Norwegian School of Economics
Bergen, Spring, 2024

Evaluating ChatGPT-4 and Bard in Categorizing Investor Risk Profiles

A Study on the Accuracy and Consistency of Chatbot Assessments

Trine Nordlie

Supervisor: Thorsten Hens

Master thesis, Economics and Business Administration, Financial Economics

NORWEGIAN SCHOOL OF ECONOMICS

Acknowledgement

First, I would like to thank my supervisor, Thorsten Hens, for guidance and for suggesting the topic of my thesis. I would furthermore like to thank Alvin Amstein who designed the client cases and collected the data in the bank. This thesis could not have been carried out without you.

Abstract

This study aims to examine the performance of OpenAI's ChatGPT-4 and Google's Bard in categorizing investors' risk profiles. The objectives were to compare the chatbots' assessments with those of financial advisors and to assess the consistency of their evaluations over time. Two research questions were stated: "How do ChatGPT and Bard categorize investor risk profiles compared to financial advisors?" and "How consistent are the chatbots' categorizations over time?"

The study included ten distinct investor descriptions (client cases), which the chatbots were asked to categorize weekly from October 7th through November 25th, 2023. The assessments of ChatGPT and Bard were compared with those from financial advisors from the same bank.

To compare assessments from ChatGPT, Bard, and the bankers, multiple Kruskal-Wallis tests were conducted, followed by Dunn's tests for post hoc analysis. Additionally, Welch's t-tests were used as an adjunct methodological measure to validate the results, checking the consistency of findings across the statistical analyses, even under varying data assumptions. A qualitative analysis of the chatbots' responses was conducted in instances where their assessments deviated from those of the bankers with statistical significance. A repeated measures ANOVA was used to assess the chatbot's consistency.

The results from the non-parametric tests indicated that ChatGPT's and Bard's assessments differed from those of bankers for half of the clients. Among these clients, both chatbots assessed the client's risk profiles more conservatively for three and higher for one. Furthermore, the results indicated that the chatbots were relatively consistent in their assessments over time. Despite some variations in their assessed risk scores for each client, these variations were relatively minor and did not indicate notable inconsistencies. The consistency of the chatbots was also supported by a lack of statistically significant difference in their assessments based on the conversational method used in the study.

The qualitative analysis revealed several weaknesses in the chatbots' reasoning, affecting their accuracy. These limitations included a lack of personalized recommendations, reliance on general principles, absence of factual support, and a lack of human-like understanding. These results highlight the importance of cautious reliance on such tools for risk profiling, especially considering the identified weaknesses in reasoning.

Table of Contents

1. Introduction

Large Language Models (LLMs) are increasingly prevalent in the financial sector and are likely to have a greater influence over the long term (Bieri et al., 2023). The possibilities for applications of LLMs in financial advisory is an emerging field now attracting attention in research. 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 (Bieri et al., 2023). ChatGPT reached 100 million users in just a few months after its launch on November 30th 2022, making it the fastest growing costumer application to date (Liu et al., 2023). ChatGPT and Bard have caught significant interest in various global forums, such as the World Economic Forum, where the importance of these technologies for the financial industry has been prominently highlighted. However, research on potential applications of chatbots like ChatGPT and Bard is still at its early stages.

Risk profiling is a crucial aspect of investment advice, and creating investment portfolios and making investment decisions require a deep understanding of the individual investor. The applicability of GPTs in risk profiling has seen limited research to date. 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 out-put consistent (Bieri 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 thesis aims to study the extent to which the current iterations of ChatGPT and Bard demonstrate accuracy and consistency in categorizing individual risk profiles for investors. The following research questions reflect the main objectives of the study:

> Research question 1 (RQ1): "How do ChatGPT and Bard categorize investor risk profiles compared to financial advisors?"

Research question 2 (RQ2): “How consistent are the chatbots’ categorizations over time?”

The first research question attempts to evaluate the precision with which ChatGPT and Bard categorize investor risk profiles, comparing their suggested risk scores to those of financial advisors from the same bank. The second research question aims to study the consistency with which these chatbots categorize investors’ risk profiles over time. The current thesis explores how such systems perform in categorizing investors’ risk profiles, and its contribution is to assess whether OpenAI’s ChatGPT-4 and Google’s Bard are suited for such rule-based case studies, and thus to explore the potential of combining LLMs with rule-based systems for investment advice.

This thesis consists of six chapters. Chapter 1 gives an introduction along with research questions. Chapter 2 provides a literature review including the theoretical background relevant for the study. Chapter 3 describes the methodology used in the study. Chapter 4 presents the results from the analysis, which are further discussed in chapter 5, including a discussing of limitations. The conclusion is presented in chapter 6.

2. Literature Review

This section provides an overview of theoretical background relevant for the research questions.

2.1 Introduction to ChatGPT and Bard

Conversational AI solutions have recently received significant attention due to new developments in natural language processing (NLP) tools. NLP is a type of AI that uses algorithms to understand and generate human-like responses (Deng & Lin, 2023). Such NLP tools are commonly referred to as "chatbots". ChatGPT by Open AI was launched in November 2022 (OpenAI, 2022), and Bard from Google in February 2023 (Pichai, 2023). Both are natural language processing systems designed to generate human-like conversations by understanding the context of a conversation and generating appropriate responses (Deng & Lin, 2023).

ChatGPT-4 is the latest version of OpenAI's GPT series of language models. It builds on the previous version, GPT-3, with improvements in comprehension, context processing, and text generation, making it more advanced in handling dialogues and answering questions (OpenAI, 2023). ChatGPT-4 is a specific application of GPT-4 technology, which is a multimodal model that can process both text and image inputs (OpenAI, 2023). Although it is less capable than humans in many real-world scenarios, it demonstrates human-level performance on various professional and academic benchmarks (OpenAI, 2023).

Google's Bard is also designed to simulate human-like interactions, answering a wide range of user inquiries and prompts (Pichai, 2023). The chatbot has been trained on a large dataset of text data, and utilizes language models to generate comprehensive and informative responses to user inputs (Pichai, 2023). Bard's functionality is based on its large language model called the Pathways Language Model 2 (PaLM 2), which is designed to comprehend facts and logical reasoning (Pichai, 2023).

2.2 Application of Chatbots in Financial Advisory

Artificial intelligence (AI) is evolving in the financial industry. Recent literature suggests positive results in the efficiency and precision of AI applications compared to human advisors within financial advisory. Huang & Rust (2018) suggest that in specific areas of financial

planning, particularly in processing data, risk assessments, and portfolio management, AI has the potential to equal or even exceed the capabilities of human advisors. This assertion is grounded in their theoretical framework, based on conceptual examination rather than empirical evidence.

Biswas et al. (2023) discussed how AIs like ChatGPT are likely to influence the investment decision-making process. Their discussion is theoretical, and the insights do not stem from empirical evidence but rather suggest directions for future research. Zaremba & Demir, (2023), referred to in their article, argued that NLP like ChatGPT can transform the future of finance. They claimed that by leveraging the power of ChatGPT, financial institutions could gain competitive advantage by improving their decision-making capabilities. Regarding potential applications in investment advice, individual investors could receive tailored financial advice through ChatGPT based on their risk tolerance, investment objectives and other considerations (Biswas et al., 2023). By processing a big amount of data, it can provide individualized investment recommendations (Biswas et al., 2023). ChatGPT may also respond to inquiries concerning asset allocation, portfolio diversification, and investment methods understanding the sentimental aspect of the investor (Biswas et al., 2023). Among different aspects of the investment decision-making process, they also mentioned that ChatGPT can assist investors in managing their portfolios and may assist in risk management by examining portfolios, detecting possible hazards, and providing advice on how to reduce those risks.

The implementation of AI-driven tools in investment advice within financial services necessitates adherence to various regulatory requirements (Bieri et al., 2023). These regulations include the Markets in Financial Instruments Directive (MiFID) in the EU and Financial Services Act (FinSA) in Switzerland. Some key obligations under regulations according to Swiss law involve fulfilling information duties, assessing client appropriateness and suitability, acting in clients' best interests, and maintaining confidentiality. The use of sensitive data raises privacy concerns and challenges in compliance with existing regulations, particularly given risks in relation to potential discrimination, bias, systematic manipulation of human actions, limited transparency of decisions, and the complicated issues associated with criminal responsibility or civil liability when using AI. Such considerations will not be discussed further in this thesis. (Bieri et al., 2023)

To the best of my knowledge, this thesis presents the first instance of empirical evidence applying Large Language Models (LLMs) for investor risk profiling. The thesis is presented with the understanding that it contributes to the exploration of LLMs within the context of risk

profiling, while acknowledging the possibility of other related research that may not have been discovered.

2.3 Limitations of Chatbot Characteristics

Conversational AI tools like ChatGPT and Bard have certain inherent limitations. To understand their operational dynamics and limitations, their foundational technology must be considered. Conversational AI tools are based on Large Language Models. A large language model (LLM) is a language model consisting of a neural network with billions of parameters trained on a large amount of unannotated data through self-supervised learning and is used, for example to predict and generate text and other content (Sejnowski, 2023). Generative Pre-Trained Transformer (GPT) is a type of LLM. Bieri et al. (2023) provides a description of the underlying architecture and behavior of GPTs:

> *[…]current GPTs attempt word-by-word text generation sequentially, based on the probability distributions of words and phrases from the training set, to give coherent responses to an input (e.g., a question). As a result, such models are probabilistic in nature, implying that the model's responses can be sensitive to details of the wording or phrasing of the prompts (Bubeck et al., 2023). A similar question can therefore be answered differently, depending on the structure of the corresponding input. Therefore, GPTs are not necessarily always output-consistent.*

Considering ChatGPT's and Bard's capabilities to categorize investors' risk profile, the probabilistic nature of GPTs holds significant implications. As such chatbots can be sensitive to nuances in the phrasing and structuring of the prompt, how the investor information and the assessment request are stated may influence the responses the chatbots provide. If variability in their responses is present, this could pose challenges as it might undermine the reliability of chatbots in assessing investors. Due to the probabilistic nature, even with similar or identical input, there's a possibility that the chatbots may generate varied responses. Inconsistent output in terms of the chatbots wording and phrasing can be compared to human advisors who express themselves differently. However, what is crucial is that the assessment remain consistent. It is decisive for the reliability of such tools that investors receive consistent assessments that do not vary depending on how the request is phrased. Potential variability in a model's output can present significant challenges in a field where consistency and accuracy are paramount, which brings us to the central research questions of this study.

Biswas et al. (2023) highlight areas that could be considered as red flags for investors to rely on them when using ChatGPT in investment decision making. They emphasize that the

accuracy of ChatGPT's output is dependent on the quality of the input data. They state that if the data used to train the model is inadequate, inconsistent, or biased, the output may be flawed, potentially leading to inaccurate investment decisions for the investor. Therefore, investors should be aware of the quality of the data being used (Biswas et al., 2023). The data on which ChatGPT and Bard is trained includes non-peer reviewed information which could be inaccurate. Biswas et al. (2023) also highlight that an overreliance on ChatGPT's output, without considering other relevant aspects, can lead to suboptimal investment decisions.

LLMs are prone to generating information that is not factual or true, a phenomenon commonly referred to as "hallucination" (Ji et al., 2023; Su et al., 2022). Such false responses that can be generated are currently a concern with applications of LLMs. Like other LLMs, both ChatGPT and Bard suffers from the problem of hallucination. ChatGPT comes with OpenAI's disclaimer that it may occasionally write plausible-sounding but incorrect or nonsensical answers (Biswas et al., 2023). GPT-4 is still not fully reliable as it "hallucinates" facts and makes reasoning errors (OpenAI, 2023). Their overconfidence in producing factually incorrect information is a potential limitation affecting their reliability (Skopeliti & Milmo, 2023). There are two types of hallucinations (Ji et al., 2023). "Intrinsic hallucination" is generated output that contradicts the input content. "Extrinsic hallucination" is when generations cannot be verified from the input content (i.e., output that can neither be supported nor contradicted by the source). The occurrence of hallucination is particular of concern in financial advisory contexts, where accuracy and trustworthiness are crucial. If hallucinations occur when assessing investor risk profiles that could lead to misclassifications or inappropriate financial advice, this could have serious implications in real-world contexts.

Another limitation of LLM based tools is that they might struggle with common-sense reasoning tasks (Li et al., 2021). Additionally, the reliability of chatbots such as ChatGPT and Bard may be affected by their inability to fully comprehend the complexity of human language and conversations (Bogost, 2022). Since such chatbots generate words sequentially based on a given prompt, they are uncapable of truly understanding the meaning behind words, implying that responses generated are likely to be shallow and lacking in depth and insight (Bogost, 2022). The inherent characteristics and potential pitfalls of GPT-based models directly inform the context and significance of the research questions in this study.

2.4 Risk Profiling

This section provides an overview of concepts of risk profiling, laying the foundation to examine the accuracy of ChatGPT and Bard in comprehending and categorizing investors risk profiles.

2.4.1 Risk Profiling from a Regulatory Perspective

There are various regulations and standards financial advisors must adhere to. Rooted in a broader regulatory ambition to improve the competitiveness of financial markets and ensure investor protection, regulators require investment firms to establish investor risk profiles before recommending financial products or investments (Klement, 2015). To ensure that investors are offered products that are suitable, every major developed financial marked has instituted suitability rules (Klement, 2015).

In the EU, MiFID II applies. According to Article 25 (2) of MiFID II (2014), investment firms are required to obtain necessary information from clients when providing investment advice or portfolio management. It follows that information must be obtained regarding the client's knowledge and experience relevant to the specific type of product, their financial situation including the ability to bear losses, and their investment objectives, including risk tolerance. This is required to ensure that investment services and financial instruments are suitable, in accordance with risk tolerance and ability to bear losses. FinSA in Switzerland, align several Swiss regulations with international standards. FinSA Section 3, Article 10 (2018), stipulates that "financial service providers that provide investment advice or portfolio management services shall perform an appropriateness or suitability review".

Both MiFID and FinSA require a suitability assessment when providing investment services. Risk profiling falls under the requirements related to the suitability assessment. The regulations emphasize the importance of comprehensively understanding a client's risk profile. While MiFID II directly emphasizes risk tolerance, FinSA approaches the concept implicitly through the requirement of a suitability assessment. This necessitates understanding an investors inclination to bear potential losses, which subtly references their risk tolerance. The assessment of risk tolerance and risk ability is essential in this process, ensuring that investment products align with the individual investor's emotional comfort level and financial resilience. While risk tolerance is featured prominently in such regulations, it is not clear how to measure it or how it influences the range of suitable investments (Klement, 2015). The

regulations provide only general rules, that the advice should suit the risk ability and risk tolerance of a client.

2.4.2 Risk Profiling Concepts

Markowitz (1952) introduced Modern Portfolio Theory (MPT), which is a method for constructing a diversified portfolio to maximize returns without accepting an unacceptable level of risk. He presented the concept of an efficient frontier, a range of optimal portfolio allocations comprising various assets. This efficient frontier is established based on the assets' returns, variances, and covariances. Investors can select a portfolio from this efficient frontier that aligns with their individual preferences for risk and return (Elton & Gruber, 1997). To construct an optimal portfolio allocation, it is important to correctly assess the individual investor's appropriate level of risk, considering both their willingness and ability to bear financial risk.

Risk profiling for investors is a fundamental concept in providing investment advice. Assessing an individual's risk profile is crucial for determining their appropriate asset allocation (Roszkowski & Davey, 2010). The term "risk profile" is frequently used by financial advisors and in financial literature. However, there is no universally accepted definition or standardized and agreed practice for risk profiling. The criteria for creating an adequate and accurate risk profile remain a subject of debate, primarily due to its complex nature. However, an investor's risk profile typically comprises a combination of relatively stable, objective and subjective factors (Nobre & Grable, 2015).

Risk capacity, as defined by Cordell (2001), refers to a client's financial ability to take on risk and is an objective element of an investor's risk profile. Cordell identifies multiple factors influencing this capacity, including portfolio goals and constraints, income, expenses, balance sheet, and financial obligations. The time horizon is a key factor; younger clients often have a higher risk capacity due to their longer time to recover from market downturns. Brayman et al. (2017) point out that financial advisors have typically used time horizon as an indicator of risk capacity. Those with a shorter time horizon, holding other factors constant, should take less risk than those with a longer time horizon, because clients with a short time horizon have less time to recover from losses (Brayman et al., 2017).

The amount and stability of an investor's income is significant: higher, stable incomes generally allow for greater risk-taking due to increased capital for investment and loss absorption (Cordell, 2001). Conversely, those with fluctuating incomes should be more cautious. Additionally, the nature of expenses, whether fixed or discretionary, affects risk

capacity (Cordell, 2001). The amount of expenses relative to income should be considered. A favourable balance sheet, characterized by lower debt and a good mix of fixed and discretionary expenses, also indicates a higher capacity for risk (Cordell, 2001). Financial obligations, such as family and retirement commitments, are important considerations in assessing an investor's ability to withstand financial risks. Accordingly, an investor with fewer financial obligations typically has a higher risk capacity compared to one with significant debts and fixed expenses.

Financial advisors commonly use the term "Risk tolerance" in relation to a client's willingness to take risks (Nobre & Grable, 2015). Grable (2000) defines financial risk tolerance as "the maximum amount of uncertainty that someone is willing to accept when making a financial decision". The concept's counterpart, "risk aversion", is also frequently used (Nobre & Grable, 2015). Unlike risk capacity, which is an objective element, risk tolerance is more subjective, and refers to an individual's level of comfort with uncertainty and potential financial losses (Nobre & Grable, 2015).

2.4.3 Methods of Risk Profiling

Risk profiling typically includes four steps: collecting information from the client, evaluating the collected information, constructing the portfolio for the client, and reporting (Hens et al., 2018, pp. 147-148). Traditional methods of risk profiling have mostly relied on questionnaires and interviews. These methods aim to assess the investor's attitudes toward risk, their past investment experiences, and reactions to hypothetical market scenarios (Grable & Lytton, 1999).

Traditional methods of risk profiling that relies on questionnaires to quantify the clients risk tolerance and capacity, have multiple limitations. (Rice (2005) examined questionnaires used by investment firms in the United States and found that the number of questions ranged from 1 to 49, with 11% of the questionnaires directly requesting the investor to choose a risk profile or portfolio themselves. He further found that the scoring and assignment of responses from questionnaires were often subjectively conducted by the advisor. Additionally, the scoring process often overlooked the importance of certain queries over others and disregards the investor's confidence in their responses to specific queries. Furthermore, he found that the suggested asset allocations tended to benefit the investment firm more than the investor.

Foerster et al. (2014) discovered that the advisor's influence was a more significant factor in the composition of investor portfolios than the factors typically evaluated in questionnaires. These findings support the idea that traditional methods of risk profiling using

questionnaires are not fully valid and are influenced by subjectivity. Considering these insights, different advisors might assess the same investor's risk profile differently, potentially leading to varied assessments from different advisors.

Another challenge advisors face is that investors often have multiple and vaguely defined goals, complicating the task of creating a clear and quantitative investment strategy (Klement, 2015). As an example, a common situation may involve an investor who wants to save for retirement in ten years, while also saving for a down payment on a house and maintaining liquid assets as emergency funds (Klement, 2015). Other limitations relate to the subjectivity of the clients self-reported data, which can lead to potential inconsistencies and biases in the risk assessment (Roszkowski & Davey, 2010).

The use of AI and machine learning in financial advisory is transforming traditional approaches, allowing for the implementation of more data-driven methods in assessing clients risk profiles. Hens et al. (2018) outlined different risk profiling methodologies, including "artificial intelligence techniques", or "machine learning techniques" (p. 151). By using such techniques, banks can utilize past observations to identify typical patterns in client behavior and predict their future reactions to risk (Hens et al., 2018). While these tools could be powerful, they could be misleading as they may not adequately differentiate between risk tolerance and other personality traits or behavior biases, and therefore, they might fail to differentiate between past and ideal behavior (Hens et al., 2018).

3. Methodology

This chapter present the research design implemented in the study. It further describes the procedures employed for data collection and analysis. An evaluation of the research design, addressing the reliability and validity of the study is then provided. Firstly, the objects studied are presented, before the chosen research design for the study is outlined.

3.1 Client Cases

To study chatbot performance in investor risk profiling, ten different client cases were provided by Amstein (2023). The selection criteria for these cases were well-aligned with the research objectives, ensuring their relevance and appropriateness. 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 a real-world tendency of investors to have limited information about their risk tolerance and related factors. Employing varied investor scenarios was instrumental to thoroughly evaluate chatbot capabilities in risk profiling, aiming for a comprehensive understanding of their performance across different investor types. The client cases can be found in Appendix A.

3.2 Research Design

3.2.1 Choice of Research Design

The research design was chosen to answer the following research questions:

> Research question 1: How do ChatGPT and Bard categorize investor risk profiles compared to financial advisors?
> Research question 2: How consistent are the chatbots' categorizations over time?

A mixed methods design that combines elements of both quantitative and qualitative research was chosen for this study (Schoonenboom & Johnson, 2017). The collected data includes both quantitative and qualitative elements. While the primary data analysis was quantitative, an

additional qualitative analysis was conducted to deepen the exploration of the results and gain insights from the qualitative data.

The study's mixed methods design follows a deductive-simultaneous design, where the core component is quantitative and the supplemental component is qualitative (Morse & Niehaus, 2009, p. 25). The inclusion of a qualitative analysis aimed to enrich and contextualize the quantitative results, in alignment with the principle of "complementarity" as described by Greene et al. (1989). Greene et al. (1989, p. 259) describes complementarity as the process of utilizing insights from one methodological approach to elaborate, enhance, clarify, and illustrate findings from another, thus offering a more detailed understanding of the research subject. In this study, the qualitative analysis aimed to elaborate on the quantitative results by offering insights into the chatbots' reasoning and identifying any potential shortcomings. This approach was deemed relevant for both the question of the chatbots' accuracy and consistency.

A longitudinal design was needed to study the chatbots' consistency. This design enabled sequential observations over time and was used to examine potential variations in the chatbots' risk assessments of the same clients across different time points. By observing potential changes in their suggested risk scores over time, the design allowed for an examination of the consistency of the chatbots' evaluations.

The methodological approach comprises characteristics of both inductive and deductive reasoning (Saunders & Bristow, 2023). Initial hypotheses were formulated after a preliminary literature review, which represents a deductive element by deriving specific hypotheses. Despite stating hypotheses, typically indicative of deductive research, they were not founded on predefined theories or clear, explicit assumptions. Instead, the hypotheses structured the study and directed the data analysis. Given limited preexisting research in the field, there was no clear theoretical framework directing the hypothesis. Their primary purpose was to direct the exploration of data, yet the actual testing leaned towards a deductive approach.

For the subsequent qualitative analysis (detailed in section 3.4), an inductive approach was adopted, focusing on identifying patterns, phrases, and arguments in the chatbots' reasoning to draw general conclusions about their reasoning patterns and potential weaknesses in their explanations. Insights were gained from specific observations, indicating inductive reasoning. (Saunders & Bristow, 2023).

Consequently, the research employed a hybrid approach as it contained elements of both deductive and inductive reasoning. While initial hypotheses guided the quantitative analysis in a deductive manner, the qualitative component adopted an inductive approach,

enabling the discovery of emergent themes and patterns in the chatbots' reasoning that were not initially anticipated. The subsequent sections detail the data collection and analysis processes.

3.3 Data Collection

Primary data, which is data explicitly collected for a study, was collected weekly between October 7th and November 25th 2023. Structured interactions with OpenAI's ChatGPT-4 and Google's Bard were used as the method of data collection. The purpose of these interactions was to prompt the chatbots to assign risk scores to the clients. These interactions simulated digital interviews but diverged notably from traditional interviews due to the non-human nature of the respondents (i.e., the chatbots), instead representing a relatively new method of automated chatbot interactions. To obtain risk assessments for each investor, two conversational methods were applied for each interaction session, conducted in separate sessions to avoid continuity in the conversation. To begin the interaction, both approaches began in a similar manner, and the client description was provided directly into their respective interfaces. The subsequent conversational methods, applied to all clients at each measurement time point, are explained in the subsequent section.

3.3.1 Two Conversational Methods in Chatbot Interactions

After providing the client description, a risk score was directly requested (Direct Method). Specifically, the chatbots were asked to categorize the investor on a scale from 1 (lowest) to 5 (highest), indicating the order of possibility to take risk. This method aimed to obtain a categorization of the investors' risk profiles without further instruction. In the second approach, two separate queries were prompted following the client description, first asking about the investor's risk ability and then their risk tolerance, with each question posed individually (Indirect Method). Subsequently, a request for a risk score was directed in the same manner as the direct approach. Details of the two interaction approaches are included in Appendix B.

The two interaction methods facilitated an examination of potential variations in risk scores after preceding queries as compared to when risk scores were requested directly. The indirect method also aimed to prompt the chatbots to differentiate between the risk dimensions, and enabled an observation of how the chatbots comprehended these dimensions.

Each interaction was brief, resembling a concise, structured interview yet incorporating an element from a survey-based approach. The question posed to prompt the chatbots to categorize the clients on a scale from one to five, is akin to a five-point Likert scale. In addition to providing a risk score, the chatbots provided an explanation, capturing the reasoning associated with the risk assessment. Therefore, the data derived from the chatbot interactions included both quantitative and qualitative information regarding each client case. The chat logs of all chatbot conversations can be found in Appendices (external).

3.3.2 Data Preprocessing and Standardization

ChatGPT and Bard each categorized the ten clients a total of 16 times over time. The risk scores from the bankers 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. In terms of the categorization scale which ranged from 1 to 5, the bankers assigned categories using whole numbers within this range. 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.

3.4 Data Analysis

As the analytical objectives were clear and pre-defined, the data analysis was predominantly confirmatory, founded on testing hypotheses regarding the efficacy of chatbots in categorizing investor risk profiles. Based on the study’s objectives, hypotheses were stated to provide a clear structure and direction, constituting confirmatory research. Subsequently, a more exploratory approach using qualitative analysis of the chatbots explanations was conducted to attain insights into the qualitative data.

To assess RQ1, a comparison with financial advisors from the same bank was used, setting these advisors’ judgements as the standard for accuracy. For RQ2, the consistency of the chatbots was evaluated by testing variations in their suggested risk scores observed over time. The quantitative data analysis was facilitated using the R statistical software, through RStudio (Version 2023.09.1+494) (RStudio Team, 2023). The hypotheses and specific tests conducted are detailed in the following section.

3.4.1 Comparison of the Assessment by Assessors

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. The following hypotheses applied regarding each investor:

H_0: There were no significant differences in the risk scores assessed by the assessors (ChatGPT, Bard, and bankers).

H_1: There was a significant difference in risk scores assessed by the assessors (ChatGPT, Bard, and bankers).

The alternative hypothesis (H_1) implies that at least Bard, ChatGPT or bankers differed in assessment compared to the other assessors, which would imply that at least one of the groups gave risk scores which were statistically different from the other groups. In testing hypotheses H_0 and H_1, the assessment of each client was tested separately to account for the variability in the client descriptions. This allowed for testing of client-specific factors and enabled identification of patterns and significant differences in the assessments that might have been overlooked in an aggregated analysis.

Multiple Kruskall-Wallis tests were conducted to evaluate the presence of statistically significant differences among Bard, ChatGPT and bankers for each client. This non-parametric test enabled a comparison with all assessors and was appropriate due to the non-normal distribution of scores (Fan & Zhang, 2012).

Post hoc Dunn's tests for multiple comparisons were applied for statistically significant results from Kruskal-Wallis to identify the specific groups that exhibit statistically significant differences. As multiple tests were conducted sequentially, the Holm-Bonferroni correction was applied to adjust the p-values (Holm, 1979). This method mitigates the inflated risk of Type I errors, i.e., false positives, that typically occur due to multiple comparisons (Holm, 1979).

Additionally, 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 statistically 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), is 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. However, the tests were conducted to assess if they yielded similar results compared to those from Kruskal-wallis and Dunn's post hoc tests, which were more appropriate given the data. The t-tests served as a robustness check for the results, as t-test results consistent with the non-parametric tests would strengthen confidence in the findings, despite the conditions not being fully met.

3.4.2 Assessing Chatbot Consistency

The hypotheses assessed by this analysis were to test whether the chatbots' scores exhibited temporal variability (time), differences in risk scores between ChatGPT and Bard (tool), and variations due to the conversational methods used, direct and indirect methods. Interaction effects of these factors were also included. The following hypotheses applied:

$H2_0$: There is no effect of the tool on the risk scores. The mean scores for Bard and ChatGPT are equal.

$H2_1$: There is an effect of the tool on the risk scores. The mean scores for Bard and ChatGPT are not equal.

$H3_0$: There is no effect of the method on the risk scores. The mean scores for the direct and indirect methods are equal.

$H3_1$: There is an effect of the method on the risk scores. The mean scores for the direct and indirect methods are not equal.

$H4_0$: There is no effect of time on the risk scores. The mean scores are equal across all eight time points.

$H4_1$: There is an effect of time on the risk scores. At least one time point's mean score is significantly different from the others.

$H5_0$: There is no interaction effect between time and tool on the risk scores.

$H5_1$: There is an interaction effect between time and tool on the risk scores.

$H6_0$: There is no interaction effect between time and method on the risk scores.

$H6_1$: There is an interaction effect between time and method on the risk scores.

$H7_0$: There is no interaction effect between tool and method on the risk scores.

$H7_1$: There is an interaction effect between tool and method on the risk scores.

$H8_0$: There are no interaction effects between tool, method, and time on the mean scores.

$H8_1$: There is at least one interaction effect between tool, method, and time on the mean scores.

Repeated Measures ANOVA was employed to test these hypotheses. The analysis tested for a time effect by assessing whether the chatbots' proposed risk scores changed over the series of eight time points. A non-significant effect would indicate that the chatbots' assessments remained consistent over time, supporting hypothesis $H4_0$, while a significant time effect would suggest inconsistency ($H4_1$). Additionally, the test evaluated a potential between-conditions effect (tool) to determine if there was a significant difference in scoring between ChatGPT and Bard. This was examined for each client in the earlier test, but here consistency data was included. A non-significant effect would indicate no significant difference in risk scores from ChatGPT and Bard ($H2_0$), whereas a significant effect would imply differential assessments ($H2_1$). The test also assessed the effect of the conversational method (direct and indirect methods) used in the chatbot interactions. A non-significant effect of method would suggest consistent risk scores regardless of the conversational method used ($H3_0$), while a significant difference would indicate an impact of the method on the chatbots' assessments ($H3_1$).

The ANOVA further tested for two-way interaction effects between the examined factors. The Time x Tool interaction effect examined whether changes in risk scores over time differed depending on the chatbot used. A significant interaction would imply that temporal changes in risk scores varied by the tool, as posited in hypotheses $H5_0$ and $H5_1$. Similarly, the Time x Method interaction effect assessed whether changes in risk scores over time differed between the methods employed. A significant finding would suggest that the changes in scores over time was method dependent, as posited in hypotheses $H6_0$ and $H6_1$. The Tool x Method interaction effect examined whether the influence of the tool on scores differed based on the method used. A significant interaction effect would indicate that the difference in direct and indirect methods was contingent on whether Bard or ChatGPT was used, as stated in hypotheses $H7_0$ and $H7_1$.

The three-way interaction effect examined the combined influence of time, tool, and method on the risk scores. This type of interaction effect assessed whether the interaction between these factors affected the risk scores, as hypothesized in $H8_0$ and $H8_1$. This tests if the combinations of the factors led to patterns in risk scores. For example, it could reveal if ChatGPT assessed differently with direct queries over time compared to Bard.

By testing for effects of time, tool, and methodological variations, the Repeated Measures ANOVA provided an assessment of the consistency and reliability of the chatbots' evaluations. The results collectively offered an understanding of how these factors influenced the chatbots' in evaluating the clients' risk profiles.

3.4.3 Qualitative Analysis of Chatbot Reasoning

Differences in assessments between chatbots and bankers prompted further investigation into the instances where the chatbots' assessments deviated from those of the advisors. For client cases with statistically significant differences, a qualitative analysis of the chatbots' responses was conducted. The analysis involved identifying phrases and arguments in the chatbots' reasoning related to their risk scores, which was either higher or lower than the advisors' assessments. This provided insights into their reasoning, and some weaknesses in their explanations were identified.

3.5 Evaluation of Research Design

3.5.1 Reliability

Throughout the data collection process, interaction queries with the chatbots were kept consistent to standardize the study's conditions, thereby contributing to its external reliability. However, the method of data collection, which involves interactions with chatbots, presents complexities in obtaining consistent data due to the inherent probabilistic characteristics of chatbot responses. Nonetheless, all interactions were documented, including both the questions posed and the corresponding responses. This documentation, along with the description of the data analysis process, facilitates potential replication of the study. Documenting the interactions ensured that the original responses were preserved in their entirety and that subsequent analyses were grounded in an accurate representation of the chatbots' output, a matter of validity.

3.5.2 Validity

Validity in this study is crucial for ensuring the accuracy and relevance of the findings. Measurement validity refers to the questionnaire's ability to measure what it is intended to measure (Saunders et al., 2016, p. 450). Since the categorization question prompted responses on a scale from 1 to 5, similar to a Likert scale, the criteria for measurement validity are

applicable. A primary concern regarding measurement validity is whether the question posed to the chatbots accurately measured what it was intended to measure (Saunders et al., 2016, p. 450). The question was formulated to obtain assessments of clients' risk profiles, with the intention that the risk scores derived from the Likert scale query would reflect the chatbots' categorizations of the clients' risk profiles.

Construct validity relates to how well the query prompted actually measures the presence of the construct intended to measure, specifically the theoretical concept of a "risk profile" (Saunders et al., 2016, pp. 450-451). The information about the clients varied in detail and was likely not comprehensive enough to cover all aspects of their risk profiles. For instance, there was not complete information on all aspects of their financial situations and risk tolerance. The chatbots made assessments based on the information received, and it is likely that the risk scores did not cover all aspects of the clients' risk profiles. Because of these considerations, the chatbots' categories were treated as "risk scores" on a scale from 1 to 5.

The accuracy of these scores in reflecting a "risk profile" also depends on how the chatbots interpreted the query. In principle, if the chatbots algorithmic interpretation closely align with the conceptual foundation of risk profiling, this may be valid. However, any misalignment, possibly due to the chatbots' limitations in comprehending the nuances of financial risk-taking behavior or the complexity of individual investor profiles, could indicate a discrepancy with the concept of a risk profile. Since the study aimed to evaluate chatbots' ability to assess risk profiles, their interpretation of and response to the specific query are integral to the study's findings. Chatbots' interpretation of the prompt is part of the results, offering insights into the tools' processing accuracy and applicability in financial risk profiling.

Consistent chatbot interactions aimed to ensure that the responses obtained were directly attributable to the chatbots' interpretation of the investor scenarios, rather than being influenced by varying questioning or interaction styles. To ensure the independence of each interaction with ChatGPT and Bard, every session was initiated as a new conversation. Starting each session in a new chat interface aimed to ensured that the chatbots' responses were not conditioned or influenced by prior exchanges, thereby maintaining data integrity, and ensuring that each response was a direct result of the specific prompt provided, free from earlier context.

The study used descriptions of investor scenarios that aimed to reflect real-world situations, ensuring that the chatbots' responses were evaluated in contexts relevant and representative of typical investor risk profiling tasks. The client cases were originally written in German and translated into English before used with the chatbots. For translation, both

Google Translate and ChatGPT-4 were utilized, aiming to maintain the original content. The linguistic differences between these two languages imply that there is a risk that the translation might have modified the nuances of the descriptions to some extent. Consequently, potential loss of linguistic nuances could impact the comprehensibility of the descriptions.

Regarding external validity, i.e., the generalizability of the findings, the results should not be generalized to be applied to categorizations of investors of chatbots on a general basis. Ten client descriptions were used. Each of ChatGPT and Bard assessed these clients multiple times, resulting in a total of 16 observations per chatbot for each client. From the bankers, there were a total of 47 to 49 risk scores. The ten client descriptions may not be representative of all investors. Furthermore, only two chatbots were tested. Therefore, the findings are likely not applicable to all chatbots or to the assessment of all types of investors. The matter of limited generalizability is further addressed in the discussion (section 5.1).

4. Results

In this chapter, the findings from the data analysis are presented. First, descriptive statistics are provided to give an overview of the assessments from Bard, ChatGPT, and bankers. Furthermore, the quantitative results from the statistical tests, and insights from the qualitative analysis are presented. Lastly, the synthesis of findings is given.

4.1 Descriptive Statistics

4.1.1 Comparison of the Assessment by Assessors

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 were used as measures of 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 indicated 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, where bankers categorized slightly higher risk scores than chatbots', and for client 7, where ChatGPT categorized a higher risk.

Despite individual differences in risk scores for each client, the average risk scores assigned by ChatGPT, Bard and the bankers were nearly the same across all clients. Bard exhibited an average score of 3.3, while both ChatGPT and bankers had an average of 3.4. This trend indicates a remarkable consistency among the evaluations of the different assessors when considering their overall average performance.

The range, 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 narrow range would indicate a strong consensus on a client's risk score, while narrow score ranges for chatbots would indicate consistent risk assessment over time. Bard's suggested score range 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 limited variability in the evaluations over time. With a range of scores consistently at one score or less for all clients, ChatGPT demonstrated a certain degree of self-consistency in assessing the clients.

Compared to the chatbots, the ranges of scores from bankers were more variable for the ten clients. The overall larger ranges indicated that the group of bankers provided a wider range of scores along the rating scale. For clients 1 and 6, bankers used the full range of scores. This suggests that there was notable variation in opinions among the bankers. Such variability in scores could suggest a more diverse interpretation of clients' risk among the bankers, underscoring the nuanced perception of risk for the clients. 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.

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

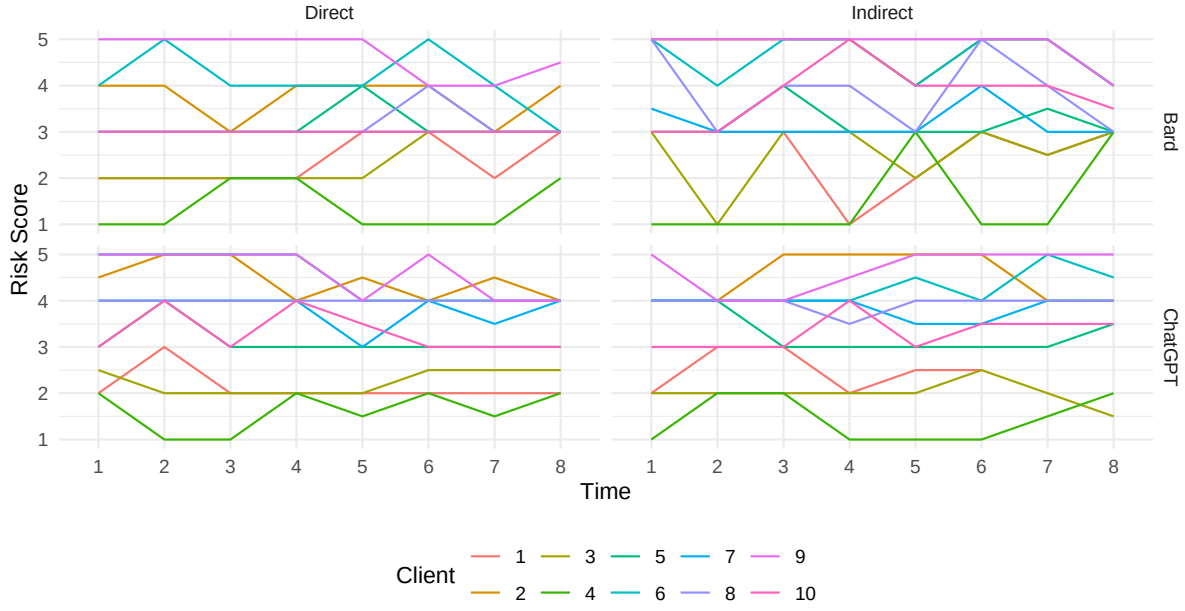
4.1.2 Chatbots' Risk Scores Over Time (Chatbot Consistency)

Figure 1 features a line plot that illustrates the risk assessments for each client as provided by ChatGPT and Bard over time. The plot differentiates between the conversational methods, i.e., the direct and indirect methods used. The plot highlights temporal trends, showing how the risk scores for each client fluctuated over the measurements. Additionally, it facilitates a

comparison of risk scores from ChatGPT and Bard for each client and delineates the variance in risk scores attributable to the use of either the direct or indirect assessment method.

Figure 1

Risk Scores for Clients by Bard and ChatGPT Over Time Using Direct and Indirect Methods



ChatGPT provided relatively consistent assessments of the clients across the eight time points, regardless of whether categorization request was made directly or after preliminary questions (Figure 1). The risk scores suggested by ChatGPT varied by less than two scores throughout the measurement period. For instance, it assigned client 1 scores of 2, “2 or 3”, or 3. For client 8, the assessment was almost entirely consistent over time using the direct method, with only a minor deviation where it stated “3 leaning towards 4” once with the indirect method.

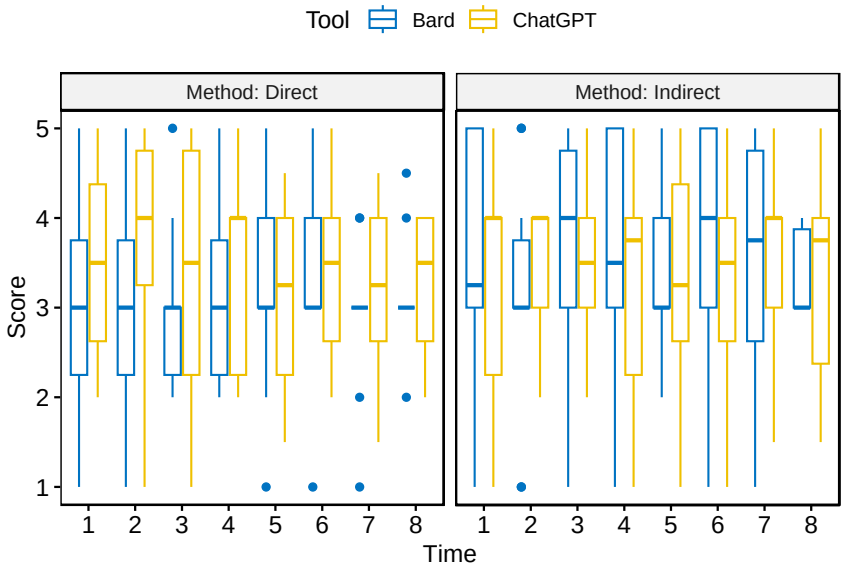
Bard’s suggested risk scores for each client mostly varied within two scores, though for some clients, the variation spanned three scores. A more distinct difference was observed between the direct and indirect methods. Bard’s risk scores were more consistent over time when using the direct method, varying within two scores for most clients and three for one client, which scores for clients 7 and 10 remained consistent. In contrast, the indirect method resulted in larger variability in the scores assigned to clients. Specifically, for client 4, Bard altered between scores of 1 and 4. The indirect method led to more frequent shifts from one score to another, resulting in less consistent scores over time.

Figure 2 illustrates the distribution of scores from Bard (blue) and ChatGPT (yellow), distinguishing between the direct and indirect methods (two panels), over the eight evaluation

sessions. The X-axis represents the assessment time points, corresponding to the eight weekly measurement occasions, while the Y-axis denotes scores ranging from one to five. The boxes in the graph aggregate the scores for all clients, categorized by each combination of tool, method, and time.

Figure 2

Box Plot of Risk Scores Over Time



Bard's scores showed relative consistency over time with the use of the direct conversational method, as indicated by the stable median of 3 throughout the rating period (Figure 2). In contrast, when the indirect method was used, the scores exhibited somewhat greater variability, and the overall median values were higher. ChatGPT's scores demonstrated a similar degree of consistency over time regardless of the method employed. Compared to Bard, ChatGPT's median scores were slightly higher when the direct method was used.

The size of the Interquartile range (IQR) reflects the variation in scores across all clients, indicating how distinctly the chatbot differentiated between the ten investors at each time point. The IQR size varied across the different measurements. Bard appeared to differentiate the clients more using the indirect method compared to the direct method, as evidenced by the larger IQRs. In the direct method, the scores were more clustered, suggesting a more homogeneous evaluation of the clients. For ChatGPT, the difference in the IQE sizes between the direct and indirect methods was minimal; however, the IQRs were slightly larger with the direct method.

When considering assessments of all clients, there was a certain degree of consistency over time, as indicated by relatively stable median values. Although there was some fluctuation between a median of 3 and 4, these changes were not dramatic, suggesting moderate variations in scores over time.

4.2 Results From Statistical Tests

4.2.1 Comparison of Assessments by Assessors

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 (Appendix C). 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 2 presents the Kruskal-Wallis test results for each of the clients. For client 1, the test did not show a statistically significant difference in scores by assessor, $\chi^2_{(2)} = 1.19$, $p = 0.551$. Similarly, no statistically 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 2

Kruskal-Wallis test statistics

Client	Kruskal-Wallis $\chi2$	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, statistically 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 statistically 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, presenting the test statistics for the specific groups that exhibited significant differences. In the subsequent sections, the results for each client are detailed separately. Accompanying each test result, a plot is presented to visualize the post hoc test results (Figures 3-7). Each plot displays the distribution of risk scores by assessors, visualizing both central tendency and variability. Individual scores are displayed with a slight horizontal jitter to avoid overlap, and error bars represent the mean and standard error for each assessor's risk assessment scores.

Table 3

Results from Dunn's tests for significant KW's

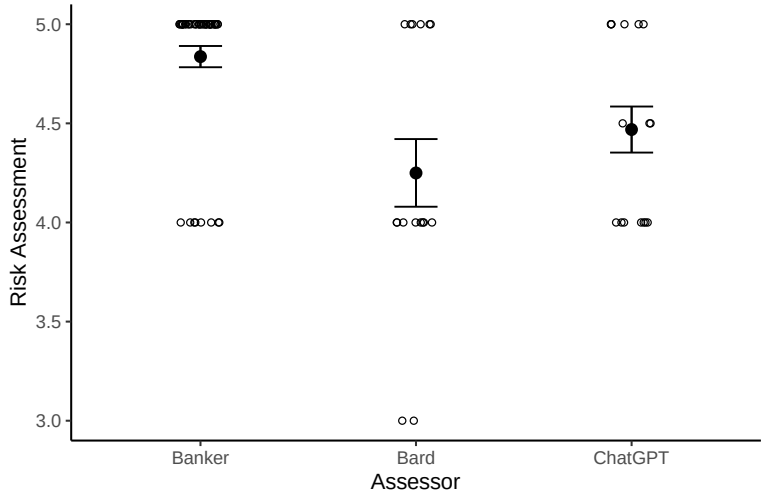
Client	Comparison	*z*	Unadjusted *p*	Adjusted *p* (Holm)
Client 2	Banker - Bard	3.72	p<.001	p<.001
	Banker - ChatGPT	2.92	.004	.007
	Bard - ChatGPT	-0.58	.565	.565
Client 3	Banker - Bard	2.91	.004	.007
	Banker - ChatGPT	5.63	p<.001	p<.001
	Bard - ChatGPT	2.21	.003	.027
Client 6	Banker - Bard	-3.59	p<.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

Client 2: The Lawyer

Dunn's test indicated a statistically significant difference between bankers and Bard, p<.001, with a z-value of 3.72. This suggested a higher ranking from bankers. The results also revealed a statistically significant difference between bankers and ChatGPT (p = .007), Z-value of 2.92, also indicating a higher rating from bankers. No significant difference was found between Bard and ChatGPT (p = .565). These results indicated a statistically significant difference in scoring, where bankers scored the lawyer higher than both chatbots. Figure 1 corroborates these statistical findings. The banker's scores were higher than those of ChatGPT and Bard, as illustrated by the non-overlapping estimates between bankers and both chatbots, indicating distinct scoring patterns. While Bard's assessment was slightly more conservative than ChatGPT's, this difference was not significant, as indicated by the overlapping estimates.

Figure 3

Risk Assessment by Assessor for The Lawyer (Client 2)

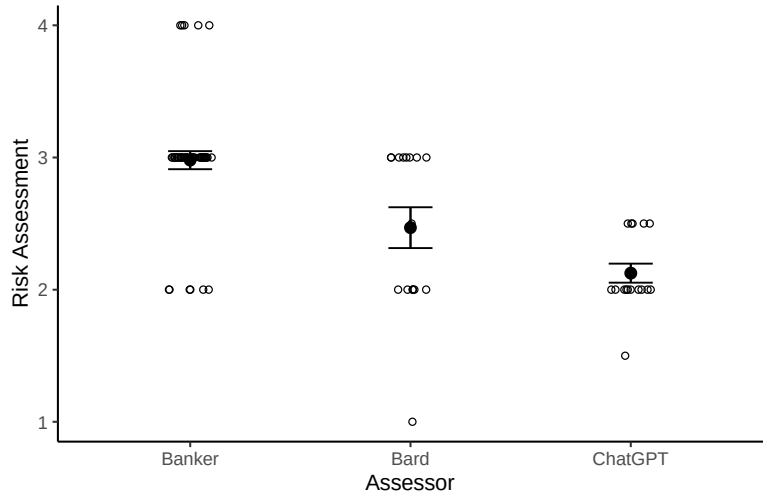


Client 3: The Architect

There was a statistically significant difference in scoring between bankers and Bard (p = .007, z = 2.91) and between bankers and ChatGPT (p<.001, z = 5.63), suggesting higher scores from bankers than both chatbots. The difference between Bard and ChatGPT was also statistically significant (p = .027, z = 2.21). The results indicated statistically significant differences between all assessors, with the strongest difference noted between bankers and ChatGPT. Figure 2 illustrates these differences, showing that the bankers assessed the architect higher than both chatbots, highlighting the disparity between their assessments.

Figure 4

Risk Assessment by Assessor for The Architect (Client 3)

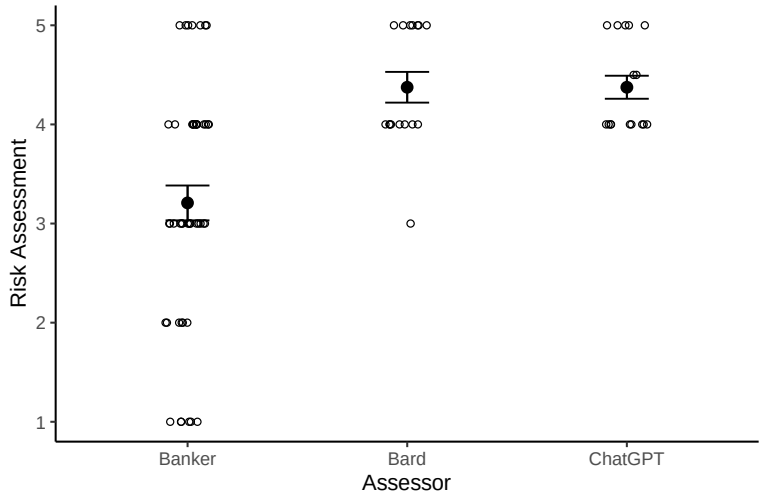


Client 6: The Engineering Student

For the engineering student, the results revealed a statistically significant difference between the bankers and Bard ($p<.001$, $z = -3.59$), and between the bankers and ChatGPT ($p<.001$, $z = -3.62$). No statistically significant difference was observed between Bard and ChatGPT ($p = .984$). These results were similar to those for the lawyer (client 2), indicating differences in assessments between bankers and the chatbots. In contrast to the lawyer's case, the bankers' assessment for this student was more conservative than that of the chatbots. The differences are evident in Figure 3, where the bankers assessed the student significantly lower than both chatbots, while Bard and ChatGPT did not differ significantly, as illustrated by their overlapping estimates.

Figure 5

Risk Assessment by Assessor for Engineering Student (Client 6)

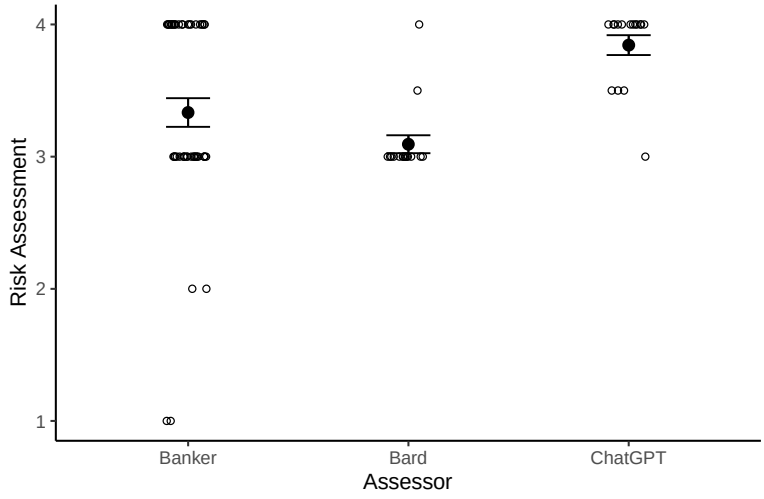


Client 7: The Insurance Consultant

For the insurance consultant, the test revealed a statistically significant difference in scores between the bankers and Bard ($p = .034$, $z = 2.12$) and between bankers and ChatGPT ($p = .009$, $z = -2.84$). There was also a statistically significant difference between Bard and ChatGPT ($p<.001$, $z = -4.05$). The test results, as illustrated in Figure 6, reveal that ChatGPT assigned significantly higher scores to this client compared to both bankers and Bard, with Bard's scores being lower than those of both bankers and ChatGPT.

Figure 6

Risk Assessment by Assessor for Insurance Consultant (Client 7)

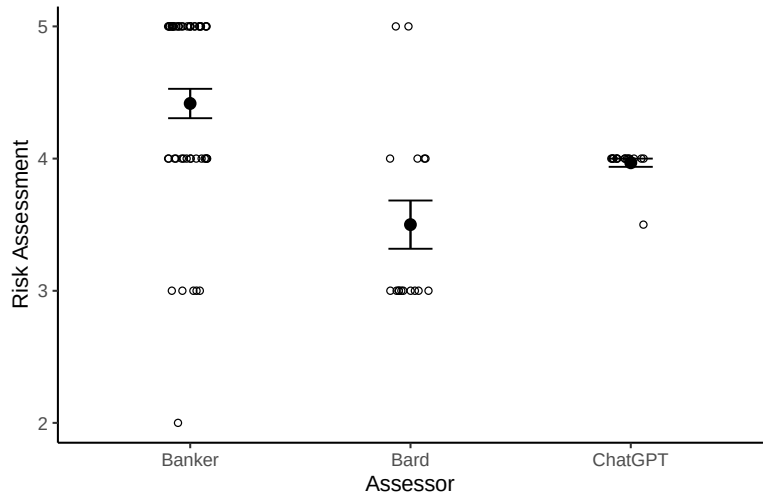


Client 8: The Restaurant Chain Owner

The results for this client indicated a trend similar to that observed in the test of the lawyer (client 2), with the banker's scores being significantly higher than those of the chatbots. There was a statistically significant difference between the scoring of the bankers and Bard ($p<.001$, $z = 4.22$) and between the bankers and ChatGPT ($p = .027$, $z = 2.47$). The difference between chatbots was not statistically significant ($p = .154$). Figure 7 shows notable lower estimates for the scores by the chatbots than those from the bankers.

Figure 7

Risk Assessment by Assessor for The Restaurant Chain Owner (Client 8)



The plots of risk assessments for the clients where the test did not uncover any statistically significant differences, are included in Appendix D. The considerable overlaps in estimates are indicative of a lack of district separation between the assessors' scoring patterns, suggesting a degree of homogeneous assessment across the assessors (D1-D5). The similar scoring estimates suggested there was a certain agreement among the assessors regarding these clients' risk levels. The presence of scores that span across common ranges for the assessors corroborates the statistical findings that there was no statistically significant difference in the risk assessments for these clients.

Welch's t-test Results

The Welch's t-test revealed statistically significant differences in the scores between ChatGPT and bankers for several clients (Table 4). 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, $t(54) = 3.89$, $p = .004$. The results for these clients aligned with the test results from Dunn's test. For Client 2, the test did not reveal a statistically significant difference after the p-value was corrected by holms method, $t(22) = 2.88$, $p = .105$. Similarly to the results from Dunn's tests, no statistically significant differences were observed in the assessments for Client 1, $t(57) = 1.32$, $p = 1.000$; Client 4, $t(36) = -0.42$, $p = 1.000$; Client 5, $t(46) = 0.65$, $p = 1.000$; Client 9, $t(20) = 1.71$, $p = 1.000$; and Client 10, $t(51) = -0.66$, $p = 1.000$.

Table 5 presents the results of Welch's t-tests for the assessment scores between Bard and bankers for each client. Statistically significant differences were found for Client 6, $t(51) = -4.98$, $p <.001$, and Client 8, $t(27) = 4.29$, $p = .003$. These assessments were also significantly different according to Dunn's test. However, for clients 2, 3 and 7, were Dunn's test indicated statistically significant differences, such differences were not observed in the Welch's t-tests after applying the Holm correction: for Client 2, $t(18) = 3.28$, $p = .058$; Client 3, $t(21) = 3.02$, $p = .084$; and Client 7, $t(61) = 1.87$, $p = .731$. Similarly to the results from Dunn's tests, no statistically significant differences were observed in the assessments for Client 1, $t(38) = 0.00$, $p = 1.000$; Client 4, $t(23) = 0.16$, $p = 1.000$; Client 5, $t(53) = 1.15$, $p = 1.000$; Client 9, $t(22) = 0.83$, $p = 1.000$; and Client 10, $t(33) = -1.02$, $p = 1.000$.

Table 4

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 5
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

4.2.2 Summary of Findings

The findings from the tests comparing assessments by assessors are summarized as follows. The primary tool of analysis, the Kruskal-Wallis test, supplemented by Dunn's post hoc analysis, revealed statistically significant differences in risk scores for five of the ten clients. These clients included the lawyer, the architect, the engineering student, the insurance consultant, and the restaurant chain owner. These findings support hypothesis H_1, which posited a significant difference in risk scores between the assessors. For the remaining clients, hypothesis H_0 was not rejected, as no significant differences in the risk scores evaluated by the assessors were found. The testing yielded mixed results: statistically significant differences in assessments were observed for half of the clients, while for the other half, no such differences were evident.

Results from Dunn's test showed that the scores from both chatbots differed significantly from those of bankers in all five instances. Specifically, the chatbots assessed the lawyer, the architect, and the restaurant chain owner, with significantly lower risk scores than the bankers. Conversely, the chatbots assessed the engineering student significantly higher. Bard's assessments of the insurance consultant were lower, whereas ChatGPT's were higher, compared to the bankers' scores. Additionally, there were differences between ChatGPT and Bard, with Bard scoring the architect higher and the insurance consultant lower.

The discrepancies in test outcomes should be noted. The Welch's t-tests, conducted as a robustness check, did not consistently reflect these findings across all clients. The Welch's t-tests supported the significant differences between ChatGPT and bankers for the same clients, except for the lawyer, where no difference was detected. Similarly, Welch's t-tests showed significant differences between Bard and bankers for the engineering student and the

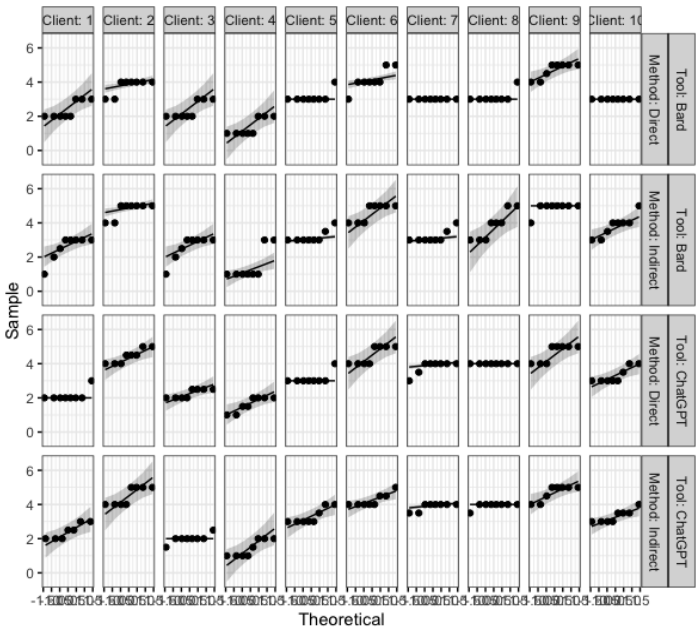
restaurant chain owner but revealed no significant differences for the rest of the clients. Notably, the Welch's t-test did not detect significant differences in risk assessments between chatbots and bankers for all the clients highlighted by the non-parametric tests, indicating a deviation from the findings established through non-parametric analysis. This indicates that while Welch's t-test corroborated the non-parametric test results for some clients, it did not for others.

4.2.3 Assessing Consistency

Repeated measures ANOVA makes several assumptions about the data that must be applicable. Multiple outliers were identified by using the "identify_outliers ()" function in R. As the test assumes no significant outliers, the twelve scores considered as outliers were removed from the dataset to uphold the assumption of the dependent variable being approximately normally distributed. The data was reassessed after removal of the outliers, and Figure 8 shows the Q-Q plot after the outliers were removed.

Figure 8

Q-Q Plot



Upon visual inspection of the Q-Q plots, the distributions predominantly conformed to normality, albeit with some observations exhibiting limited variability. Considering the repeated measures ANOVA's robustness to minor deviations from normality, it was deemed appropriate to proceed with the analysis. The decision aligns with the understanding that the

test can tolerate certain departures from the normal distribution without substantial detriment to the validity of the results. Thus, the assumptions required for the test were judged to be sufficiently met. Table 6 shows the results from the repeated measures ANOVA performed to evaluate the effects of tool, method, and time on score.

Table 6

Results from Repeated Measures ANOVA (Type III test)

Effect	df (error)	F	*p*	η^2_G
Tool	1 (2)	0.65	.506	0.00
Method	1 (2)	3.14	.218	0.01
Time	7 (14)	0.73	.648	0.01
Tool x Method	1 (2)	12.96	.069	0.02
Tool x Time	7 (14)	0.33	.928	0.00
Method x Time	8 (14)	0.46	.851	0.01
Tool x Method x Time	9 (14)	0.15	.991	0.00

The main effect of tool was not statistically significant, $F_{(1,2)} = 0.65$, $p = .506$, generalized eta squared (ges) = 0.00, indicating that there was no significant difference in scores between the two chatbots, ChatGPT and Bard. The main effect of the method was not statistically significant, $F_{(1,2)} = 3.14$, $p = .218$, ges = 0.01, which indicates that the direct and indirect methods of querying did not yield significantly different scores. The effect of time was also not significant, $F_{(7, 14)} = 0.73$, $p = .648$, ges = 0.01, indicating that the scores did not significantly change over the eight time points measured. Due to the absence of statistically significant effects of tool, method, and time, $H2_1$, $H3_1$ and $H4_1$, could not be concluded.

No significant interaction effects were found between tool and method, $F_{(1, 2)} = 12.96$, $p = .069$, ges = 0.02, the tool and time, $F_{(7, 14)} = 0.33$, $p = .928$, ges = 0.00, or the method and time, $F_{(7, 14)} = 0.46$, $p = .851$, ges = 0.01. Additionally, the three-way interaction between tool, method, and time was not significant, $F_{(7, 14)} = 0.15$, $p = .991$, ges = 0.00, suggesting that the interaction between these factors did not significantly affect the scores. The results implied that hypotheses $H5_1$, $H6_1$, $H7_1$, $H8_1$ was not supported as the related null hypotheses was not rejected.

As there was no significant change in risk scores over the eight time points tested, this indicated that ChatGPT and Bard provided relatively stable assessments, consistent over time. The lack of a significant tool effect suggested that there was no substantial difference in the average risk scores given by ChatGPT and Bard. The result implied that the chatbots performed similarly in terms of providing risk scores on a scale from 1 to 5, with neither

consistently assigning higher or lower risk scores than the other across the measurements. These results suggests that there is no basis for preferring one chatbot over the other. The non-significant method effect indicated that the two querying methods used to obtain risk assessments from the chatbots did not significantly differ in their impact on the scores. This implied that the way questions were posed– whether directly asking for a categorization or leading up to it with preliminary questions about the client's risk ability and tolerance – did not result in statistically different scoring outcomes. The additional context provided by the preliminary questions did not significantly alter the chatbots' final risk score compared to direct requests for categorization. In terms of the final numerical risk score assigned, the assessments seemed to be relatively stable across the different querying approaches. The chatbots' assessments were consistent and not significantly influenced by the two different querying methods.

The non-significant interaction effects between tool, method, and time point suggest there was no significant interaction influencing the risk scores assigned by the chatbots. The lack of a significant interaction between the chatbot type and the questioning method indicated that ChatGPT and Bard provided similar risk scores, regardless of whether they were asked directly or after preliminary questions. Essentially, the way in which questions were posed did not lead to any noticeable difference in how the chatbots assessed risk for the clients. The tool by time and method by time interactions suggested that the scoring remained consistent over time, regardless of the tool or method employed. The chatbots' scoring remained consistent over the various time points, regardless of which chatbot was used or whether the question was asked directly or indirectly. This consistency across tools and methods over time indicated that the chatbot evaluations were stable and reliable.

The non-significant three-way interaction between tool, method and time suggested that no discernible or consistent patterns were found in the risk scores that could be attributed to a specific combination of these three factors. The results implied that the risk assessments remained even and predictable, unaffected by any combination of chatbot, questioning method, and point of evaluation. This indicated a level of robustness in the chatbots' ability to assess the client's risk, as their risk scores were not dependent on how and when the questions were posed. Consistency was not influenced by these variables.

In conclusion, the repeated measures ANOVA provided an assessment of the chatbots' reliability in categorizing clients. The results indicated that the chatbots provided stable risk assessments over time, which was not significantly affected by which of them was used or the method of questioning. Since no significant interaction was observed between the tool, method

and time, this suggested that chatbot evaluations were robust and reliable. The absence of significant effects implies that the chatbots evaluated the clients with a degree of consistency, regardless of the time point or method employed.

4.3 Insights from Qualitative Analysis of Chatbot Reasoning

Statistically significant differences were observed in the assessments provided by the chatbots and bankers for five of the clients (see section 4.2.1). The responses of the chatbots when assessing these clients were utilized for qualitative analysis. Complete descriptions of these clients are found in Appendix A. All monetary values referenced are in Swiss Francs (CHF) and the following quotes are found in the chat logs of conversations with the chatbots in Appendices (external).

Both ChatGPT and Bard tended to consider the same factors in their assessments of the clients, based on the information they received. For example, the chatbots typically justified their proposed risk score with the clients' goals, investment horizon, financial situation, knowledge, and experience, among other factors. Across the interactions, it was apparent that the emphasis on different factors, and at times the interpretation of how these factors influence the clients' risk capacity and tolerance, varied. The chatbots' use of the terms "risk capacity", "risk ability", and "risk tolerance" was at times interchangeable, suggesting a misalignment with the definitions found in the literature. This was not surprising, given the numerous similar and often inconsistent applications of terms in describing risk attitudes (Nobre & Grable, 2015). However, this aspect was not further explored in the analysis. Some weaknesses in the chatbots' reasoning, pertinent to their accuracy, was observed. These will be further described in the next section.

4.3.1 Lack of Personalized Recommendations

Lack of personalized recommendations was suggested when the chatbots provided general recommendations or arguments without fully utilizing all the available client information. Bard's responses were typically shorter and less detailed than those of ChatGPT, often containing less information relevant to the specific case and instead offering more general advice. Some statements were generalized, indicating that specific client information was not utilized. For example, when assessing the engineering student (client 6), Bard stated, "Students typically have lower incomes and fewer financial obligations, which can give them more flexibility to invest in riskier assets". When evaluating the insurance consultant (client 7), Bard

stated, “You are divorced and have two adult children. This suggests that you have financial obligations to others.”. This reasoning appeared to lack personalization, especially since the case detailed that the consultant’s adult children have completed their studies and, after years of paying alimony, the consultant now has more funds available for himself and wants to invest for retirement.

4.3.2 Reliance on General Principles

The chatbots’ reasoning often relied on general principles when assessing how certain factors affected the clients’ risk. Age was commonly used as an indicator of the investment horizon’s length, and the chatbots tended to struggle with comprehending the horizon when it was not directly related to the client’s age. Both chatbots rated the engineering student’s risk higher than bankers did, frequently arguing for a long investment horizon, allowing for higher risk taking. While the student just started their master’s program and wants to invest 50,000 to get closer to the 80,000 required for the start-up initiation after studies, both chatbots struggled to comprehend their short investment horizon. For instance, Bard gave responses such as, “They are young and have a long investment horizon” or “The investor is likely investing for the long term, as they are not planning to use the money in the short term”. In one instance when assigning a risk score of 5, Bard recognized the goal by stating, “They are currently starting their Master’s program at ETH and have a goal of starting a tech start-up with friends after completing their studies. For the start-up, each person would need to contribute about 80,000 CHF initially […]”. However, in the same response it stated, “They are young and have a long time horizon for their investments. This means that they can afford to take on more risk in the short term, as they have time to recover from any losses.”. Although Bard in this response acknowledged the client’s goal of starting a start-up, it inadequately grasped that the investment horizon is short, approximately two years. Such contradictory reasoning suggested that Bard struggled to comprehend the investment horizon when its general principle did not apply to the specific case at hand. Defining the investment horizon based solely on age was considered as an oversimplification that did not match the real situation.

ChatGPT’s responses also regularly considered the student’s age, suggesting a longer investment horizon as it allows for recovery from potential short-term losses. However, a more nuanced understanding of the horizon tended to appear later in the same responses, such as, “[…] The investor is looking to almost double the current assets in a relatively short time frame (typically the duration of a Master’s program, which can be 1-2 years). Shorter horizons can constrain the ability to wait out market downturns.”. Such contradictory reasoning related

to the investment horizon within the same response suggested a lack of integrated analysis of the factors involved. Limited contextual and holistic comprehension was suggested based on these findings. The factors emphasized in the responses were often summarized at the conclusion, which at times expressed that the chatbots were considering each factor separately rather than holistically.

Similarly, the chatbots argued with a long investment horizon based on the client's age when assessing the lawyer. This also applied to Bard in assessing the restaurant chain owner (client 8), despite it was described that a significant portion of his wealth could be invested for long term. In contrast to assessments of other clients where the described horizon was not recognized, ChatGPT usually comprehended the architect's investment horizon correctly. A longer investment horizon was argued and related to the investment objective of saving for children's education and the ages of the children. For instance, ChatGPT stated, "Given that your children are 5 and 6 years old, you likely have at least a 10–15-year investment horizon until they start their higher education. A longer investment horizon usually allows for more potential risk-taking because it provides time to recover from market downturns.". Bard's comprehension of this client's horizon varied, sometimes related to age, and sometimes more correctly related to their investment goal.

4.3.3 Lack of Factual Support

Lack of factual support was suggested when the chatbots used information not provided or made assumptions based on unknown data. Some statements from Bard lacked factual support from the prompt. For instance, when assessing the engineering student, Bard claimed, "[…] they are also aware of the risks involved and are considering diversifying their portfolio", although there was no information in the case about the student's plan for diversification. Bard also mentioned a "stable income", contradicting the case details which indicated the student had no income. Although the prompt made it clear that the student has no regular expenses, Bard commented that the student's savings constitutes a significant amount of money, but not enough to cover the entire education costs or the cost of starting a tech start-up, which also lacked support from the prompt. Furthermore, Bard´s assertion of the student´s risk tolerance in an instance were not backed by the case information, as it stated, " […] All of these factors suggest that the investor is very comfortable with the possibility of losing some of their money in the short term in order to potentially achieve higher return in the long term. They are also willing to invest in riskier assets, such as stocks and cryptocurrencies". The prompt did not

mention any willingness to invest in such risky assets and the assertion that the investor is comfortable with the possibility of short-term losses contradicted the information provided.

4.3.4 Lack of Human-Like Understanding

Lack of human-like understanding was observed in some ChatGPT interactions. Its higher assessment of the engineering student was justified in a more aggressive risk approach as necessary to approach their goal, despite the case specifying the aim was to "get closer to" rather than fully achieve the 80,000 goal. For instance, ChatGPT stated, "You have a specific and relatively near-term goal to fund a startup. This may require a more aggressive investment approach if you're looking to grow your 50,000 CHF to 80,000 CHF within the time frame of your Master's program and thereafter.".

ChatGPT suggested that the student could absorb potential financial losses without immediate personal consequences, based on their lack of financial obligations. This reasoning did not consider the personal consequences of losses hindering the startup initiation. This aspect was not adequately addressed in the responses.

In assessing the lawyer, both chatbots consistently argued for a high-risk tolerance, based on their expressed awareness of value fluctuations and acceptance as long as not extreme. The lawyer's expressed belief in the future of cryptocurrencies and "not hesitant to make a significant investment in this asset class" was also used to support this evaluation. However, in one instance, Bard mentioned that the investor is "[…] only willing to invest a portion of their assets in cryptocurrencies" and interpreted this as an unwillingness to take on excessive risk. This was the sole argument for a lower risk score associated with assigning a score of 3. This reasoning seemed inconsistent with the investor's statement, as being "not hesitant to make a significant investment" suggests a willingness to allocate a significant portion of the fund to cryptocurrencies, indicating a higher risk tolerance. Here, Bard's reasoning associated with a perceived lower risk tolerance, appeared to be partly based on an assumption that the investor will limit his exposure to cryptocurrency, which was not directly evident from the statements.

The insurance consultant, knowledgeable in the financial industry, expressed awareness that value appreciation involves greater risk and acceptance of risk to achieve a 3-5% annual return. Both chatbots frequently acknowledged the aim for a 3-5% annual return, arguing that achieving this return do not necessitate an overly aggressive approach. However, ChatGPT tended to justify the investor's expressed willingness to accept some risk related to 3-5% return as indicative of a high risk tolerance and its high ratings appeared to be linked to

this perceived high risk tolerance. For instance, ChatGPT stated, “The investor states they have a willingness to accept higher risks in pursuit of value appreciation. This suggests a higher risk tolerance.”. As the investor did not express a willingness to undertake significantly high risk, but rather an acceptance of some level of risk for the return target, such reasoning seemed to overestimate the risk tolerance. In contrast, Bard’s reasoning was more conservative, often directly linking the willingness to take risks to the goal of 3-5% annual return. An example of Bard’s reasoning was “You are willing to accept some risk in order to achieve a 3-5% annual return. This is a reasonable goal that is achievable with a diversified portfolio of investments.”.

4.4 Synthesis of Findings

4.4.1 Results from Analysis Answering RQ1

In this study, no statistically significant differences were observed in the risk scores assigned by the chatbots and the financial advisors for five of the clients. This suggests that, for half of the clients, the chatbots aligned with the advisors’ assessments to a certain extent. For the other five clients, statistically significant differences were observed between the assessments of chatbots and the advisors. Both chatbots were more conservative than advisors in assessing the lawyer, architect, and restaurant chain owner, but assessed the engineering student more aggressively. Additionally, ChatGPT rated the insurance consultant higher, whereas Bard rated them lower compared to the bankers. Overall, the certainty of how ChatGPT and Bard categorize investor risk profiles compared to financial advisors remains unclear, as the results varied among different clients. For certain clients, the Welch’s t-tests did not identify significant differences that the Kruskal-wallis supplemented by Dunn’s tests did.

Multiple limitations in the chatbots’ reasoning were observed in the qualitative analysis, which may have impacted their accuracy. These limitations were classified into themes: “lack of personalized recommendations”, “reliance of general principles”, “lack of factual support”, and “lack of human-like understanding”. In some cases, these weaknesses were associated with the chatbots’ tendency to categorize clients at higher or lower risk scores compared to the assessments made by bankers.

When comparing the assessments of ChatGPT and Bard, significant differences were observed in their evaluations of two clients: the architect and the insurance consultant. ChatGPT rated the architect lower, but the insurance consultant higher compared to Bard.

Despite significant differences between all assessors in evaluating these clients, Bard's assessment of the architect and insurance consultant were closer to the advisors' compared to ChatGPT. For the remaining seven clients, there were no significant differences between ChatGPT and Bard, indicating a general agreement in their evaluations. Their similar assessments were further supported by the results of the Repeated Measures ANOVA, which showed no significant differences in scores depending on the chatbot used when tested across assessments of clients and time.

In conclusion, the analysis indicated that ChatGPT and Bard exhibit varying degrees of alignment with financial advisors when categorizing investor risk profiles. While there were no significant differences in risk scores for half of the clients, indicating a certain level of alignment, discrepancies were evident in the other half. This divergence in assessments, along with the identified weaknesses in the chatbots' reasoning, suggests that ChatGPT and Bard do not consistently match the accuracy of financial advisors' evaluations.

4.4.2 Results from Analysis Answering RQ2

Descriptive statistics showed minor variations in the risk scores provided by ChatGPT and Bard for each individual client over time. ChatGPT's risk scores showed variations within a narrow range of less than two risk scores across the eight measurements, which can be considered relatively consistent. Bard's suggested risk scores varied to a slightly greater degree and appeared to be somewhat influenced by the interaction method used.

The results from the repeated measures ANOVA showed no statistically significant effect of time on the scores, indicating that the scores remained relatively consistent throughout the observed period. Additionally, no significant effect of the conversational method was detected when examining data across clients and over time. This indicated that the method of prompting – whether following queries about risk ability and risk tolerance or posed directly – had no significant impact on the scores.

The qualitative analysis indicated that the factors highlighted by the chatbots were relatively consistent across interactions. However, the emphasis and sometimes the interpretation of different aspects could vary to some extent over the interactions.

Overall, the study found that both ChatGPT and Bard exhibited a degree of consistency in their assessments over time. Despite some variations in their assessed risk scores for each client, these variations were relatively minor and did not indicate significant inconsistencies. The analysis, including both descriptive statistics and repeated measures ANOVA supported the conclusion that the chatbots maintained relatively stable risk scores over the period

observed. This consistency was further emphasized by the lack of significant differences in the scores based on the conversational method used. Therefore, the overall findings indicated that ChatGPT and Bard were relatively consistent in their risk categorisations of investors over time.

5. Discussion

The clarity of how ChatGPT and Bard categorize investors compared to financial advisors was not evident, as it varied for different investors in this study. Consequently, the chatbots' accuracy in categorizing investor risk profiles is not definitive. The results suggested that ChatGPT and Bard may be adequate in certain situations, indicating a potential to match financial advisors' ability to categorize investor risk profiles with comparable accuracy. However, the reasoning and factors underlying the chatbots' assessments for these clients, and whether these aligned with those of the advisors, remain uncertain, as it was not investigated in this study.

Despite statistically significant differences were determined for half of the clients with the non-parametric test, such statistically significant differences are not a sufficient condition for the results to be meaningful. The statistical results should not be interpreted as indicating importance or providing meaningful support for the findings. The economic significance of these differences remains to be explored. Practitioners in the field could determine the economic significance of the results, indicating whether the results are meaningful in a more practical sence.

For applications of chatbots in risk profiling, it is apparent that considerations extend beyond providing risk scores. The qualitative analysis provided insights into the chatbots' reasoning when categorizing the clients, for the clients where the non-parametric tests uncovered statistically significant differences between both chatbots and bankers. The weaknesses identified in their reasoning impact their accuracy and practical suitability. Although the advisors' reasoning was not used for comparison, the identified weaknesses where in some instances suggested to be linked to their divergent scores from the advisors'.

The chatbots' frequent use of age in assessing the clients' investment horizon indicated a comprehension that younger clients often have a higher risk capacity due to their longer time to recover from market downturns (Cordell, 2001). However, their frequent reliance on such general principles sometimes resulted in a poor match for specific cases. For instance, the chatbots' interpretation that the engineering student has a longer time horizon appeared to underlie the higher risk scores assigned and may explain the chatbots' higher assessments of this client compared to that of advisors.

Financial obligations related to children and family were repeatedly understood as affecting the clients' risk capacity (Cordell, 2001). Additionally, the chatbots' assessment of the clients reflected an understanding that higher income was linked to a higher risk capacity,

and correspondingly, that high expenses were linked to a lower risk capacity (Cordell, 2001). However, it was suggested that they struggled with integrated analysis of these factors. This applied to assessments of investment horizon and at times, they also seemed to struggle to relate and weight elements of the clients' financial situation to each other.

Instances where the chatbots' reasoning seemed to lack nuance where it was suggested to be a "lack of human-like understanding", can be related to the chatbots inability to fully comprehend the complexity of human language and conversations (Bogost, 2022). These findings supports the notion that chatbots cannot fully understand the meaning behind words, which might lead to responses that are shallow, lacking depth and insight (Bogost, 2022). In some cases, this weakness in responses may be associated with divergent ratings compared to financial advisors. For instance, ChatGPT's higher assessment of the insurance consultant may relate to this limitation, as the ratings was associated with a perceived higher risk tolerance, likely overestimating their actual risk tolerance. This was also supported by Bard's reasoning, which directly linked the risk tolerance to the goal of 3-5%, corresponding with lower assessments compared to both ChatGPT and bankers.

Bard's arguments made based on unprovided information (lack of factual support), are likely related to the phenomenon of extrinsic hallucination (Ji et al., 2023). This is supported by the statements in their generated responses that contained information that could not be verified from the input. This was the case when Bard stated that the engineering student was considering diversifying their portfolio, despite no information provided about this, asserted a "stable income" without details of any income, or claimed that the student was willing to invest in riskier assets, such as stocks and cryptocurrencies. As these claims were based on unknown information and contained details that could neither be confirmed nor denied based on the original input, it was considered indicative of extrinsic hallucinations (Ji et al., 2023).

Considering the limitations identified in chatbots' reasoning, one could argue that the chatbots are not suited at all. Their frequent oversimplification and misinterpretation of nuanced information supports the notion of an inherent lack of human-like understanding, which raises concerns about the depth and reliability of their assessments and overall suitability for tasks that require judgment and personalized evaluation. The potential occurrence of extrinsic hallucinations further impairs these concerns. These errors reflect a broader issue of LLMs current limitations in fully grasping and responding to the intricacies of human language and individual circumstances. Therefore, their use in sensitive areas such as risk profiling requires careful considerations. The shortcomings identified indicate that the

human capacity to fully interpret, emphasize, and adapt to different situations of individuals cannot yet be fully replicated by these chatbots.

The overall risk score assessments by ChatGPT and Bard were relatively consistent, despite minor variations in their assessments of individual clients over time. The reliability of chatbots in categorizing investors risk profiles also depends on the premise that the choice of conversational method does not influence the assessments. Any variations in the chatbots' suggested risk scores, depending on the conversational method used, could imply inconsistencies in their assessments. This study further supports the chatbots' consistency, demonstrating that the conversational method used had no statistically significant effect on their assessments.

Descriptive statistics of the bankers' scores revealed disparities in their assessments of the same clients, indicating disagreement among them. This observation supports the theory that the process of assessing clients is affected by subjectivity (Rice, 2005; Foerster et al., 2014), leading financial advisors to assess the same clients differently. Compared to the group of bankers, the chatbots appeared to be more consistent in their assessments. However, a definitive conclusion cannot be made since the higher number of observations among bankers could, to some extent, account for a broader range observed in the bankers' evaluations.

ChatGPT was asked for a self-assessment: "Is ChatGPT good at risk profiling?" It should be clarified that "self-assessment" in this context means ChatGPT "evaluates" its own capabilities, reflecting its programming and the data on which it has been trained, rather than exhibiting self-awareness. The language model's statements are based on its training data, not on self-reflection. In response to the question, ChatGPT indicated that while it can provide information on a wide range of topics, including risk profiling to some extent, it is not specifically designed for risk profiling. Furthermore, it explained that while it can offer general information and guidance, it lacks access to real-time data or specialized knowledge in specific domains. Moreover, it advised, "For accurate risk profiling, especially in sensitive or specialized areas, it's advisable to consult with experts, use dedicated risk assessment tools, or leverage platforms designed for risk management that consider a broader range of factors and up-to-date information.". ChatGPT gives a cautionary note. If you ask ChatGPT for medical or legal advice, it you will receive similar responses.

The self-assessment by ChatGPT underscores the need for caution when considering its use for risk profiling, mirroring the findings of this study. While it was demonstrated that ChatGPT (and Bard) could keep up with bankers in certain scenarios with comparable risk

scores, limitations in their reasoning were identified, impacting their accuracy and practical applicability.

5.1 Limitations of the Study

The present study has several limitations. Firstly, its generalizability is restricted. The scope of the conclusions is limited due to the dataset which comprised 16 categorizations from ChatGPT and Bard, and mostly 48 by bankers for each client. Moreover, the ten client cases may not fully be representative of all investor descriptions. The study was confined to the specific chatbots, ChatGPT-4 and Bard, and the findings may not apply to other LLMs, particularly those with specific training within finance.

There was also a selection bias due to the group of advisors limited to those in one specific bank. Consequently, the study relied on their subjective ratings, which may introduce a bias, as judgments could vary among advisors. Such variability was also observed within the group of bankers in this study, where their assessments of the same clients diverged, resulting in some clients receiving scores spanning from one to five. As the chatbot data was compared with a sample of data from bankers in one specific bank, the topic should be studied further comparing findings with a wider range of samples to include the appraisal of advisors from other countries to assess cross-cultural differences.

Other limitations include the subjective choice of questions. The study did not investigate the ideal questioning method for risk assessment and was limited to two dialogue approaches. From a theoretical understanding, the output of chatbots is influenced by the quality and formulation of the input (Biswas et al., 2023). Therefore, it remains unclear whether different conversational approaches would have resulted in alternative outcomes. Additionally, the questions prompted to obtain risk scores from chatbots are not extensive of all possible inquires that investors could pose to get an assessment. Further, the method used to obtain risk scores was structured and did not fully resemble a real dialogue due to lack of interaction. Thus, the study provides insight only into the chatbots' performance based on the employed methodology, and the results are not universally applicable for how they work, which needs to be studied further.

The study focused on versions of ChatGPT and Bard available between October 7th and November 25th, 2023. Since then, more recent versions of both models have been released. The study's findings are time dependent, and the quick advancement of chatbot technology may change and improve their effectiveness. Future studies should consider these updates.

To address the limitation of limited generalizability, the subject matter and context of the study was described, and caution is informed in generalizing the findings to be applied to chatbot categorizations of investors in general. A fundamental consideration which further limits the generalizability of this study is the question of whether language models are suited for tasks such as risk profiling. As discussed in previous literature (see section 2.3) and supported by findings from the qualitative analysis (see section 4.3), chatbots lack the ability to perform genuine cognitive processing or to "think" in the same way humans do. However, the premise of the study has been that chatbots can be utilized in risk profiling, especially since people use them for financial advice.

5.1.1 Considerations

Some considerations must be addressed regarding the results of Welch's t-tests, which complemented the non-parametric tests. The data did not fully meet the assumptions required for parametric tests. Violations of assumptions can influence Type I and Type II errors, and can result in over- or under-estimation of the inferential measures and effect sizes (Osborne & Waters, 2002), implying that the results may not be valid. Therefore, the non-parametric tests, being more compatible with the data, informed the subsequent qualitative analysis.

In the context of qualitative analysis, another aspect to consider is the nature of the explanations provided by the chatbots in their responses. Since the responses of chatbots result from complex interactions derived from their training data, it is not feasible to attribute definitive "reasoning" behind their assessments. Therefore, such reasoning should not be interpreted as full explanations for their suggested scores. Although certain statements were frequently associated with particular scores, there was no clear association, but rather a tendency observed. In some cases, the reasoning appeared similar although different scores were given. Analyzing these responses provided insights into the types of information and arguments ChatGPT and Bard tended to use when assigning risk scores.

The bankers used for comparison received the same limited information as the chatbots and made their assessments without client interaction. This approach was adopted to establish a benchmark for comparison. Typically, financial advisors require more comprehensive information and interactive dialogues for accurate risk profiling. The bankers' approach in this study likely deviates from their real-world practices, suggesting that their assessments might not reflect standard procedures in actual scenarios. Therefore, the study's results should be understood within its specific context. The lack of statistically significant differences in

assessments between chatbots and advisors should not be interpreted as evidence of the chatbots' proficiency being on par with that of financial advisors.

In analysing the consistency of chatbots, adjustments were made to the dataset prior to the analysis, particularly in relation to the treatment of outliers. Multiple outliers were excluded prior to the execution of the repeated measures ANOVA to align with the statistical test prerequisites. This procedural refinement may reduce score variability, which may impact the sensitivity of the ANOVA to detect subtle effects. It is conceivable that this adjustment may slightly change the mean scores, possibly leading to a conservative estimation of effect sizes. The resulting non-significant findings should be considered with the understanding that the exclusion of outliers may marginally increase the probability of a Type II error. While the data suggests consistency in chatbot evaluations across time, the implications of outlier removal warrant a cautious approach to the conclusiveness of these observations.

The repeated measures ANOVA primarily tested whether there was a significant change in mean scores over time collectively for the group of clients. Individual changes in scores for each client over time were not necessarily detected by this test, as the analysis assessed overall changes in mean scores for the entire client group. In case of a significant time effect, it denoted a general change in scores across the measured time points but did not identify the specific investors responsible for this variation. If there were significant individual variations that offset each other across the investor group, the collective effect of time may not be significant in the results. The test confirmed the consistency of the chatbots' scoring collectively but did not determine the consistency of scores at the individual client level. This implied the following: Although no statistically significant time effect was detected in the repeated measures ANOVA, there may still be some score variations in the categorization for clients, as this was not specifically tested in this test. Descriptive results showed some variation in scores from chatbots for each client, but these differences were neither systematic nor substantial enough to change the overall mean significantly, thus eluding detection of statistically significant effects in the test.

6. Conclusion

This thesis addressed two primary research questions, “How do ChatGPT and Bard categorize investor risk profiles compared to financial advisors?” and “How consistent are the chatbots’ categorizations over time?”

The study revealed no statistically significant differences in the risk scores assigned by ChatGPT and Bard compared to those assigned by bankers for half of the clients. This suggests that, in some instances, chatbots may align with financial advisors’ in categorizing clients’ risk profiles to a certain extent. However, for the remaining clients, significant differences were observed, indicating that the chatbots’ ability to match advisors’ risk categorizations varied across different clients. For these clients, both chatbots assessed the clients’ risk profiles more conservatively for three clients and higher for one. While ChatGPT and Bard showed potential for providing accurate risk assessments in certain contexts, their overall accuracy in comparison to advisors is not definitive across all clients.

Comparing the outcomes of non-parametric tests with those of Welch’s t-tests, discrepancies were observed. Specifically, Welch’s t-test did not reveal statistically significant differences for some clients, whereas the non-parametric test did. However, the assumptions required for Welch’s t-test were not met, and the smaller sample size may have reduced the sensitivity of the test and lowered its power.

While statistically significant results were determined for half of the clients with the non-parametric test, this does not imply that the differences are meaningful. The economic significance of these differences remains to be explored, and practitioners could determine whether the results are meaningful.

Multiple limitations in the chatbots’ reasoning were observed in the qualitative analysis, affecting their accuracy. These limitations included a lack of personalized recommendations, reliance on general principles, absence of factual support and a lack of human-like understanding. The practical application of chatbots should be approached with an understanding of these limitations. The significant differences in the chatbots’ assessments for half of the clients and the identified weaknesses in reasoning emphasizes the need for further research into their capabilities for risk profiling.

ChatGPT and Bard exhibited a certain degree of consistency in their risk score assessments over time. Although there were minor variations in the scores assigned to clients across measurements, these did not signify major inconsistencies. The absence of a significant effect of the conversational method used further supports the chatbots’ consistency in their

evaluations. While no significant effect of time or method was detected with the repeated measures ANOVA, the removal of outliers and the focus on collective rather than individual client scores may mask subtle variations. This suggests a cautious interpretation of the chatbots' consistency. Moreover, insight from the qualitative analysis indicated that the factors the chatbots' deemed important and took into considerations were quite consistent across interactions, though some variations in emphasis and interpretation of certain factors were noted.

This study provided insights into the capabilities of ChatGPT-4 and Bard in providing risk assessments for investors, an important aspect of providing investment advice. Users should be aware of the inherent limitations and potential risks associated with such AI-driven tools. Sole reliance on the current iterations of ChatGPT and Bard as the only source for risk profiling is not advisable. Users should consider these limitations and assess chatbot-generated information and advice along with trusted, established, and reliable financial sources. ChatGPT and Bard are only as reliable as the data on which they are trained. However, if trained by reliable, experienced financial professionals and on a comprehensive dataset of validated financial information (including investment theories and investor behaviour studies etc.), LLMs could have the potential to significantly advance the process of risk profiling for investors. With further validation, tools like ChatGPT and Bard could serve as valuable and efficient sources for rapid access to assessments for investors. Further research is needed to validate the reliability of such LLMs for risk profiling, with validation from professionals and diverse question formulations. Integrating LLMs like ChatGPT and Bard into the practise of risk profiling also requires various considerations, including ethics and privacy. Further development and refinement are required to improve the reliability and robustness of tools like ChatGPT and Bard before they could be efficiently applied.

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: Conversational Methods

Direct Method

"Categorize the investor between one and five indicating the order: 1 lowest to 5 highest possibility to take risk."

Indirect Method

"What is the risk ability of this investor?"
"What is the risk tolerance of this investor?"
"Categorize the investor between one and five indicating the order: 1 lowest to 5 highest possibility to take risk."

Appendix C: Shapiro Wilk test statistics and Q-Q-plots

Table C1

Results from Shapiro Wilk test

	W	*p*
Client 1	0.849	p<.001
Client 2	0.654	p<.001
Client 3	0.813	p<.001
Client 4	0.714	p<.001
Client 5	0.775	p<.001
Client 6	0.860	p<.001
Client 7	0.736	p<.001
Client 8	0.825	p<.001
Client 9	0.495	p<.001
Client 10	0.796	p<.001

Figure C1

Distribution of Data for Client 1

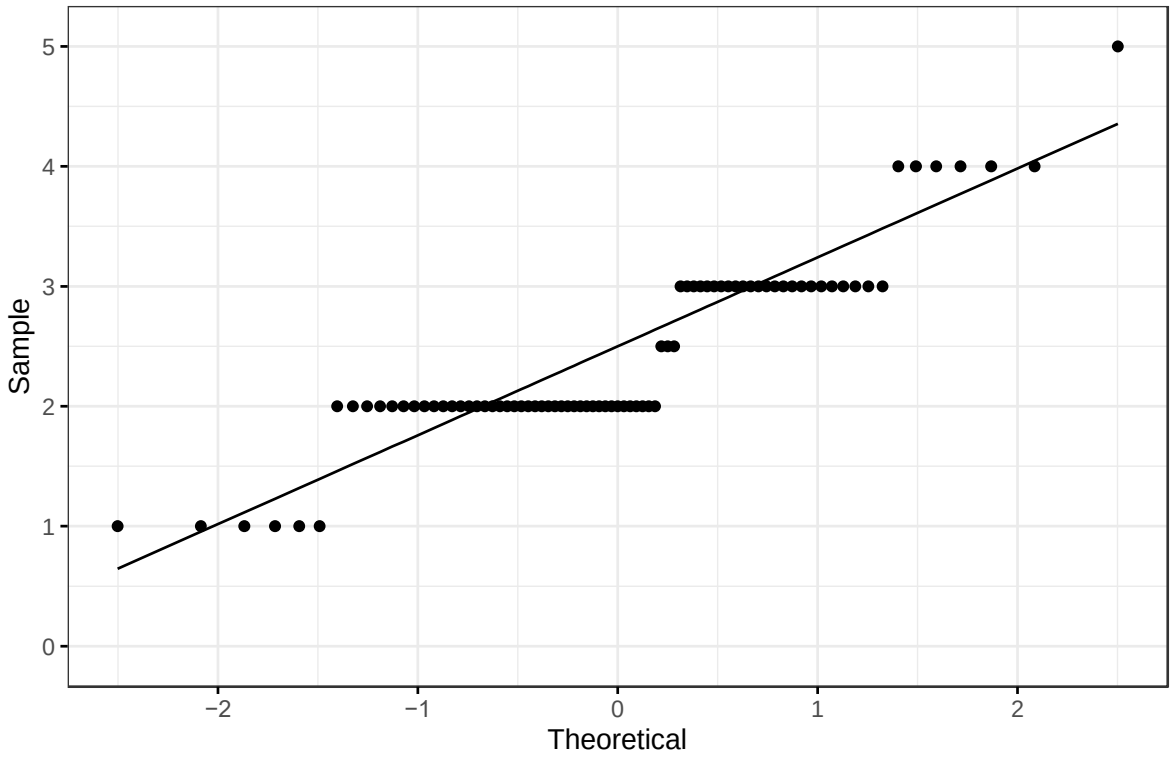


Figure C2

Distribution of Data for Client 2

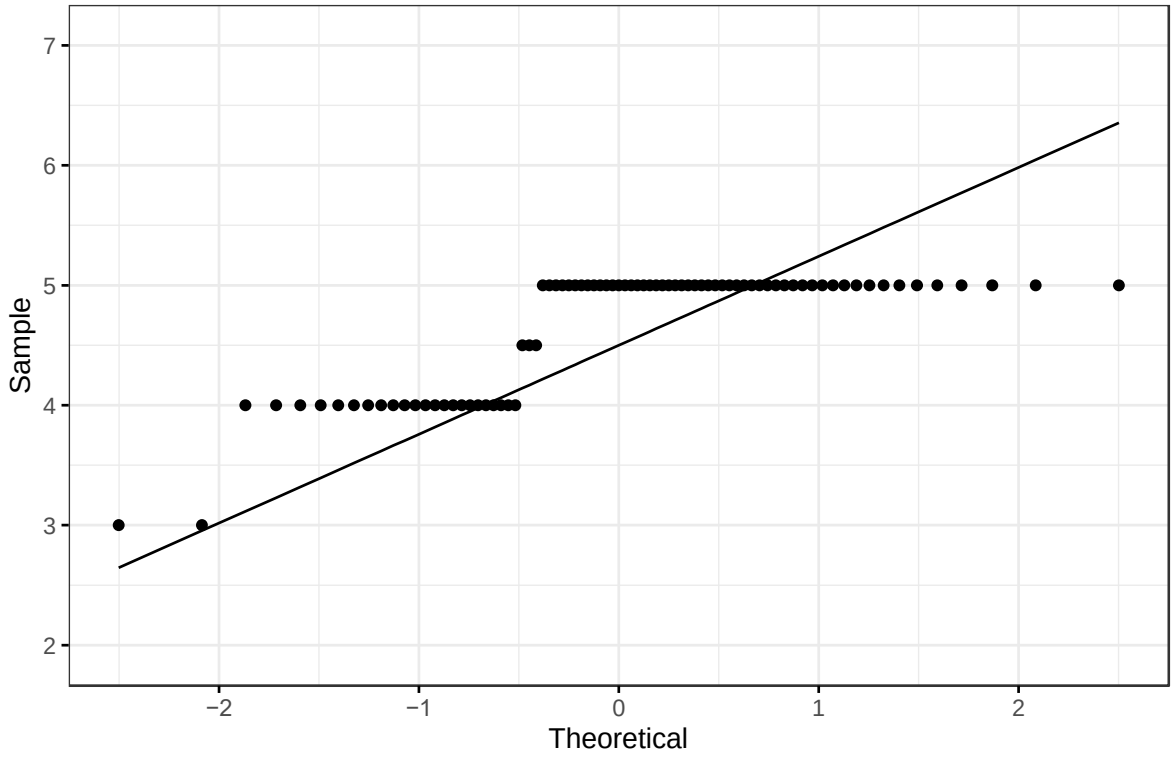


Figure C3

Distribution of Data for Client 3

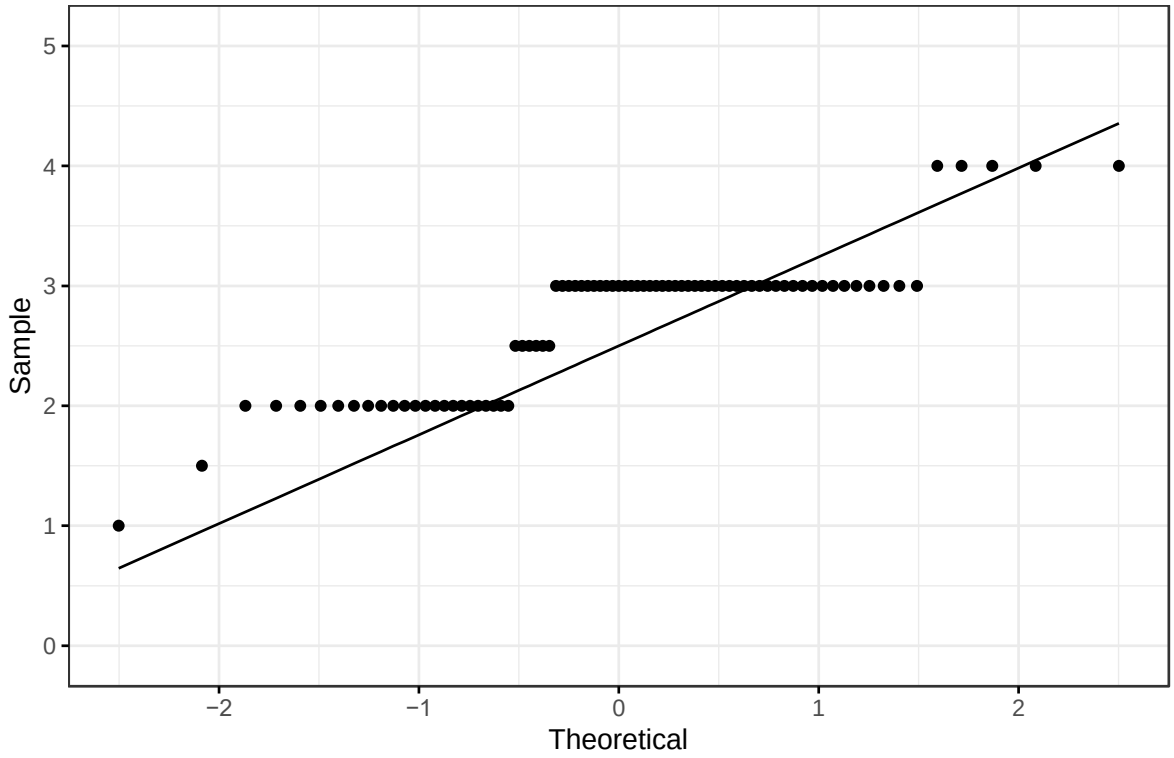


Figure C4

Distribution of Data for Client 4

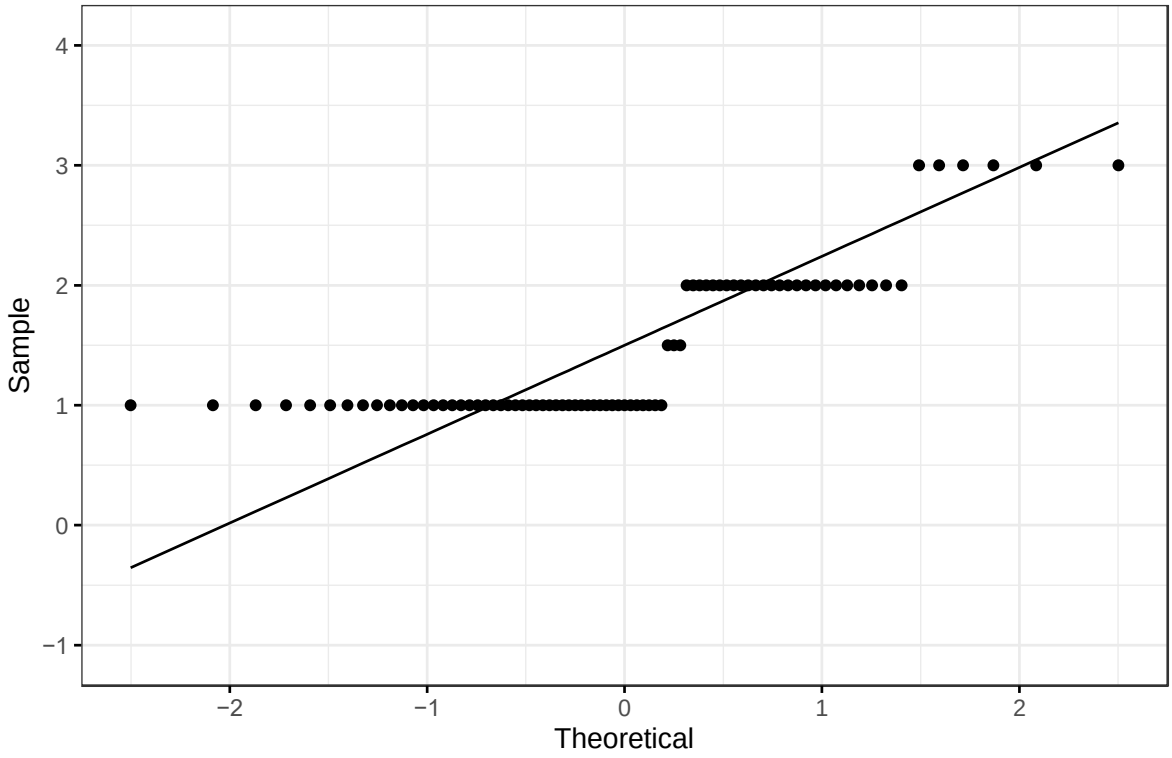


Figure C5

Distribution of Data for Client 5

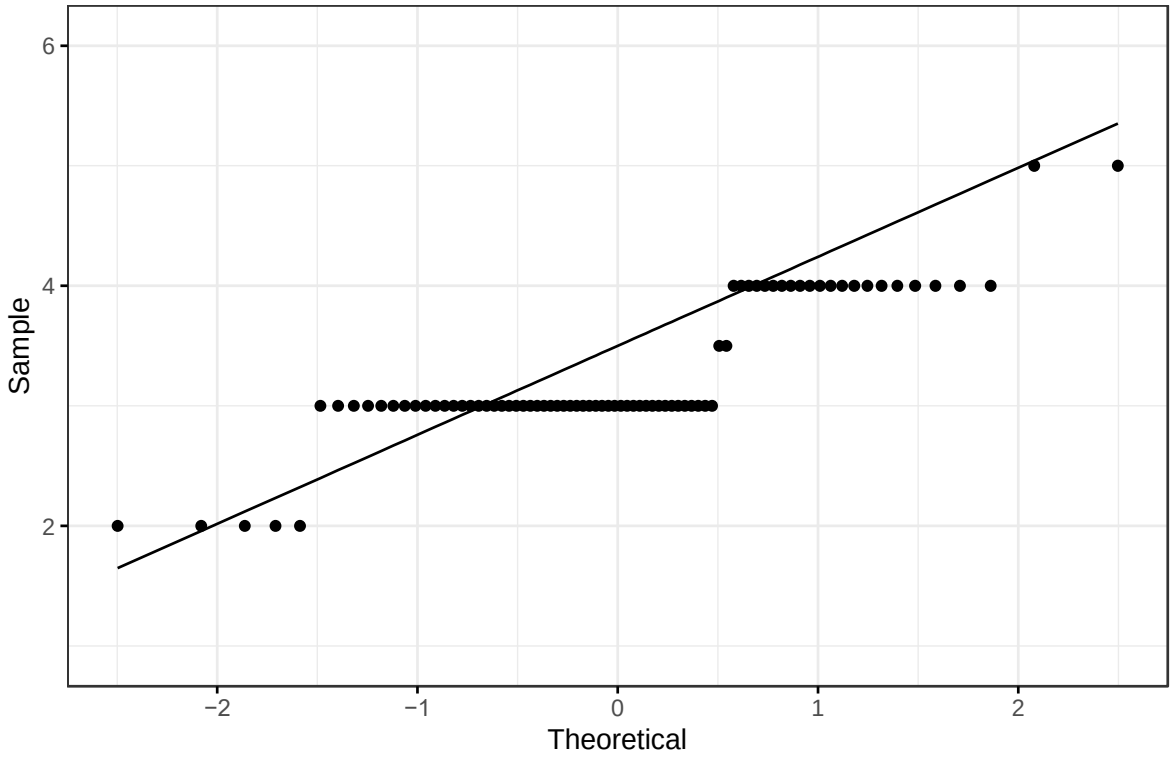


Figure C6

Distribution of Data for Client 6

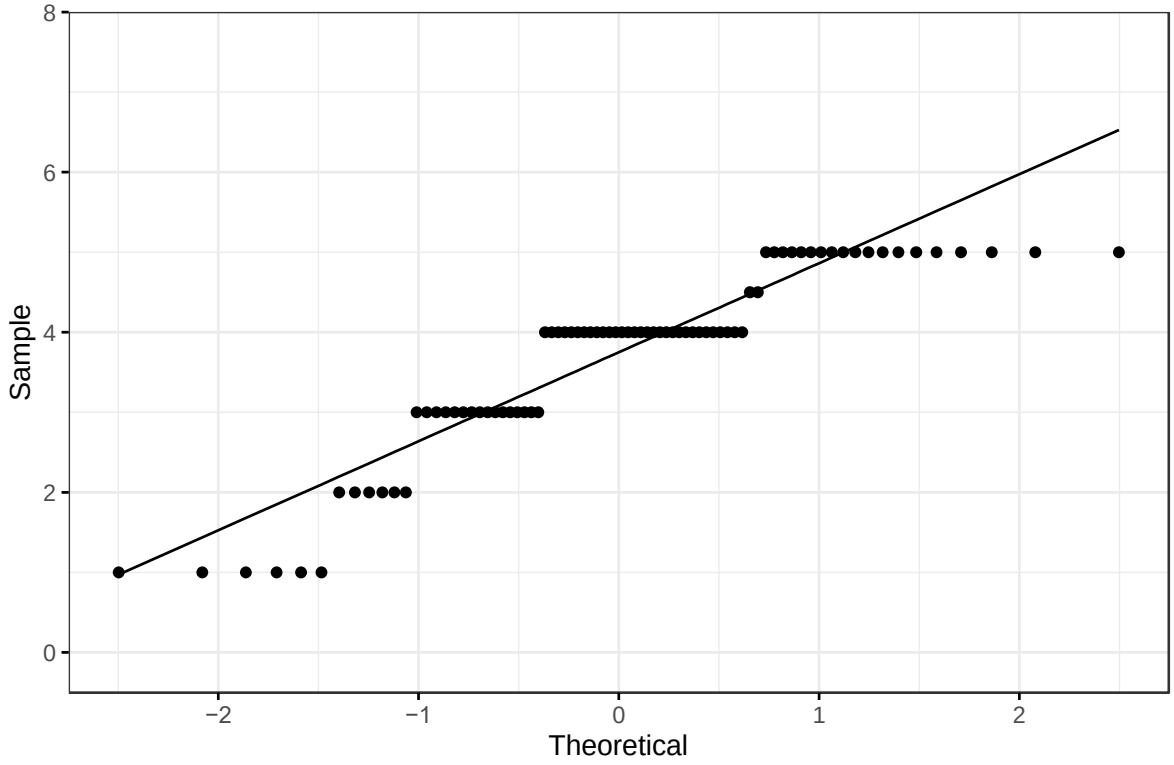


Figure C7

Distribution of Data for Client 7

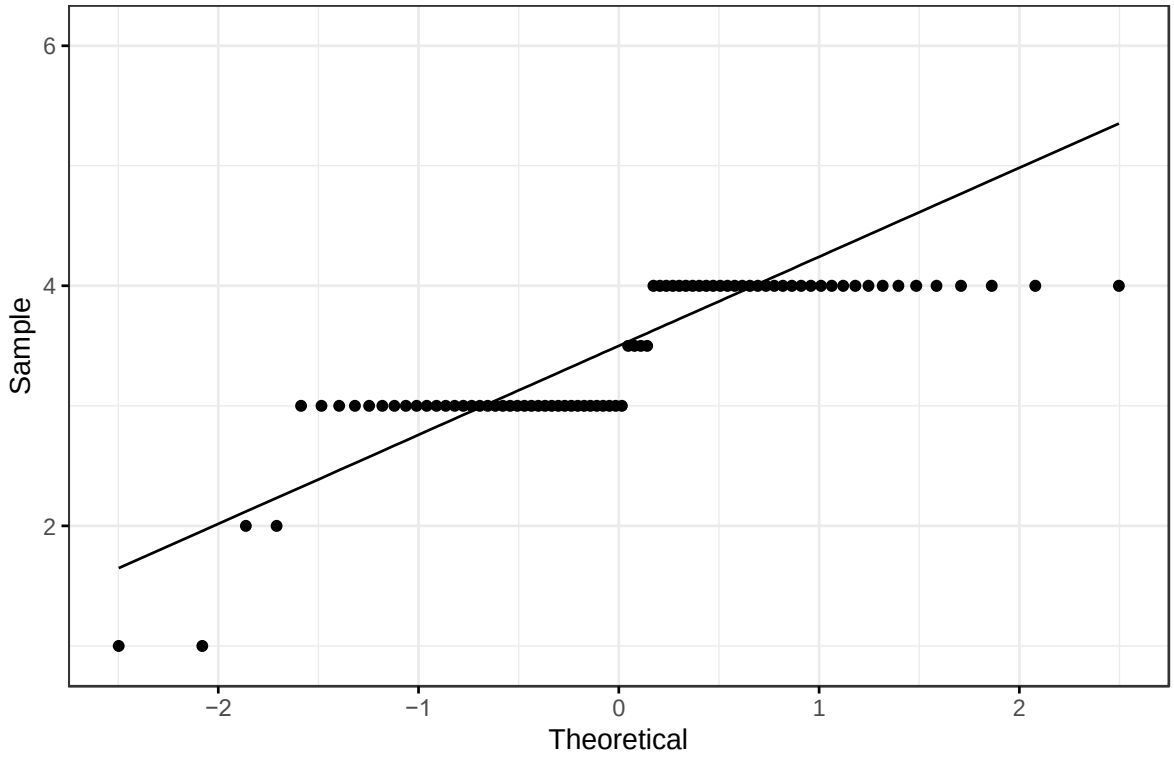


Figure C8

Distribution of Data for Client 8

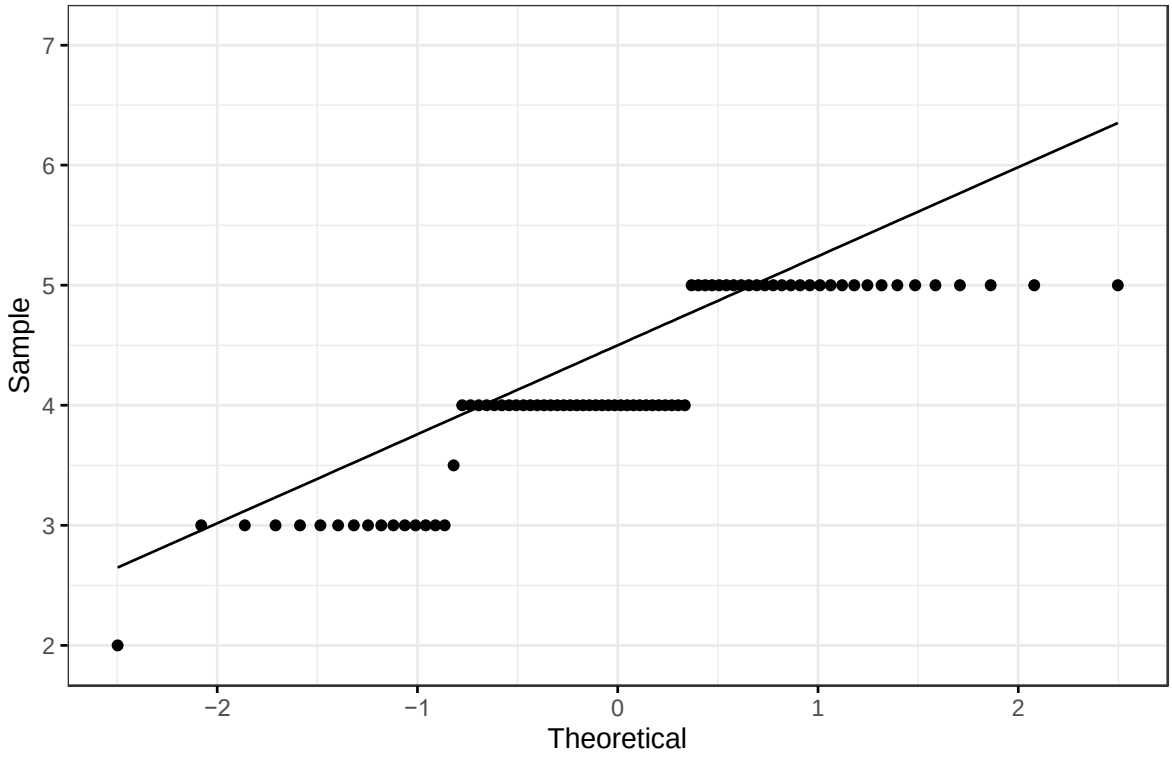


Figure C9

Distribution of Data for Client 9

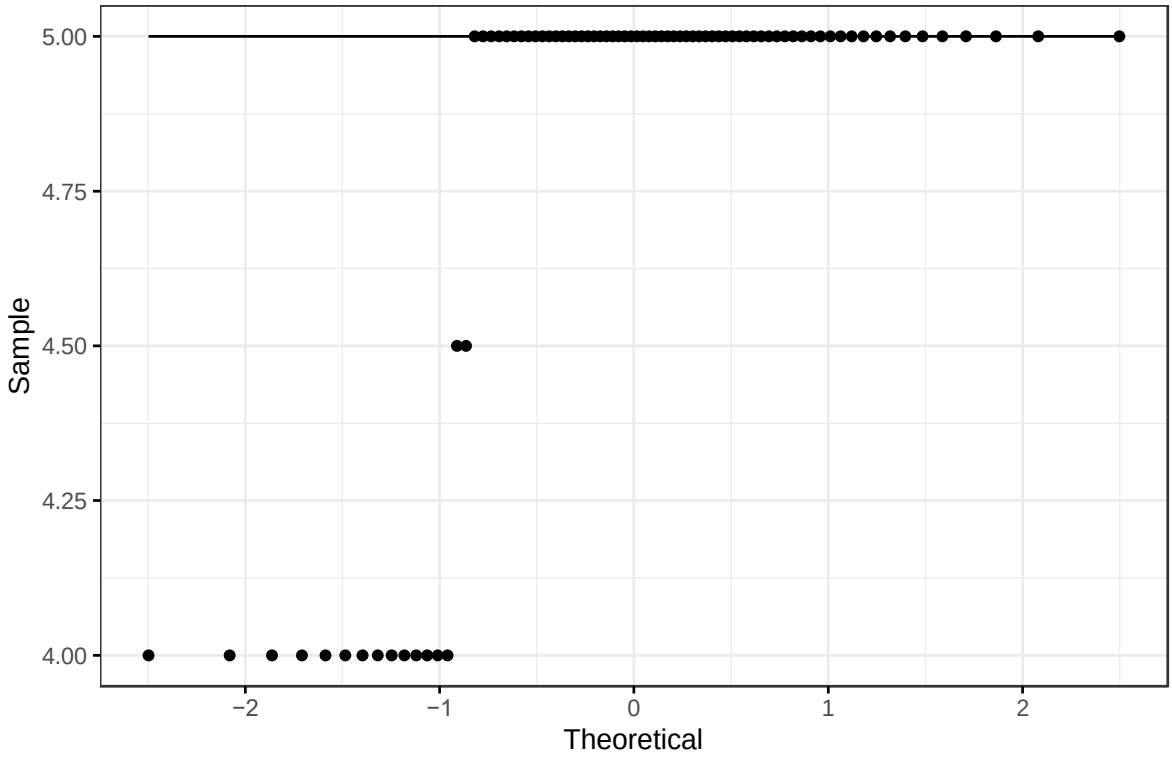
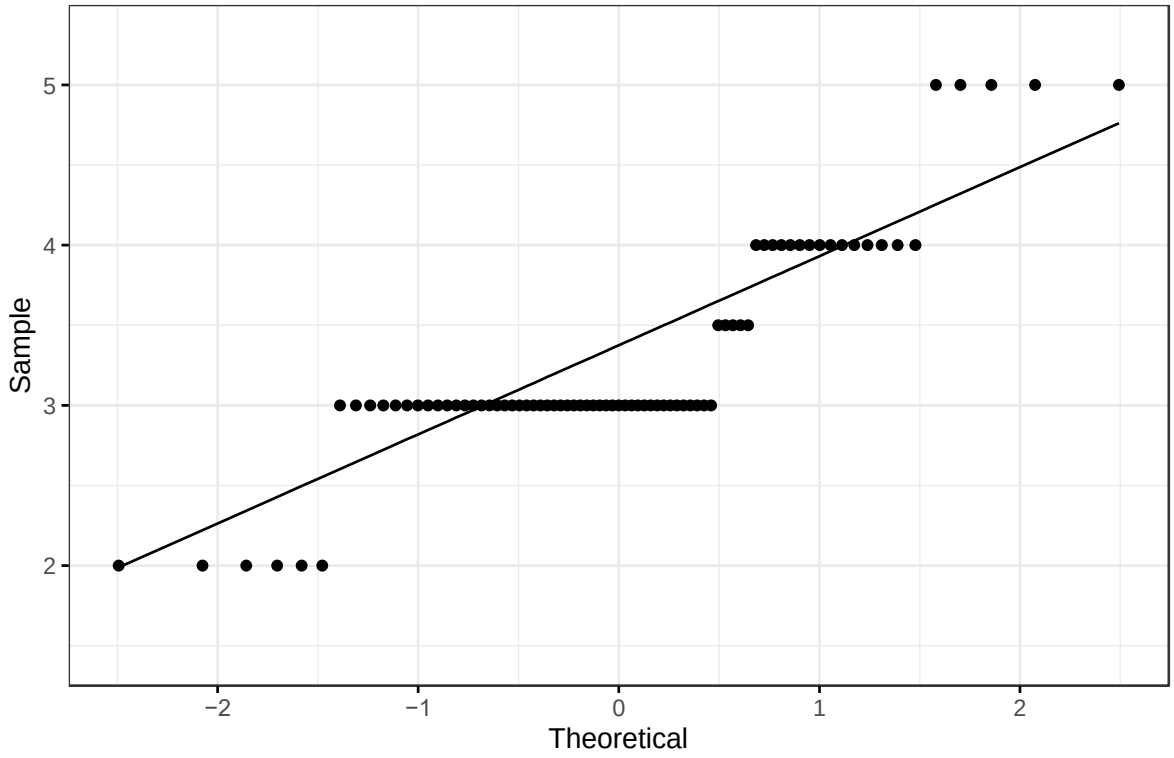


Figure C10

Distribution of Data for Client 10



Appendix D: Plots of Risk Assessments by Assessor for Clients 1, 4, 5, 9 and 10

Figure D1

Plot of Risk Assessment by Assessor for Client 1

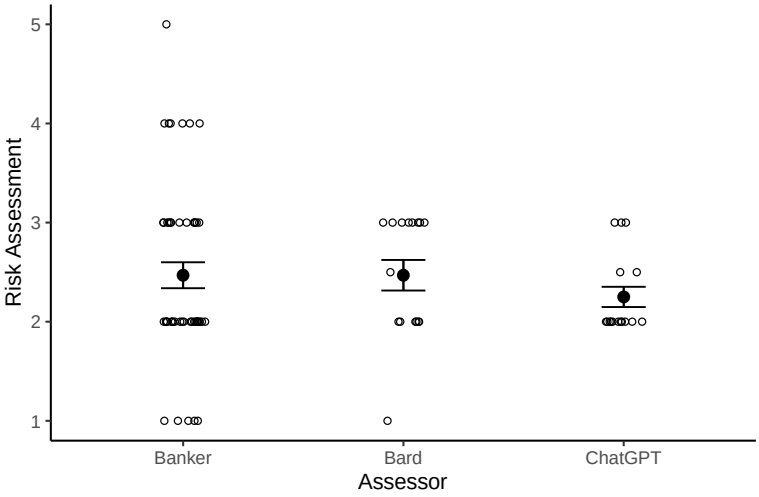


Figure D2

Plot of Risk Assessment by Assessor for Client 4.

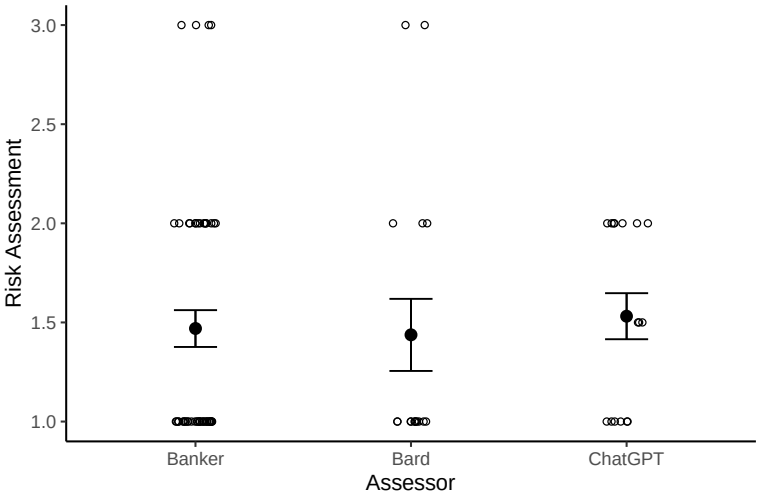


Figure D3

Plot of Risk Assessment by Assessor for Client 5

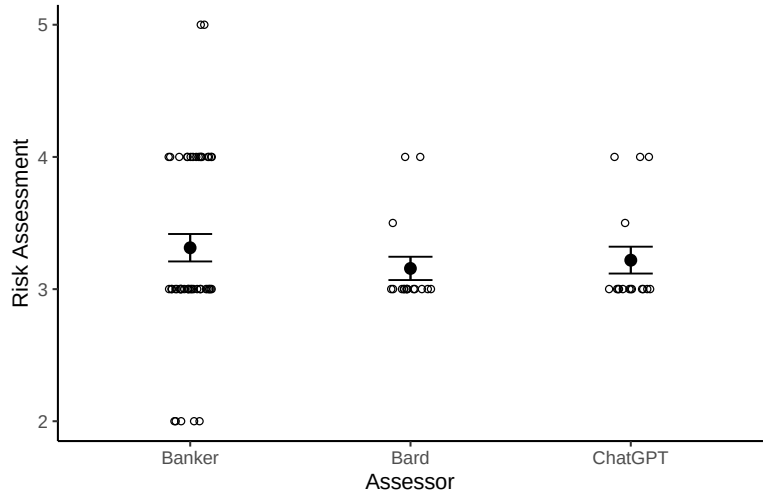


Figure D4

Plot of Risk Assessment by Assessor for Client 9

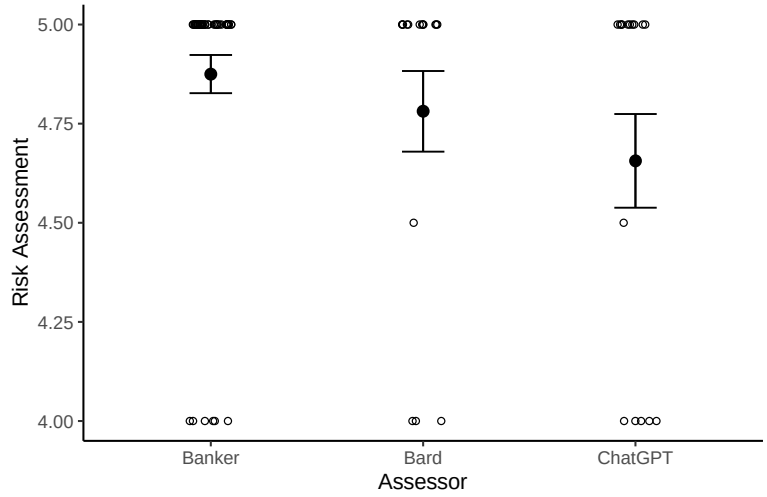
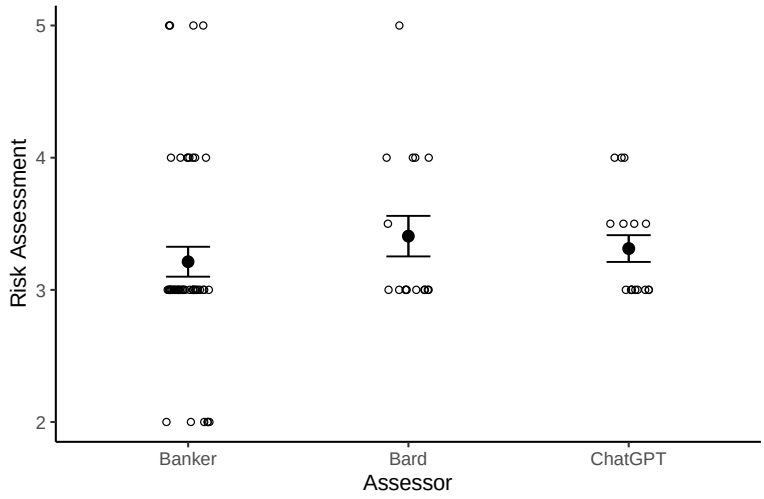


Figure D5

Plot of Risk Assessment by Assessor for Client 10



Appendix E: 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	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	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	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	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	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	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	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	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	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	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.

References

Amstein, A. (2023). *Client cases* (Provided by Thorsten Hens for research purposes. The translated cases can be found in Appendix A.) [datasett].

Bieri, D., Reichmuth, L., Stengel, C., Wickihalder, S., Yilmaz, E., & Ankenbrand, T. (2023). *GPT for Financial Advice*.

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

Bogost, I. (2022). ChatGPT Is Dumber Than You Think. *The Atlantic*. https://www.theatlantic.com/technology/archive/2022/12/chatgpt-openai-artificial-intelligence-writing-ethics/672386/

Brayman, S., Grable, J. E., Griffin, P., & Finke, M. (2017). Assessing a Client's Risk Profile: A Review of Solution Providers. *Journal of Financial Service Professionals*, *71*(1), 71–81.

Cordell, D. M. (2001). RiskPACK: How to Evaluate Risk Tolerance. *Journal of Financial Planning*, *14*(6), 36.

Deng, J., & Lin, Y. (2023). The Benefits and Challenges of ChatGPT: An Overview. *Frontiers in Computing and Intelligent Systems*, *2*(2), 81–83. https://doi.org/10.54097/fcis.v2i2.4465

Elton, E., & Gruber, M. (1997). Modern Portfolio Theory, 1950 to Date. *Journal of Banking & Finance*, *21*, 1743–1759. https://doi.org/10.1016/S0378-4266(97)00048-4

Fan, C., & Zhang, D. (2012). A Note on Power and Sample Size Calculations for the Kruskal-Wallis Test for Ordered Categorical Data. *J Biopharm Stat, 22*(6), 1162–1173. https://doi.org/10.1080/10543406.2011.578313

Fedlex. (2018). *950.1 Federal Act of 15 June 2018 on Financial Services (Financial Services Act, FinSA)*. https://www.fedlex.admin.ch/eli/cc/2019/758/en#tit_2/chap_2/sec_3

Foerster, S., Linnainmaa, J. T., Melzer, B. T., & Previtero, A. (2014). Retail Financial Advice: Does One Size Fit All? *NBER Working Paper Series*, 20712-n/a. ABI/INFORM Global. https://doi.org/10.3386/w20712

Grable, J. (2000). Financial Risk Tolerance and Additional Factors That Affect Risk Taking in Everyday Money Matters. *Journal of Business and Psychology*, *14*, 625–630. https://doi.org/10.1023/A:1022994314982

Grable, J., & Lytton, R. H. (1999). Financial risk tolerance revisited: The development of a risk assessment instrument☆. *Financial Services Review*, *8*(3), 163–181. https://doi.org/10.1016/S1057-0810(99)00041-4

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a Conceptual Framework for Mixed-Method Evaluation Designs. *Educational Evaluation and Policy Analysis*, *11*(3), 255–274. JSTOR. https://doi.org/10.2307/1163620

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

Hens, T., Bachmann, K., & De Giorgi, E. (2018). *Behavioral Finance for Private Banking* (2nd edition). Wiley Finance.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Huang, M.-H., & Rust, R. T. (2018). Artificial Intelligence in Service. *Journal of Service Research*, *21*(2), 155–172. https://doi.org/10.1177/1094670517752459

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

Li, X., Kuncoro, A., d'Autume, C., Blunsom, P., & Nematzadeh, A. (2021). *A Systematic Investigation of Commonsense Understanding in Large Language Models*.

Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., & Liu, Y. (2023). Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study. *arXiv*. https://doi.org/10.48550/arXiv.2305.13860

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

Morse, J. M., & Niehaus, L. (2009). *Mixed Method Design: Principles and Procedures*. Left Coast Press. https://books.google.fr/books?id=zYxWAAAAYAAJ

Nobre, L. H. N., & Grable, J. E. (2015). The Role of Risk Profiles and Risk Tolerance in Shaping Client Investment Decisions. *Journal of Financial Service Professionals*, *69*(3), 18–21.

Official Journal of the European Union. (2014). *Directive 2014/65/EU of the European Parliament and of the Council of 15 May 2014 on markets in financial instruments and amending Directive 2002/92/EC and Directive 2011/61/EU* (Bd. L173). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014L0065&from=EN

Open AI. (2023). *GPT-4*. https://openai.com/research/gpt-4

OpenAI. (2022). *Introducing ChatGPT*. https://openai.com/blog/chatgpt

Osborne, J., & Waters, E. (2002). Four Assumptions of Multiple Regression That Researchers Should Always Test. *Practical Assessment, Research & Evaluation*, *8*.

Pichai, S. (2023). *An important next step on our AI journey*. Goolge. https://blog.google/technology/ai/bard-google-ai-search-updates/

Rice, D. (2005). *Variance in risk tolerance measurement: Toward a uniform solution*. ProQuest Dissertations Publishing.

Roszkowski, M., & Davey, G. (2010). Risk Perception and Risk Tolerance Changes Attributable to the 2008 Economic Crisis: A Subtle but Critical Difference. *Journal of Financial Service Professionals*, *64*.

RStudio Team. (2023). *RStudio: Integrated Development Environment for R*. RStudio, PBC. https://www.rstudio.com/

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

Saunders, M., & Bristow, A. (2023). *2023 Research Methods for Business Students Preface and Chapter 4* (s. xxvii, 128).

Saunders, M., Lewis, P., & Thornhill, A. (2016). *Research methods for business students: 7. Ed* (Seventh edition.). Harlow u.a: Pearson.

Schoonenboom, J., & Johnson, R. (2017). How to Construct a Mixed Methods Research Design. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *69*. https://doi.org/10.1007/s11577-017-0454-1

Sejnowski, T. J. (2023). Large Language Models and the Reverse Turing Test. *Neural Computation*, *35*(3), 309–342. https://doi.org/10.1162/neco_a_01563

Skopeliti, C., & Milmo, D. (2023). ChatGPT needs a huge amount of editing': Users' views mixed on AI chatbot. *The Guardian*.

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

Zaremba, A., & Demir, E. (2023). ChatGPT: Unlocking the Future of NLP in Finance. *Modern Finance*, *1*(1), 93–98. https://doi.org/10.61351/mf.v1i1.43