected. The exams occurred between March 25, 2018, and October 15, 2023 (KNF, n.d.). The exam consists of 120 test questions, with a total duration of 3 hours. There are four possible answers, but only one is correct. For each correct answer, two points are awarded, and for each incorrect answer, one point is deducted. No points are given for unanswered questions. To pass the exam, one needs to score 160 points.

Two methods were used to test how GPT-4 chat would perform in the exam. In the first method, GPT-4 chat received all questions at once; however, in the second method, the chat received question by question (not the whole set at once).

In the first method, the entire securities stockbroker exam was passed, preceded by the instructions:

"The securities stockbroker exam is a single-choice test consisting of 120 questions. To pass the exam, a minimum score of 160 points is required. The scoring system for the exam is as follows:

Correct Answer: +2 points

Incorrect Answer: -1 point

No Answer: 0 points

Your task is to answer the questions and receive a minimum of 160 points."

In this approach, GPT chat was presented with the whole test at once and then proceeded to solve the tasks by choosing one out of four correct answers. Each exam was solved in a separate instance of GPT chat to avoid interactions between the already completed tests. Before passing the test, the rules were explained to the GPT chat.

In contrast, the second method involved presenting GPT chat with instructions on how to solve the test, followed by pasting questions one by one. This method allowed GPT chat more time to respond to each question. However, it is essential to note that the total response time taken by GPT chat was still below the time limit set for the actual test conducted by the Financial Supervision Authority (3 hours).

After the calculation, we conducted a t-statistics test with 99% confidence to determine whether the GPT-4 chat can pass the test.

- Null Hypothesis (H0): The test taker's average score is equal to or greater than the points required to pass.
- Alternative Hypothesis (H1): The test taker's average score is less than the points required to pass.

The research findings were subsequently verified, and the number of correct answers, incorrect answers, and refusals to respond were tallied. Additionally, the questions in the test were categorized into different sections to ascertain whether GPT chat exhibited similar proficiency across all types of questions. Four categories were distinguished: law, finance-theory, finance-mathematical tasks, and specific knowledge (KNF, 2024):

- The legal tasks encompassed a range of topics, including civil law, commercial law, tax and foreign exchange law, issues related to securities and other financial instruments, matters concerning public offerings and public companies, issues related to trading in financial instruments, matters concerning supervision over the financial and capital markets, issues related to the creation and functioning of investment companies and funds as well as management of alternative investment funds, matters concerning the commodity market exchange, issues related to the settlement-depository system, the Accounting Act, and accounting issues.
- The second category, finance theory, included theoretical topics (not requiring calculations). These issues covered financial mathematics, analysis and valuation of debt instruments, financial analysis of enterprises and stock valuation, analysis of derivative instruments, and investment strategies.
- The third category covered the same range of material as the second category but required mathematical calculations.
- The fourth category of questions pertained explicitly to the work of a stockbroker, including stock exchange and over-the-counter trading, setting prices of listed financial instruments, professional ethics, and prevention of crimes in the capital market.

III. Results and Discussion

This study aimed to assess the performance of the GPT chat system in the Polish Stockbrokers' examination over multiple iterations spanning from March 25, 2018 to October 15, 2023. The evaluation metrics included the number of correct and incorrect answers, questions not answered, questions canceled by the Polish Financial Supervisory Authority (KNF), total points achieved, and the points required for passing. The results of the first testing method are presented below.

TABLE 1. GPT-4's Results in Solving the Stockbroker Exam Over the Years - First Testing Method

Exam date	Correct answers	Wrong answers	Questions not answered	Questions canceled by KNF	Total points achieved by GPT chat	Points required	Result
15 October 2023	66	54	0	0	78	160	Fail
19 March 2023	68	52	0	0	84	160	Fail
9 October 2022	73	47	0	0	99	160	Fail
27 March 2022	72	48	0	0	96	160	Fail
20 June 2021	63	57	0	0	69	160	Fail
13 September 2020	63	56	0	1	70	158	Fail
27 October 2019	61	58	0	1	64	158	Fail
24 March 2019	65	54	0	1	76	158	Fail
21 October 2018	63	54	0	3	72	154	Fail
25 March 2018	61	55	0	4	67	152	Fail
Average	65,5	53,5	0,0	1,0	77,5	158,0	
Median	64,0	54,0	0,0	0,5	74,0	159,0	
Standard deviation	4,1	3,4	0,0	1,3	11,4	2,7	

Source: own calculation.

The sample mean is 77.5 points, the required point to pass 138,0 points, and the sample standard deviation is 11.4 points. The calculated t-statistic is approximately -16.78, and the critical t-value for a 99% confidence level in a one-sided test with 9 degrees of freedom is approximately -2.82. Since the absolute value of the t-statistic is greater than the absolute value of the critical t-value ($|16.78| > |2.82|$), we reject the null hypothesis. This indicates that there is significant evidence at the 99% confidence level to conclude that the test taker's average score is significantly lower than the points required to pass. It indicates that GPT-4's score would not be sufficient to pass the Stockbroker Exam.

A closer examination of the performance metrics reveals that GPT-4 did not pass any of the 10 exams. Moreover, the interesting is the strategy taken by the test taker. There were no instances where questions were left unanswered by GPT chat in any of the exams. However, there were instances of questions being canceled by the Polish Financial Supervisory Authority (KNF), ranging from none in the earlier exams to a maximum of four questions in the March 2018 exam.

Following the results, the second testing method was implemented. In the second series of tests where GPT chat was asked each question individually. The results show a distinct pattern compared to the first test series where GPT chat was asked to answer the entire test in one go. This second approach, spanning from March 2018 to October 2023, still resulted in GPT chat failing to meet the required threshold for passing the Polish Stockbrokers exam, yet it demonstrates a noteworthy change in performance dynamics. Below the results of the second testing method are presented.

TABLE 2. GPT-4'S RESULTS IN SOLVING THE STOCKBROKER EXAM OVER THE YEARS - SECOND TESTING METHOD

Exam date	Correct answers	Wrong answers	Questions not answered	Questions canceled by KNF	Total points achieved by GPT chat	Points required	Result
15 October 2023	69	44	7	0	94	160	Fail
19 March 2023	70	44	6	0	96	160	Fail
9 October 2022	76	38	6	0	114	160	Fail
27 March 2022	73	37	10	0	109	160	Fail
20 June 2021	66	46	8	0	86	160	Fail
13 September 2020	67	46	6	1	88	158	Fail
27 October 2019	64	48	7	1	80	158	Fail
24 March 2019	66	45	8	1	87	158	Fail
21 October 2018	64	44	9	3	84	154	Fail
25 March 2018	64	40	12	4	88	152	Fail
average	67,9	43,2	7,9	1,0	92,6	158,0	
median	66,5	44,0	7,5	0,5	88,0	159,0	
standard deviation	3,9	3,5	1,9	1,3	10,4	2,7	

Source: own calculation.

The calculated t-statistic with the updated data is approximately -13.80, and the critical t-value for a 99% confidence level in a one-sided test with 9 degrees of freedom is approximately -2.82. Similar to the previous analysis, since the absolute value of the t-statistic is greater than the t-value ($|13.80| > |2.82|$), we reject the null hypothesis. This indicates that there is significant evidence at the 99% confidence level to conclude that the test taker's average score is significantly lower than the points required to pass, even with the updated data. It indicates that GPT-4's score would not be sufficient to pass the Stockbroker Exam.

The total points achieved by GPT chat in the second testing method are higher than in the first method. For instance, in October 2023, GPT chat scored 94 points as opposed to 78 points in the first testing method. This trend of increased scoring is consistent across all test dates, suggesting that providing more time for each question positively impacts GPT chat's performance.

The number of correct answers in the second testing method is higher compared to the first testing method. Similarly, there is a noticeable decrease in the number of wrong answers, with the second testing method recording a lower count of incorrect responses compared to the first testing method.

However, we can observe that the GPT chat took a different strategy in answering questions. In the first testing method, the AI model answered all questions. On the other hand, the number of unanswered questions in the second testing method ranged from 6 to 12 across different exam dates. This factor could be attributed to the altered testing methodology, where GPT chat might have taken more time to consider each question, leading to some questions being left unanswered within the given timeframe.

In conclusion, while the altered testing approach in the second series improved the GPT chat's total points, it also introduced the occurrence of unanswered questions. Despite these performance improvements, the AI model was still unable to pass the Polish stockbrokers' examination.

TABLE 3. GPT-4'S RESULTS IN SOLVING THE STOCKBROKER EXAM OVER THE YEARS BY QUESTION TYPE - FIRST TESTING METHOD

Question type	Share of questions by type (%)	Share of correct answers (%)	Share of wrong answers (%)	Share of questions not answered (%)
Legal	25	68	32	0
Finance Theoretical	33	69	31	0
Finance Calculation	18	52	48	0
Specific Knowledge	24	27	73	0

Source: own calculation.

TABLE 4. GPT-4'S RESULTS IN SOLVING THE STOCKBROKER EXAM OVER THE YEARS BY QUESTION TYPE -SECOND TESTING METHOD

Question type	Share of questions by type (%)	Share of correct answers (%)	Share of wrong answers (%)	Share of questions not answered (%)
Legal	25	66	27	8
Finance Theoretical	33	66	26	7
Finance Calculation	18	72	19	9
Specific Knowledge	24	26	70	4

Source: own calculation.

The analysis of GPT chat's performance on the Polish Stockbrokers exam, categorized by question type, reveals distinct patterns and variations between the two methods of testing.

In the first method, where GPT chat was provided the full test at once. GPT chat demonstrated relatively high proficiency in Legal Questions (25% of the total), with 68% correct answers and 32% wrong answers. Similarly, the AI model showed proficiency in Finance Theoretical Questions (33% of the total) with 69% correct answers and 31% wrong answers. The performance of the chat was mediocre in Finance Calculation Questions (18% of the total), with only 52% correct answers and 48% wrong answers. Specific Knowledge Questions (24% of the total) were the most challenging for GPT chat, with only 27% correct answers and a high 73% wrong answers.

In the second method of testing, where GPT chat was given questions one at a time, GPT chat showed similar performance in

legal questions (a slight decrease in correct answers to 66%, with wrong answers at 27% and 8% of questions not answered). Similarly, in finance theoretical questions, GPT chat performed well (a slight decrease to 66% correct answers, 26% wrong answers, and 7% not answered). However, GPT chat increased

performance in finance calculation questions (72% correct answers, 19% wrong answers, and 9% not answered). As in the first testing method, GPT chat performed poorly in specific knowledge questions (with 26% correct answers, 70% wrong answers, and 4% not answered).

In comparison, the second testing method revealed a notable effect on the performance of GPT chat. This method demonstrated an improvement in GPT chat's ability to answer financial calculation questions. However, during the first testing method, in which GPT chat was presented with the entire set of questions, the AI model committed a significant error. In certain calculation questions, GPT chat attempted to retrieve the answers from its memory instead of performing the calculations. Consequently, GPT chat incorrectly interpreted the task required to respond to the question. Giving GPT chat more time to analyze each question might be particularly beneficial for complex calculation-based questions.

However, the overall performance in the specific knowledge category, particularly concerning Polish laws and regulations, remained consistently low across both testing methods. This indicates an ongoing challenge in this area. The questions in this category demanded not only specific knowledge but also the ability to perform complex tasks, such as setting appropriate prices and executing correct orders for buying or selling stocks.

The second testing method also revealed that GPT chat employed a new strategy. Unlike the first method, where GPT chat attempted to answer all questions, the second method left some questions unanswered. This change could be attributed to the AI model giving more thoughtful consideration to each question when time constraints were less pressing.

In summary, although the alteration in testing methodology did result in improvements in specific categories, it also underscored the limitations of GPT chat in consistently and comprehensively responding to questions across various types of content featured in the Polish Stockbroker exam.

IV. Conclusions

This comprehensive study aimed to evaluate the performance of GPT chat, specifically the GPT-4 model, in passing the Polish Stockbrokers' examination. The assessment was conducted through two distinct methodologies over multiple iterations between 25 March 2018 and 15 October 2023. The results offer several key insights into the capabilities and limitations of GPT-4 in this professional context.

The study revealed that GPT-4 performed better when given more time to answer each question individually. This approach led to a higher number of correct answers and a notable decrease in wrong answers compared to the first method, where the entire test was presented at once. However, this improvement in accuracy came at the cost of an increased number of unanswered questions, suggesting a trade-off between accuracy and completeness.

Despite the observed improvements in certain aspects, GPT-4 consistently failed to achieve the passing score in all iterations of the Polish Stockbrokers' examination. This underperformance highlights the model's limitations in fully grasping and applying the specialized knowledge and analytical skills required for this specific professional certification.

The analysis of performance based on question type unveiled distinct patterns. GPT-4 showed relatively higher proficiency in legal and finance theoretical questions, but struggled significantly with specific knowledge questions. Interestingly, the model demonstrated a marked improvement in finance calculation questions when given more time, underscoring its potential in handling complex, calculation-based queries.

These findings emphasize the potential and constraints of AI applications like GPT-4 in professional and academic fields. While GPT-4 shows promise in understanding and processing complex information, its application in passing professional certifications like the Polish Stockbrokers' examination is currently limited. This suggests that while AI can be a valuable tool for learning and preliminary analysis, it cannot yet replace the nuanced understanding and decision-making skills of human professionals.

The study underscores the need for ongoing research and development in AI. Improvements in AI models, particularly in their ability to handle specialized, context-specific information and in decision-making under time constraints, could enhance their applicability in professional certifications and other complex tasks.

In conclusion, the study of GPT-4's performance in the Polish Stockbrokers' examination provides valuable insights into the current capabilities of AI in complex, professional settings. While there are notable strengths, particularly in processing and analyzing information, the limitations in achieving the required proficiency for professional certification indicate the need for further advancements in AI technology. This exploration serves as a critical step in understanding and shaping the future role of AI in professional and educational domains.

V. References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Balona, C. (2023). ActuaryGPT: Applications of large language models to insurance and actuarial work. Available at SSRN 4543652.

Bashynska, I., Prokopenko, O., & Sala, D. (2023). Managing Human Capital with AI: Synergy of Talent and Technology. Zeszyty Naukowe Wyższej Szkoły Finansów i Prawa w Bielsku-Białej, 27(3), 39-45.

Beerbaum, D. O. (2023). Generative Artificial Intelligence (GAI) with Chat GPT for Accounting–a business case. Available at SSRN 4385651.

Blair-Stanek, A., Carstens, A. M., Goldberg, D. S., Graber, M., Gray, D. C., & Stearns, M. L. (2023). GPT-4's Law School Grades: Con Law C, Crim C-, Law & Econ C, Partnership Tax B, Property B-, Tax B. Crim C-, Law & Econ C, Partnership Tax B, Property B-, Tax B (May 9, 2023).

Callanan, E., Mbakwe, A., Papadimitriou, A., Pei, Y., Sibue, M., Zhu, X., ... & Shah, S. (2023). Can gpt models be financial analysts? an evaluation of chatgpt and gpt-4 on mock cfa exams. arXiv preprint arXiv:2310.08678.

Eulerich, M., Sanatizadeh, A., Vakilzadeh, H., & Wood, D. A. (2023). Is it All Hype? ChatGPT's Performance and Disruptive Potential in the Accounting and Auditing Industries. SSRN Electronic Journal.

Fares, O. H., Butt, I., & Lee, S. H. M. (2023). Utilization of artificial intelligence in the banking sector: A systematic literature review. Journal of Financial Services Marketing, 28(4), 835-852.

Farhat, F., Chaudry, B. M., Nadeem, M., Sohail, S. S., & Madsen, D. O. (2023). Evaluating AI models for the National Pre-Medical Exam in India: a head-to-head analysis of ChatGPT-3.5, GPT-4 and Bard. JMIR Preprints.

Geerling, W., Mateer, G. D., Wooten, J., & Damodaran, N. (2023). ChatGPT has aced the test of understanding in college economics: Now what?. The American Economist, 05694345231169654.

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Medical Education, 9(1), e45312.

Jang, D., Yun, T. R., Lee, C. Y., Kwon, Y. K., & Kim, C. E. (2023). GPT-4 can pass the Korean National Licensing Examination for Korean Medicine Doctors. PLOS Digital Health, 2(12), e0000416.

Jung, L. B., Gudera, J. A., Wiegand, T. L., Allmendinger, S., Dimitriadis, K., & Koerte, I. K. (2023). ChatGPT passes German state examination in medicine with picture questions omitted. Deutsches Ärzteblatt International, 120(21-22), 373.

Karmańska, A. (2022). Artificial Intelligence in audit. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 66(4), 87-99.

Kilic, M. E. (2023). AI in Medical Education: A Comparative Analysis of GPT-4 and GPT-3.5 on Turkish Medical Specialization Exam Performance. medRxiv, 2023-07.

Loubier, M. (2023). ChatGPT: A Good Computer Engineering Student?: An Experiment on its Ability to Answer Programming Questions from Exams.

Malladi, R. K. (2023). Emerging Frontiers: Exploring the Impact of Generative AI Platforms on University Quantitative Finance Examinations. arXiv preprint arXiv:2308.07979.

Martínez, E. (2023). Re-Evaluating GPT-4's Bar Exam Performance. Available at SSRN 4441311.

Nametala, C. A., Souza, J. V. D., Pimenta, A., & Carrano, E. G. (2023). Use of econometric predictors and artificial neural networks for the construction of stock market investment bots. Computational Economics, 61(2), 743-773.

Pursnani, V., Sermet, Y., Kurt, M., & Demir, I. (2023). Performance of ChatGPT on the US fundamentals of engineering exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. Computers and Education: Artificial Intelligence, 5, 100183.

Rosoł, M., Gąsior, J. S., Łaba, J., Korzeniewski, K., & Młyńczak, M. (2023). Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Scientific Reports, 13(1), 20512.

Takagi, S., Watari, T., Erabi, A., & Sakaguchi, K. (2023). Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Medical Education, 9(1), e48002.

Terwiesch, C. (2023). Would chat GPT3 get a Wharton MBA. A prediction based on its performance in the operations management course.

Wang, X., Hu, Z., Lu, P., Zhu, Y., Zhang, J., Subramaniam, S., ... & Wang, W. (2023). Scibench: Evaluating college-level scientific problem-solving abilities of large language models. arXiv preprint arXiv:2307.10635.

Yeadon, W., Inyang, O. O., Mizouri, A., Peach, A., & Testrow, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. Physics Education, 58(3), 035027.

Yeadon, W., and Douglas P. Halliday. "Exploring durham university physics exams with large language models." arXiv preprint arXiv:2306.15609 (2023).

Polish Financial Supervision Authority (n.d.). Examinations for Securities Brokers. Available at: https://www.knf.gov.pl/dla_rynku/egzaminy/Maklerzy_papierow_wartosciowych_egzaminy/testy (Accessed: January 28, 2024).

KNF, Examination Commission for Securities Brokers (2023). Communication No. 4 regarding the thematic scope of the exam for securities brokers and the skills test [Announcement No. 4 on the subject scope of the securities broker exam and skills test]. Available at: https://www.knf.gov.pl/knf/pl/komponenty/img/Komunikat_4_2023_87292.pdf (Accessed: January 25, 2024).

Minister of Finance Regulation on examinations for securities brokers and investment advisors and the skills test (2016) Journal of Laws (Dziennik Ustaw), 707, Poz. 707.