KIER DISCUSSION PAPER SERIES

KYOTO INSTITUTE
OF
ECONOMIC RESEARCH

Discussion Paper No.1103

“How good are LLMs in risk profiling?”

Thorsten Hens and Trine Nordlie

April 2024

KYOTO UNIVERSITY
KYOTO, JAPAN

How good are LLMs in risk profiling?

April 15, 2024

Thorsten Hens[1,2,3]

Trine Nordlie[2]

1 Department of Finance, University of Zurich, Plattenstrasse 32, CH-8032 Zurich, Switzerland.

2 Department of Finance, Norwegian School of Economics, NHH, Helleveien 30, 5045 Bergen.

3 Institute of Economic Research, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan.

Abstract

This study compares OpenAI's ChatGPT-4 and Google's Bard with bank experts in determining investors' risk profiles. We find that for half of the client cases used, there are no statistically significant differences in the risk profiles. Moreover, the economic relevance of the differences is small.

JEL Classification D8, D14, D81, G 51

Keywords Large Language Models, ChatGPT, Bard, Risk Profiling

Introduction

Large Language Models (LLMs) are increasingly prevalent in the financial sector and are likely to have an even greater influence over the long term (Ankenbrand et al., 2023). The possibilities for applications of LLMs in financial advisory is an emerging field now attracting attention in research (cf. Biswas et al., 2023). Increasingly powerful LLM based chatbots, such as OpenAI's ChatGPT and Google's Bard, represent the latest developments in natural language processing and are currently being used and explored in a wide range of applications (Ankenbrand et al., 2023).

Risk profiling is a crucial aspect of financial advice. It builds the base for the strategic asset allocation which is known to be the most important determinant of the investment success. As the studies of Brinson, Hood, & Beebower (1995) and Brinson, Singer, & Beebower (1991) have shown, the strategic asset allocation determines 90% of the investment success. Therefore, to ensure optimal financial advice, regulators in Europe and Switzerland, as mandated by the MiFID and FIDLEG laws, require that professionals conduct risk profiles during the practice of financial advice to ensure suitability and protection for the client.

Creating investment portfolios and making investment decisions requires a deep understanding of the individual investor. While ChatGPT and Bard have showed promising theoretical potential (Guo et al., 2023), their capabilities in comprehending and assessing investor risk profiles are not yet clear. Within the academic discourse, it is widely acknowledged that LLM based systems have certain inherent limitations. For instance, they are not necessarily output consistent (Ankenbrand et al., 2023) and are susceptible to producing hallucinations, information that is incorrect or untrue (Ji et al., 2023; Su et al., 2022). Despite their power and transformative potential in many applications, the complexities and intricacies of these tools give rise to specific constraints that must be acknowledged and addressed in their implementation and use. For these reasons, it is relevant to study how such systems perform in risk profiling, and to what extent they can be effectively applied. This study investigates the accuracy of the current iterations of ChatGPT and Bard in categorizing individual risk profiles for investors.

Methodology

This section presents the research design implemented in the study. It further describes the procedures employed for data collection and analysis.

Client Cases

To study chatbot performance in investor risk profiling, ten different client cases were provided by Amstein (2023). The cases encompassed a variety of investor descriptions, each detailing information about financial situation, investment objectives, risk preferences and knowledge and experience. While there were variations in the specific details like age, profession, investment objectives, the cases shared consistent overarching features for comparability. The investor descriptions were characterized by brief and direct statements, reflecting real-world tendencies of investors to have limited information about their risk tolerance and related factors. The client cases can be found in the online Appendix A.

Data Collection

The data t from the bankers was collected by Amstein (2023) based on an online survey in the fall 2023. The bankers were all employed at the same bank in Switzerland which has operations nationwide. They were incentivised to score the risk profiles of the 10 clients. The number of risk scores for each client differed; clients 1 through 4 were categorized 49 times, clients 5 through 9 were categorized 48 times, and client 10 was categorized 47 times. This variation in frequency was due to participant attrition during the survey. The banker were familiar with the categorization scale which ranged from 1 to 5. The bankers assigned categories using whole numbers within this range.

The data from the LLMs was collected weekly between October 7th and November 25th, 2023. ChatGPT and Bard each categorized the ten clients a total of 16 times over time. To achieve comparability, the LLMs were asked the same questions as the bankers. We required the LLMs to: «Categorize the investor between one and five indicating the order: 1 lowest to 5 highest possibility to take risk.» Although the chatbots were requested to provide similar categorical responses, they in some instances provided non-integer answers such as "2 or 3" or "2 leaning towards 3". In such cases, the mean value of their suggested ranges was used for quantitative analysis.

Data Analysis

As the analytical objectives were clear and pre-defined, the data analysis was confirmatory, founded on testing hypotheses regarding the efficacy of chatbots in categorizing investor risk profiles. The underlying hypothesis is that a client is randomly matched with a banker and that the point in time a LLM would be applied is also random. Thus, in both cases the client faces some randomness.

Table 1 presents the descriptive statistics for the risk scores assigned to each client by Bard, ChatGPT, and bankers. The average score (mean) and median were used as measures of central tendency of the scores for each client by assessors. Minimum and maximum scores measure the spread. The descriptive measures for the bankers were based on individual assessments by each banker. In contrast, the measures for the chatbots were derived from all evaluations conducted throughout the period.

Table 1

Descriptive Statistics for Each Client by Assessor; Mean (average score), median, minimum (min) and maximum (max) score

	Bard				ChatGPT				Bankers			
Client	Mean	Median	Min	Max	Mean	Median	Min	Max	Mean	Median	Min	Max
Client 1	2.5	2.8	1.0	3.0	2.3	2.0	2.0	3.0	2.5	2.0	1.0	5.0
Client 2	4.3	4.0	3.0	5.0	4.5	4.5	4.0	5.0	4.8	5.0	4.0	5.0
Client 3	2.5	2.8	1.0	3.0	2.1	2.0	1.5	2.5	3.0	3.0	2.0	4.0
Client 4	1.4	1.0	1.0	3.0	1.5	1.5	1.0	2.0	1.5	1.0	1.0	3.0
Client 5	3.2	3.0	3.0	4.0	3.2	3.0	3.0	4.0	3.3	3.0	2.0	5.0
Client 6	4.4	4.0	3.0	5.0	4.4	4.0	4.0	5.0	3.2	3.0	1.0	5.0
Client 7	3.1	3.0	3.0	4.0	3.8	4.0	3.0	4.0	3.3	3.0	1.0	4.0
Client 8	3.5	3.0	3.0	5.0	4.0	4.0	3.5	4.0	4.4	5.0	2.0	5.0
Client 9	4.8	5.0	4.0	5.0	4.7	5.0	4.0	5.0	4.9	5.0	4.0	5.0
Client 10	3.4	3.0	3.0	5.0	3.3	3.0	3.0	4.0	3.2	3.0	2.0	5.0

Note. Bard and ChatGPT have N = 16, Bankers N = 48 (mostly)

When comparing the assessments from Bard, ChatGPT, and the bankers across clients, the average scores generally indicated a degree of conformity, with disparities predominantly fluctuating by less than one risk score between the assessors (Table 1). An exception was observed for client 6, where both chatbots' risk scores were marginally higher, exhibiting an average of 4.4 and a median of 4, in contrast to bankers, which had an average of 3.2 and a median of 3. Further, other more substantial discrepancies were observed for clients 2, 3 and 8, for which bankers categorized slightly higher risk scores than chatbots', and for client 7, for which ChatGPT determined a higher risk score than the other assessors.

The spread, representing the difference between the maximum and minimum scores, generally indicates the level of consensus or lack thereof, within the group of assessors. For bankers, a small spread would indicate a strong consensus on a client's risk score, while for chatbots this would indicate consistent risk assessment over time. Bard's suggested risk profile scores frequently covered an interval of two or three scores, indicating that its assessments varied within two or three rating points on the scale when assessing each client. ChatGPT's scores for each client appeared relatively consistent, varying by two scores or less, indicating more consistency in the evaluations over time. With a range of scores equal or smaller than one for all clients, ChatGPT demonstrated a certain degree of self-consistency in assessing the clients.

Compared to the chatbots, the bankers' risk profiling scores for the ten client profiles were more variable. The overall larger spread indicates that the group of bankers is less consistent in their evaluation of the client according to the rating scale. For clients 1 and 6, bankers answers cover the full range of scores. This suggests that there was significant variation in opinions among the bankers. The fact that some clients received both the lowest and highest scores suggests that bankers' perceptions of risk vary considerably from one banker to another, or that the investor profiles were diverse enough to justify a wide range of assessments, underscoring the nuanced perception of risk profiles for the clients.

It must be noted that the number of observations within each assessor group potentially impacted the measured ranges. There were more observations from the bankers (N=483) compared to each chatbot (N=160). A larger number of observations could potentially, to some extent, account for a broader range observed in the bankers' evaluations.

The hypothesis assessed in the first analysis was to test if there were any statistically significant differences between the assessments of ChatGPT, Bard, and bankers for each of the ten investors.

Welch’s t-tests were conducted to compare ChatGPT and bankers, and Bard and bankers, for each client. Holm-Bonferroni correction was applied to the p-values to reduce the risk of false significant results due to the large number of tests. Welch’s t-test was chosen over Student’s t-test due to its reliability when the two samples have unequal variances and sample sizes (Ruxton, 2006). The assumption of normal distribution was not met. Although the test in practice can be relatively robust against deviations from normality as long as the distribution is reasonably symmetric, i.e. the distribution is not skewed (West, 2021), it should be considered that the smaller sample sizes, especially from the chatbots, likely influence the precision of the test. This consideration is a general statistical principle, as the robustness to these violations as discussed by West (2021) is under conditions that do not explicitly address the complexities introduced by small sample sizes. In the online Appendix C alternative tests were conducted to assess if they yielded similar results compared to Welch’s t-tests: Kruskal-Wallis and Dunn’s post hoc tests, which were more appropriate given the data. But the results were basically the same.

Results

Statistical differences

The Welch’s t-test revealed significant differences in the scores between ChatGPT and bankers for several clients (Table 2). Specifically, the test revealed: for Client 3, t(43) = 8.60, p<.001; for Client 6, t(61) = -5.53, p<.001; for Client 7, t(60) = -3.86, p = .004; and for Client 8, t(54) = 3.89, p = .004.

Table 2
Results from Welch't t-tests Comparing ChatGPT and Bankers

							95% CI	
Client	*t*	df	Unadjusted *p*	Adjusted *p* (Holm)	M Banker	M ChatGPT	*LL*	*UL*
Client 1	1.32	57	.191	1.000	2.5	2.3	-0.11	0.55
Client 2	2.88	22	.008	0.105	4.8	4.5	0.1	0.63
Client 3	8.60	43	p<.001	p<.001	3.0	2.1	0.65	1.05
Client 4	-0.42	36	.679	1.000	1.5	1.5	-0.36	0.24
Client 5	0.65	46	.522	1.000	3.3	3.2	-0.2	0.39
Client 6	-5.53	61	p<.001	p<.001	3.2	4.4	-1.59	-0.75
Client 7	-3.86	60	p<.001	0.004	3.3	3.8	-0.77	-0.25
Client 8	3.89	54	p<.001	0.004	4.4	4.0	0.22	0.68
Client 9	1.71	20	.102	1.000	4.9	4.7	-0.05	0.49
Client 10	-0.66	51	.514	1.000	3.2	3.3	-0.4	0.21

Table 3 presents the results of Welch’s t-tests for the assessment scores between Bard and bankers for each client. Significant differences were found for Client 6, t(51) = -4.98, p <.001, and Client 8, t(27) = 4.29, p = .003.

Table 3

Results from Welch's t-tests Comparing Bard and Bankers

							95% CI	
Client	*t*	df	Unadjusted *p*	Adjusted *p* (Holm)	M Banker	M Bard	*LL*	*UL*
Client 1	0.00	38	.998	1.000	2.5	2.5	-0.41	0.41
Client 2	3.28	18	.004	.058	4.8	4.3	0.21	0.96
Client 3	3.02	21	.006	.084	3.0	2.5	0.16	0.86
Client 4	0.16	23	.877	1.000	1.5	1.4	-0.39	0.45
Client 5	1.15	53	.256	1.000	3.3	3.2	-0.12	0.43
Client 6	-4.98	51	p<.001	p<.001	3.2	4.4	-1.64	-0.7
Client 7	1.87	61	.066	.731	3.3	3.1	-0.02	0.5
Client 8	4.29	27	p<.001	.003	4.4	3.5	0.48	1.35
Client 9	0.83	22	.414	1.000	4.9	4.8	-0.14	0.33
Client 10	-1.02	33	.317	1.000	3.2	3.4	-0.58	0.19

Economic Relevance

In this section we compute the economic relevance of the differences found above. We follow DeGiorgi and Hens (2009) but only use a mean-variance utility (cf. Markowitz ,1952).

$$U^i(\lambda^i) = (\mu - r_f)\lambda^i - \frac{\gamma^i}{2}\sigma^2(\lambda^i)^2$$

To keep things simple, we assume the clients,$i = 1, \ldots, 10$, can invest in a risky asset with expected return $\mu = 7\%$ and volatility $\sigma = 20\%$ -- which are typical numbers for equity indices. Alternatively, they can invest risk-free at $r_f = 2\%$. Each risk profile RP^i results in an asset allocation λ^i. To fix ideas, we assume $\lambda^i = 0.25 * (RP^i - 1)$. Thus, the lowest (highest) risk profiles results in 0% (100%) equity. Also, we assume that the bankers get the risk profile right – on average. Based on the average risk profile of the bankers we can determine the risk aversion of the clients, $\gamma^i = \frac{(\mu - r_f)}{0.25*(RP^i - 1)\sigma^2}$. This puts us in a position to compare the utility loss – measured in returns – that results from deviations of the LLMs to this average: $U^i(\lambda^c) - U^i(\lambda^i)$, where c indicates the asset allocation from the risk profile of the LLM c assessed on some time period. Table 4 shows the utility losses for the two LLMs for each client– as well as the losses averaged over clients. The median losses are 12 basis points for Bard and 11 basis points for ChatGPT. To put this in perspective notice that the average cost of advice (all-in fee, total expense ratio of products, ticket fees, …) in private banking in Switzerland is 1%, i.e. 10 times as much. Thus – on average the differences between the bankers and the LLMs models are not economically relevant. But – one might also be unlucky and the LLM gets is totally wrong in which case the utility loss would be 87 basis points on average and 3.12% in one instance for Bard!

Table 4	Diff Bard				Diff ChatGPT			
utility	mean	median	max	min	mean	median	max	min
0.92%	-0.15%	-0.12%	-0.92%	0.00%	-0.09%	-0.09%	-0.12%	0.00%
2.40%	-0.13%	-0.11%	-0.55%	0.00%	-0.05%	-0.02%	-0.11%	0.00%
1.24%	-0.20%	-0.04%	-1.24%	0.00%	-0.26%	-0.30%	-0.69%	-0.07%
0.29%	-0.66%	-0.29%	-3.12%	-0.29%	-0.27%	-0.29%	-0.37%	0.00%
1.40%	-0.03%	-0.02%	-0.16%	-0.02%	-0.04%	-0.02%	-0.16%	-0.02%
1.34%	-0.55%	-0.21%	-1.01%	-0.01%	-0.50%	-0.21%	-1.01%	-0.21%
1.42%	-0.03%	-0.02%	-0.15%	-0.02%	-0.12%	-0.15%	-0.15%	-0.02%
2.08%	-0.22%	-0.33%	-0.33%	-0.02%	-0.03%	-0.02%	-0.13%	-0.02%
2.36%	-0.03%	-0.01%	-0.10%	-0.01%	-0.04%	-0.01%	-0.10%	-0.01%
1.30%	-0.14%	0.00%	-1.10%	0.00%	-0.06%	0.00%	-0.25%	0.00%
average diff	-0.21%	-0.12%	-0.87%	-0.04%	-0.15%	-0.11%	-0.31%	-0.04%

Conclusion

This study asked “How do ChatGPT and Bard categorize investor risk profiles compared to financial advisors?”

For half of the clients the study revealed no statistically significant differences in the risk scores assigned by ChatGPT and Bard compared to those assigned by bankers. Moreover, on average, the differences had minor economic relevance.

Certainly, this was just one – but the first – study to assess those differences. Further studies should be based on client advisors from a different bank, and they should repeat the risk scoring of the LLMs since they improve over time. Moreover, one should look more deeply in the explanations LLMs give for their risk scores as Nordlie (2024) has done. For a client it is not sufficient to know her risk profile but to hold through an investment strategy, the client needs to understand why this risk profile is most suitable for her. Nordlie (2024) sheds some doubt on the performance of LLMs in this respect.

References

Amstein, A. (2023). *Financial Incentives in Risk Profiling: An Experimental Study*. Master Thesis, Department of Finance, University of Zurich.

Ankenbrand, T., Bieri, D., Reichmuth, L., Stengel, C., Wickihalder, S.and Yilmaz, E. (2023). *GPT for Financial Advice*. The Combination of Large Language Models and Rule-Based Systems. IFZ, HSLU, Lucerne.

Biswas, S., Joshi, N., & Mukhopadhyaya, J. (2023). *ChatGPT in Investment Decision Making: An Introductory Discussion*. https://doi.org/10.13140/RG.2.2.36417.43369

Brinson, G. P., Hood, L. R., & Beebower, G. L. (1995). *Determinants of portfolio performance*. Financial Analysts Journal, 51(1), 133–138.

Brinson, G. P., Singer, B. D., & Beebower, G. L. (1991). *Determinants of portfolio performance II: An Update*. Financial Analysts Journal, 47(3), 40–48.

De Giorgi, E. and Th. Hens (2009). *Prospect theory and mean-variance analysis: does it make a difference in wealth management?* Investment Management and Financial Innovations, Volume 6, Issue 1.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). *How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. arXiv*. https://doi.org/10.48550/arXiv.2301.07597

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). *Survey of Hallucination in Natural Language Generation*. ACM Computing Surveys, *55*(12), 1–38. https://doi.org/10.1145/3571730

Klement, J. (2015). *Investor Risk Profiling: An Overview.* SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2597691

Markowitz, H. (1952). *Portfolio Selection***.** The Journal of Finance, *7*(1), 77–91. https://doi.org/10.2307/2975974

Nordlie, T. (2004). *Evaluating ChatGPT-4 and Bard in Categorizing Investor Risk Profiles A Study on the Accuracy and Consistency of Chatbot Assessments.* Master thesis, Economics and Business Administration, Financial Economics. NorwegianSchool of Economics, Bergen.

Ruxton, G. D. (2006). *The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test*. Behavioral Ecology, *17*(4), 688–690. https://doi.org/10.1093/beheco/ark016

Su, D., Xiaoguang, L., Zhang, J., Shang, L., Jiang, X., Liu, Q., & Fung, P. (2022). *Read before Generate! Faithful Long Form Question Answering with Machine Reading.* Findings of the Association for Computational Linguistics, 744–756. https://doi.org/10.48550/arXiv.2203.00343

West, R. M. (2021). *Best practice in statistics: Use the Welch t-test when testing the difference between two groups.* Annals of Clinical Biochemistry, *58*(4), 267–269. https://doi.org/10.1177/0004563221992088

Online Appendices

Appendix A: Client Cases

Client 1:

Age: 24

Profession: Student (Economics, beginning of Master's)

Part-time job at the department approximately: 1800.- CHF per month

Fortune: 350,000 CHF

I inherited 350k from my father and am unsure how to manage this money. My goal is to keep most of this money as a backup until I enter the workforce (about 3 years from now). I would also like to put this wealth aside for later and invest it for the long term so that I can start planning to have children without any worries (approximately in 10 years). My rent for my shared Apartment is 700.- per month, which I would like to cover with the inheritance. Since the house will be demolished in about a year, I'm concerned I won't be able to find such an affordable shared apartment again, as I don’t want to leave my location in the city. I would like to finance my remaining expenses through my part-time job, but this income is not always guaranteed because I sometimes work more and sometimes less. It would be great, of course, if I achieve an increase in wealth, but what's most important to me is that I still have a significant portion of the wealth at the end of my studies. I have some knowledge about investments, but I don't have the confidence to manage the money myself. I also attach great importance to ESG criteria.

Client 2:

Age: 33

Status: Single

Profession: Lawyer (250k annual income)

Fortune: 500,000 CHF

I am completely focused on my career and do not plan to start a family in the future. I would like to emigrate at 50 and buy a house in South America to enjoy my retirement there. I would like to invest my assets for this purpose. My current wealth amounts to 500,000. -, but I regularly set aside larger sums per month and would like to invest them as well. Currently, I am renting a place in the city for 2600.- CHF per month, and my other monthly expenses range from 5000.- to 7000.-. The goal of my investment is substantial wealth growth, so that I can approach my retirement plans without financial worries. I am aware that fluctuations in value are part of the process, and I can handle them as long as they do not become extreme. I personally believe in the future of cryptocurrencies and am not hesitant to make a significant investment in this asset class.

Client 3:

Age: 37

Occupation: Architect

Assets: 150,000 CHF in savings plus a house (valued at 600,000, with a 200k mortgage)

We are a young family consisting of two children (5 & 6 years old) and us parents. We own a small architectural firm and both work full-time. In addition to our everyday expenses, we have a mortgage on our house, which still amounts to 200,000. All in all, we can cover all our monthly expenses with our salaries and occasionally set aside some money for holidays. The workload at our architectural firm seems to be secure for the next few years, which is why we would like to invest our savings of 150,000. - to secure the academic education of our children. A decent increase in wealth would be nice, but we are not willing to take big risks. We are somewhat skeptical about the financial world, but admittedly, we ourselves have little knowledge. If this money predominantly flows into national companies, we would be perfectly happy.

Client 4:

Age: 65

Occupation: Retired, former owner of a painting business

Assets: 1.6 million CHF + House valued at 1.1 million CHF (no mortgage)

I have just sold my painting business and retired. I am extremely sceptical about the world of banks and do not trust them, especially after the media reports of recent years. However, my children have advised me to invest the money, which is why I am turning to you. I live in my fully paid-off house in the countryside, which is valued at approximately 1.1 million. Furthermore, I will now start receiving my pension fund, which covers my everyday expenses well. If any costs arise with the house, I must be able to cover them myself. Since the house has recently been renovated, this should not be a major problem for the time being. In my old age, I might consider moving to a nice retirement apartment, and the house would then be up for sale. I would like to invest my wealth of 1.6 million so that my two children and their families can inherit well later. It is important to me that I do not incur any losses and that my hard-earned money is preserved.

Client 5:

Age: 29

Occupation: Professional Football Player

Assets: 2.1 million CHF / Salary: 1,100,000 CHF per annum

I am currently playing football in the Super League and I am a regular player in my club. Since one never knows how long a football career lasts, I want to build security for the future and invest my money. My average monthly expenses for rent and daily life amount to around 12,000.-. I try to save as much as possible, but there are always larger expenses that come up. Currently, I still have a contract until the summer of 2025 and I believe it should be extended if my health situation doesn't change. My investment goal is to generate income and wealth growth, so that after my football career, I can pursue an education to become a psychologist without financial worries. I am well aware of the risks of losses, as there is no possibility for long-term financial planning in my profession. Additionally, it is extremely important to me that my investments are ESG compliant, as being a public figure comes with media attention and I do not want to cause any unnecessary controversy. Once my football career is over, I will likely need to use parts of my wealth for my livelihood, and I do not want to completely change my lifestyle during the transitional phase.

Client 6:

Age: 23

Occupation: Student (Electrical Engineering at ETH)

Assets: 50,000 CHF

I am currently starting my Master's program at ETH and still live at home. My parents are currently providing full support for me, which means I have no regular expenses in any form. Currently, I have about 50,000 in cash myself. My goal is to start a tech start-up with friends after completing my studies, for which each of us would need to contribute about 80,000 initially. We are very confident that our idea will be successful and therefore, I am willing to take this risk. Now, I would like to invest my 50,000 to get closer to this goal. I am aware of the risks of fluctuating values and am willing to take them in order to achieve increased value appreciation.

Client 7:

Age: 52

Marital Status: Divorced, with two adult children

Occupation: Consulting in the insurance industry

Assets: 200,000 CHF

Income: 11,000 CHF per month

After years of paying alimony to my ex-wife, our children are now grown up and have completed their education, which means I now have more money available for myself. I want to save and invest my money for my retirement. My investment horizon is therefore my retirement age. Since I am relatively knowledgeable in the financial industry, I am fully aware that high value appreciation comes with more risk. However, I am willing to accept this to achieve a 3-5% annual return. I do not plan to touch my assets but would like to make monthly contributions if possible. At some point, I would like to afford a nice home and move out of my city apartment. This could be in a year or in ten years. For now, that would be the only reason why I would need liquid assets.

Client 8:

Age: 40

Marital Status: Married, with one child

Occupation: Owner of a restaurant chain

I am the founder and CEO of a restaurant chain, and up until now, I have invested my wealth in a bank's defensive fund. Now, I would like to have more control and overview of my investments, so I am asking you to restructure my investment portfolio. Specifically, I would like to allocate a significant portion of my money into impact investment solutions. My liquid assets, including the money in a fund, amount to 2.1 million CHF. Additionally, I still hold a 10% stake in my privately held, non-publicly traded restaurant chain. My wife and our 8-year-old daughter live in our house, which is 60% mortgaged and valued at 1.6 million CHF, located by Lake Zurich. Over the past few years, after deducting our expenses, we have been able to set aside approximately 200,000 from my share in the restaurant chain. Since I do not anticipate facing financial difficulties in the future, I can invest a significant portion of this wealth for the long term. Fluctuations in value are not a problem if I can ethically justify the investment. It would be nice, though, if we could soon pay off our mortgage.

Client 9:

Age: 30

Occupation: Entrepreneur

Assets: 1.8 million CHF

I have just sold my start-up, which I invested a lot of blood and sweat into, for a low seven-figure amount. My main interest is maximizing my financial returns. While I don't have a specific investment goal in mind, my primary objective is to achieve significant profits and impressive growth with my investments. My knowledge of financial markets and investment products is limited, but I am willing to learn and make informed decisions. I am aware of the risks, I am inclined towards risk-taking, and I am ready to explore both risky and rewarding investment strategies. I have heard about the potential gains from cryptocurrencies and would like to allocate a significant portion of my portfolio to this exciting asset class. I am open to discussions about various investment opportunities, including stocks, bonds, real estate, and alternative investments, as long as they have the potential for significant returns. The sooner, the better.

Client 10:

Age: 52

Occupation: Entrepreneur

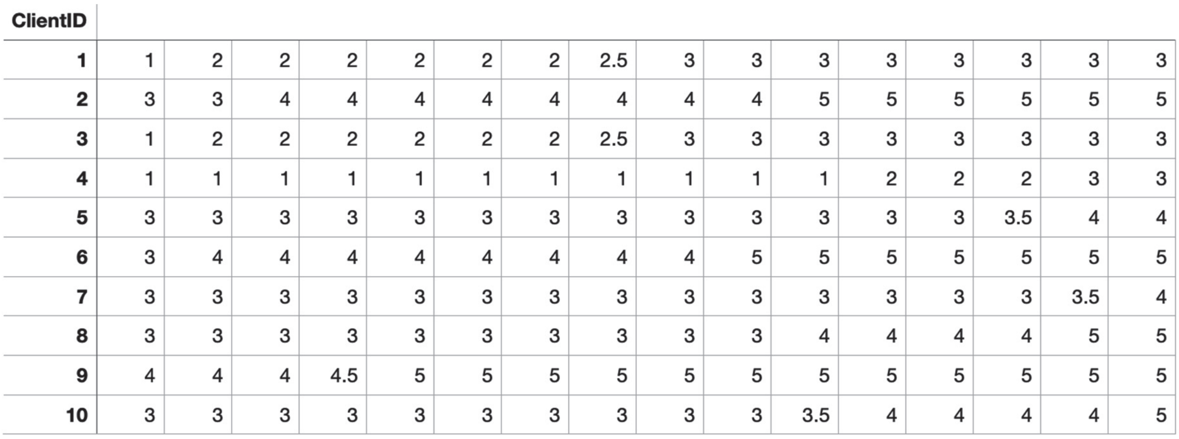
I have been successfully running a business, holding various properties, and investing consistently in start-ups/companies. Now is the time when I would like to take a step back to spend more time with my two children, as this has been somewhat neglected in recent years. I am now in search of a successor for my company and would like to entrust the management of my portfolio to your hands. My wealth is in the eight-figure range and consists of shares in companies (approximately 2.5 million CHF), real estate holdings (14 million CHF), and my investment portfolio (6 million CHF). The goal should be steady wealth growth, allowing me to acquire additional properties in the future. I have always followed a 'safety first' approach and carefully selected my investments to minimize losses, especially in times of increasing market uncertainties! However, I have also taken risks with real estate purchases when my intuition advised me to do so. My investment strategy should cover both

Value Investing and Growth Investing and may also include smaller amounts in speculative investments. There should always be enough liquid assets available, as I do not hesitate when it comes to purchasing lucrative properties.

Appendix B: Raw Data

Table E1

Matrix of Scores from Bard for Each Ten Clients (ClientID)

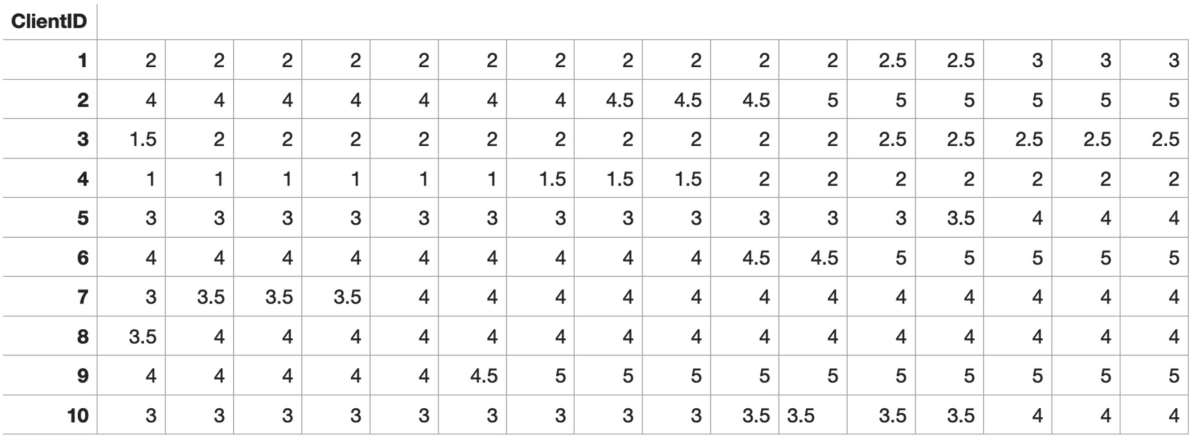


ClientID																
1	1	2	2	2	2	2	2	2.5	3	3	3	3	3	3	3	3
2	3	3	4	4	4	4	4	4	4	4	5	5	5	5	5	5
3	1	2	2	2	2	2	2	2.5	3	3	3	3	3	3	3	3
4	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	3
5	3	3	3	3	3	3	3	3	3	3	3	3	3	3.5	4	4
6	3	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5
7	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3.5	4
8	3	3	3	3	3	3	3	3	3	3	4	4	4	4	5	5
9	4	4	4	4.5	5	5	5	5	5	5	5	5	5	5	5	5
10	3	3	3	3	3	3	3	3	3	3	3.5	4	4	4	4	5

Note. Columns are scores from 1 to 5.

Table E2

Matrix of Scores from ChatGPT for Each Ten Clients (ClientID)

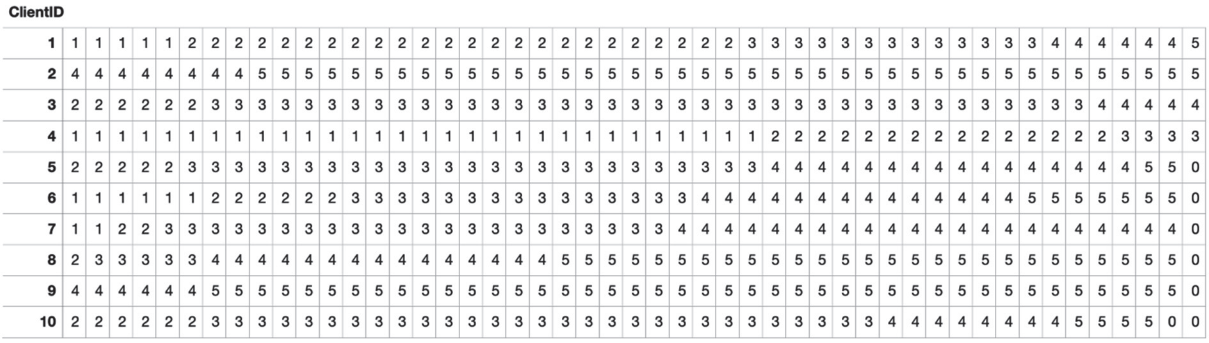


ClientID																
1	2	2	2	2	2	2	2	2	2	2	2	2.5	2.5	3	3	3
2	4	4	4	4	4	4	4	4.5	4.5	4.5	5	5	5	5	5	5
3	1.5	2	2	2	2	2	2	2	2	2	2	2.5	2.5	2.5	2.5	2.5
4	1	1	1	1	1	1	1.5	1.5	1.5	2	2	2	2	2	2	2
5	3	3	3	3	3	3	3	3	3	3	3	3	3.5	4	4	4
6	4	4	4	4	4	4	4	4	4	4.5	4.5	5	5	5	5	5
7	3	3.5	3.5	3.5	4	4	4	4	4	4	4	4	4	4	4	4
8	3.5	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
9	4	4	4	4	4	4.5	5	5	5	5	5	5	5	5	5	5
10	3	3	3	3	3	3	3	3	3	3.5	3.5	3.5	3.5	4	4	4

Note: Columns are scores from 1 to 5.

Table E3

Matrix of Scores from Bankers for Each of Ten Clients (ClientID)



ClientID																																																		
1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	5
2	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
3	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3
5	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	5	5	0
6	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	0
7	1	1	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	0
8	2	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	0
9	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	0
10	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	5	5	0	0

Note. Each column represents scores from 1 to 5 given by different bankers, with each client being evaluated by 47 to 50 bankers in total. Individual banker evaluations for each client cannot be tracked across the table.

Appendix C: Alternative Statistical Tests

Shapiro-Wilk normality tests, in addition to visual assessment of Quantile-Quantile (Q-Q) plots, were conducted to check the normality assumption within the data. Results from the Shapiro-Wilks tests indicated significant deviations from normality across the tested datasets for each client. For all tests, p<.001, suggesting non-normality. For some of the datasets, plots of residuals versus fitted values indicated a bit of heterogeneity, suggesting that the data were not equally spread across the dataset. The failure to meet the assumptions, a common occurrence across the ten distinct datasets, made the Kruskal-Wallis test, a non-parametric alternative to one-way ANOVA, more appropriate for conducting the analyses.

Table C1 presents the Kruskal-Wallis test results for each of the clients. For client 1, the test did not show a significant difference in scores by assessor, $\chi^2_{(2)}$ = 1.19, p = 0.551. Similarly, no significant difference was found in the assessments for clients 4 ($\chi^2_{(2)}$ = 1.19, p = 0.551), 5 ($\chi^2_{(2)}$ = 0.96, p = 0.620), 9 ($\chi^2_{(2)}$ = 4.61, p = 0.100) and 10 ($\chi^2_{(2)}$ = 2.21, p = 0.331).

Table C1

Kruskal-Wallis test statistics

Client	Kruskal-Wallis χ2	df	*p*
Client 1	1.19	2	0.551
Client 2	17.49	2	p<.001
Client 3	34.21	2	p<.001
Client 4	1.19	2	0.551
Client 5	0.96	2	0.620
Client 6	20.81	2	p<.001
Client 7	16.59	2	p<.001
Client 8	19.92	2	p<.001
Client 9	4.61	2	0.100
Client 10	2.21	2	0.331

Conversely, significant differences were observed for assessments of clients 2, $\chi^2_{(2)}$ = 17.49, p<.001; Client 3, $\chi^2_{(2)}$ = 34.21, p<.001; Client 6, $\chi^2_{(2)}$ = 20.81, p<.001; Client 7, $\chi^2_{(2)}$ = 16.59, p<.001; and Client 8, $\chi^2_{(2)}$ = 19.92, p<.001. These significant results indicated that there were differences in scores attributed by assessors for these clients.

The Kruskal-Wallis test results suggested that there were statistically significant differences in the assessments among ChatGPT, Bard, and bankers for five of the ten clients. Results from the post hoc analysis using Dunn's test are shown in Table 3 in the paper, presenting the test statistics for the specific groups that exhibited significant differences.

Table C2

Results from Dunn's test for significant Kruskal Wallis test results

Client	Comparison	*z*	Unadjusted *p*	Adjusted *p* (Holm
Client 2	Banker - Bard	3.72	p<.001	p<.001
	Banker - ChatGPT	2.92	.004	.007
	Bard - ChatGPT	-0.58	.565	.565
Client 3	Banker - Bard	2.91	.004	.007
	Banker - ChatGPT	5.63	p<.001	p<.001
	Bard - ChatGPT	2.21	.003	.027
Client 6	Banker - Bard	-3.59	p<.001	p<.001
	Banker - ChatGPT	-3.62	p<.001	p<.001
	Bard - ChatGPT	-0.02	.984	.984
Client 7	Banker - Bard	2.12	.034	.034
	Banker - ChatGPT	-2.84	.005	.009
	Bard - ChatGPT	-4.05	p<.001	p<.001
Client 8	Banker - Bard	4.22	p<.001	p<.001
	Banker - ChatGPT	2.47	.013	.027
	Bard - ChatGPT	-1.42	.150	.154