risks

MDPI

Article

Assessing ChatGPT's Proficiency in Quantitative Risk Management

Marius Hofert

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China;
mhofert@hku.hk

Abstract: The purpose and novelty of this article is to investigate the extent to which artificial intelligence chatbot ChatGPT can grasp concepts from quantitative risk management. To this end, we enter a scholarly discussion with ChatGPT in the form of questions and answers, and analyze the responses. The questions are classics from undergraduate and graduate courses on quantitative risk management, and address risk in general, risk measures, time series, extremes and dependence. As a result, the non-technical aspects of risk (such as explanations of various types of financial risk, the driving factors underlying the financial crisis of 2007 to 2009, or a basic introduction to the Basel Framework) are well understood by ChatGPT. More technical aspects (such as mathematical facts), however, are often inaccurate or wrong, partly in rather subtle ways not obvious without expert knowledge, which we point out. The article concludes by providing guidance on the types of applications for which consulting ChatGPT can be useful in order to enhance one's own knowledge of quantitative risk management (e.g., using ChatGPT as an educational tool to test one's own understanding of an already grasped concept, or using ChatGPT as a practical tool for identifying risks just not on one's own radar), and points out those applications for which the current version of ChatGPT should not be invoked (e.g., for learning mathematical concepts, or for learning entirely new concepts for which one has no basis of comparison to assess ChatGPT's capabilities).

Keywords: ChatGPT; quantitative risk management; risk; risk measures; time series; extremes; dependence

MSC: 68T50

JEL Classification: C45; G32; D83

check for updates

Citation: Hofert, Marius. 2023. Assessing ChatGPT's Proficiency in Quantitative Risk Management. *Risks* 11: 166. https://doi.org/10.3390/risks11090166

Academic Editor: Mogens Steffensen

Received: 6 August 2023
Revised: 15 September 2023
Accepted: 18 September 2023
Published: 19 September 2023

Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

ChatGPT is an artificial intelligence language model developed by OpenAI. *GPT* stands for generative pre-trained transformer. A *transformer* is a specific class of neural networks, which is *pre-trained*, so trained on huge amounts of past observations. Being *generative* refers to ChatGPT's step-by-step generation of *tokens* (a group of characters such as a word or part of a word; use https://platform.openai.com/tokenizer (accessed on 17 September 2023) to see how words are split into tokens) to build sentences, paragraphs or whole articles of text based on the learned predictive distribution of tokens and initial tokens. *Chat* stands for the fact that the transformer was optimized for conversational chat. ChatGPT was launched on 30 November 2022 and had one million users within five days (as a comparison, it took Instagram two and a half months to reach the same userbase), making it the fastest growing consumer application yet.

With ChatGPT widely available by now and the interest from both academia (for example, from an educational point of view for teaching, or by students as an auxiliary tool for learning) and the industry (for example, for summarizing market information on risk or automating internal processes), our goal is to address the following question:

> "How proficient is the current version of ChatGPT
> in aspects of quantitative risk management?"

In answering this question through a conversation with ChatGPT, the reader (besides our own, subjective, assessment) will be able to assess the current status of ChatGPT as a tool to be applied in the context of quantitative risk management with the aforementioned purposes in mind. Remotely related references are the following. Kashettar (2023) mentions opportunities (for decision making based on large amounts of data; for increasing operational effectiveness by freeing risk managers to concentrate on other tasks; and for risk communication via natural language) and challenges (concerning bias in training data, and concerning data security and privacy as risk management can involve private client, as well as corporate-related information) of using ChatGPT in risk management practice. Gomez (2023) puts ChatGPT in the position of Chief Risk Officer and asks it to conduct a basic risk assessment. And Sidorenko (2023) provides sample prompts for ChatGPT in the wider realm of risk management. Note that none of these online articles considers *quantitative* risk management, though.

Our conversation with ChatGPT also has elements of a technical report (when ChatGPT is treated as a new technology), of a user experience report (when ChatGPT is treated like a search engine or Wikipedia page), of an interview or case study (when ChatGPT is treated as a person), and of a pedagogical publication (when ChatGPT is treated like a person) and should thus be readable by a large audience. The interested reader can also consult Hofert (2023), which is written in a similar style, on the topic of pitfalls when working with the stochastic notion of correlation.

Section 2 briefly mentions the setup we used when working with ChatGPT. Section 3 contains our conversation with ChatGPT on various aspects of risk and its modeling. We also analyzed and rated ChatGPT's responses with regard to the above goal. Section 4 provides a (numerical) summary of our assessment of ChatGPT's performance, and Section 5 provides concluding remarks.

2. The Setup

The conversation in this section took place over the course of two weeks in April 2023. Access to the ChatGPT version 3.5 we used was provided by The University of Hong Kong (note that OpenAI does not provide accounts to the public in Hong Kong at the moment). Concerning hyperparameters, ChatGPT allowed to freely choose the following (as most users, we left all as the defaults):

- Max response: 800. The maximal number of tokens used per response (an integer between 1 and 4000).
- Temperature: 0.7. This hyperparameter (a real number between 0 and 1) controls the randomness of the output (from deterministic to creative responses). Low temperature values (0 to 0.3) provide more repetitive and deterministic outputs, medium temperature values (0.3 to 0.7) provide a balance between repetitive and creative outputs, and high temperature values (0.7 to 1) provide highly creative and diverse, but potentially less coherent, outputs.
- Top P: 0.95. Similar to the temperature, this hyperparameter (a real number between 0 and 1) affects the randomness of the output but uses a different method. Low values narrow down the output's token selection to likelier tokens only, large values allow the output's tokens to range from likely to less likely ones. See OpenAI Community (2023) for a comparison of this and the temperature parameter.
- Past messages included: 10. This hyperparameter (an integer between 1 and 20) controls how many past messages are included in each new query, thus providing context for new queries.

Besides a short analysis, we assessed ChatGPT's responses with regard to the aforementioned goal with a 5-star rating in the margins (0 stars = poor answer; 5 stars = excellent answer; half-stars allowed). Furthermore, for readability, we converted ChatGPT's responses to technical questions to LaTeX and formatted the output slightly to save space; as addressed later, we shortened one response. We chose not to ask ChatGPT to provide LaTeX output right away, as it is often not required for understanding an answer, and we did not want to restrict ChatGPT in its freedom to respond. Also note that we did not format our questions in LaTeX for ChatGPT (neither below nor when we asked them).

3. The Conversation and Its Analysis

3.1. Reproducibility

Hi ChatGPT, what version are you?

> As an AI language model, I do not have a specific version number. I am constantly being updated and improved by my developers, so my abilities and knowledge are always expanding.

★★★☆☆ We see that ChatGPT's answer starts with the phrase "As an AI language model", which, by now, is so common that it can even reveal AI usage, for example in Amazon reviews; see Vincent (2023). Due to the usage of ChatGPT, it may become more difficult to detect spam in the near future, but at least cases involving this sentence are easy to spot. However, these sentences may also make it harder for automatic spam detection software to be triggered, as they are more and more common and there is nothing obviously wrong about them.

Of course, seeing ChatGPT's "As an AI language model" phrase does not mean that every response must start like that. In fact, one major problem we learn from ChatGPT's answer is the issue of non-reproducibility; it also plagues other machine learning procedures. There are two aspects to this problem: one concerns the G in ChatGPT. Generating a sequence of tokens from the predictive distribution of tokens is non-reproducible, except when the temperature hyperparameter is chosen as 0, in which case the answer is always the maximum likelihood estimator based on the predictive distribution. Note that the answer may still depend on the past messages considered for creating the current response (the *context*). So asking the same question another time may thus (and indeed often does) lead to a different answer. Second, and worse, not even knowing the rough version in use implies that we do not even know whether or when the predictive distribution potentially changed. In short, neither the exact outcome nor the whole predictive distribution is typically reproducible. The latter, as ChatGPT implies, can change at any point. Asking the same question the next day led to the following answer:

> As an AI language model, I don't have a specific version number as I am constantly updated and improved by my developers. However, my current model is OpenAI's GPT-3.

★★★☆☆ We now learn that ChatGPT is based on GPT-3. The actual version implemented is indeed "GPT-3.5", though.

That is interesting, so what you are answering will not necessarily be reproducible?

> As an AI language model, my responses are generated based on the input I receive and the algorithms that are programmed into me. So while my responses may not be exact replicas of previous responses, they are generated through a consistent set of rules and processes. However, it is possible that some variations may occur due to the nature of machine learning and natural language processing.

★★★★☆ Here, ChatGPT reports that responses are not necessarily reproducible. The reason given ("nature of machine learning and natural language processing") refers to the generative modeling (the first aspect mentioned before). Depending on the predictive distribution (and hyperparameters influencing it), a correct/yes/positive (incorrect/no/negative) answer can potentially change to a incorrect/no/negative (correct/yes/positive) one, leaving the conversation partner with little to no information to learn from but rather some mistrust in the validity of future responses.

But you said you are constantly being updated and improved. So if I ask you the same question in the near future, your answer might be quite different because of such an update or improvement *and not just because of some variation due to the nature of machine learning and natural language processing, right?*

> Yes, that's correct. As an AI language model, I am constantly learning and improving based on the data and feedback provided by my developers. So, if you were to ask me the same question in the future, my response might be different than it is now because I may have learned new information, improved my algorithms, or been updated with new features or capabilities.

★★★★☆ This response refers to the change of the predictive distribution (the second aspect mentioned before) through an update.

Don't you think it would be important to know your version number then? Reproducibility is one of the core principles underlying the scientific method.

> I understand your point, but as an AI language model, my development and improvement process is continuous and dynamic. Therefore, I don't have a single version number like traditional software applications do. Instead, my developers use a version control system to track changes and updates to my algorithms and capabilities. While reproducibility is indeed important in scientific methods, the nature of machine learning and natural language processing is such that my responses can vary based on the input and context, even with the same version number. Therefore, my developers focus on ensuring that I produce reliable and accurate responses based on my current state, rather than on version numbers.

★★★★☆ Pressed, ChatGPT reveals that its development is based on a version control system (as are most, if not all, bigger software development projects). As such, knowing the version number (the unique identifier of ChatGPT's released software) would at least allow one to track major changes in the predictive distribution and thus alleviate the problem of non-reproducibility.

3.2. Risk

With the reproducibility issue in mind, we now start to ask about risk in general and aspects of its modeling in particular.

Can you define risk, please?

> Risk can be defined as the probability or likelihood of an event or circumstance occurring, and the potential negative consequences or impacts that may result from that event. In other words, risk is the chance that something undesirable or harmful may happen, and the degree of uncertainty or unpredictability associated with that event. Risk can be assessed and managed in various contexts, such as financial investments, business operations, health and safety, and other areas where there is a need to weigh potential benefits against potential harm.

★★★★⯪ This is a rather good answer, given the open-endedness of the question. It is hard to come up with a single-sentence definition of risk, as there are various types of risks (environmental, health, safety, financial, insurance, etc.) that have different meanings in different contexts; see also Kloman (1990). This aspect was well captured.

I'd like to talk about financial and insurance risk modeling. Can you define market risk, credit risk, operational risk, systemic risk and model risk, please?

> Sure, I can provide definitions for each of those types of risks in the context of financial and insurance risk modeling:
>
> 1. Market risk: Refers to the potential for financial losses due to changes in market conditions, such as interest rates, exchange rates, stock prices, and commodity prices. Market risk is typically associated with investments in financial assets, such as stocks, bonds, and derivatives.
> 2. Credit risk: Refers to the potential for financial losses due to the failure of a borrower to repay a loan or meet other financial obligations. Credit risk

> is typically associated with lending activities and is influenced by factors such as the creditworthiness of the borrower, the terms of the loan, and the economic environment.
>
> 3. Operational risk: Refers to the potential for financial losses due to failures or disruptions in an organization’s internal processes, systems, or controls. Operational risk can arise from a variety of sources, including human error, technology failures, fraud, and external events.
> 4. Systemic risk: Refers to the potential for widespread financial disruption or instability due to the failure of one or more large financial institutions or other systemic factors. Systemic risk can be difficult to predict and manage, as it can be influenced by a wide range of factors, including macroeconomic conditions, regulatory policies, and market sentiment.
> 5. Model risk: Refers to the potential for financial losses due to errors or limitations in the mathematical or statistical models used to assess and manage financial risks. Model risk can arise from a variety of sources, including incomplete or inaccurate data, flawed assumptions, or the misuse or misinterpretation of models. Managing model risk requires ongoing monitoring and validation of models, as well as robust governance and control frameworks.

★★★★⯪ The first part of our question provides a context for ChatGPT (financial and insurance risk) in which to answer our actual question. The definition of market and credit risk is very good. For operational risk, a precise definition has been known since its introduction in (BIS 2006, p. 144): “Operational risk is defined as the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events. This definition includes legal risk, but excludes strategic and reputational risk”. One would have, perhaps, expected ChatGPT to provide this definition. The answer did provide all aspects of the main definition though; just the inclusion/exclusion of certain types of risk is missing. Systemic risk was overall captured well, too. Given that model risk is not easy to model and there is much less literature available on model risk, ChatGPT also captured this type of risk very well.

Can you briefly summarize the financial crisis of 2007 to 2009, provide reasons for why it happened and address which of the above five types of risk played a role in its unfolding?

> The financial crisis of 2007 to 2009 was a major global economic downturn that resulted in widespread financial losses, bank failures, and a severe recession in many countries. The crisis was triggered by a combination of factors, including:
>
> 1. The housing market bubble: A surge in housing prices in the early 2000s led to a boom in subprime mortgage lending, which allowed many people with poor credit histories to buy homes they could not afford.
> 2. Risky financial practices: Banks and other financial institutions engaged in risky lending and investment practices, such as securitizing subprime mortgages into complex financial instruments known as collateralized debt obligations (CDOs), which were then sold to investors.
> 3. Lack of regulation and oversight: The financial sector was largely unregulated, and government agencies failed to adequately monitor and regulate risky financial practices.
> 4. Global economic imbalances: A lack of balance in international trade and investment flows, particularly between the US and China, contributed to global economic instability.
>
> All of the above five types of risk played a role in the unfolding of the financial crisis of 2007 to 2009. Market risk was a factor, as the collapse of the housing market and the value of CDOs led to massive losses for investors. Credit risk was also a factor, as many subprime borrowers defaulted on their mortgages, leading

> to widespread losses for lenders. Operational risk was present in the failure of banks and other financial institutions, while systemic risk was a factor in the global nature of the crisis and its impact on the broader economy. Finally, model risk played a role in the crisis, as flawed assumptions and flawed models were used to assess and manage financial risks.

★★★★★ For a first dive into this difficult topic, this response is quite good, which may be explained through the vast amount of literature available on the subject that must have entered ChatGPT's training data in some form or another. We can also try to dig into more details from here.

Why was there a surge in housing prices in the early 2000s?

> There were several factors that contributed to the surge in housing prices in the early 2000s, which ultimately led to the housing market bubble that helped trigger the financial crisis of 2007 to 2009. Some of the main factors that contributed to the housing price boom include:
>
> 1. Easy access to credit: In the early 2000s, it became easier for people to obtain mortgages and other forms of credit, as banks and other lenders relaxed their lending standards and offered more subprime mortgages to borrowers with poor credit histories.
> 2. Low interest rates: The Federal Reserve kept interest rates low in the early 2000s in an effort to boost the economy after the dot-com bubble burst in 2000. Low interest rates made it cheaper for people to borrow money and led to a surge in demand for housing.
> 3. Speculation and investor demand: The housing market became a popular investment vehicle, as many people bought homes with the expectation that they would be able to sell them for a profit in the future. This speculative demand drove up housing prices.
> 4. Demographic shifts: The early 2000s saw a demographic shift as the baby boomer generation reached their peak earning years and began to invest more heavily in real estate. This increased demand for housing also contributed to rising prices.
>
> All of these factors combined to create a housing market bubble that eventually burst, leading to widespread foreclosures, bank failures, and a severe recession.

★★★★★ Also here, ChatGPT's response is great for learning about the reasons behind the rising housing prices, even relating the financial crisis to the dot-com bubble. Certainly, for some expert readers, these answers may not be satisfying, but for a, say, graduate student encountering this large topic for the first time and trying to glean some pointers on the underlying factors, this answer is more than adequate; also, experts could of course just ask more specific follow-up questions. As a fact check, one could continue with a quick online search on the mentioned points. The lack of "As an AI language model" is refreshing. But let us dig a bit deeper into one of the more technical aspects of the financial crisis.

You already mentioned that CDOs played a role. What is a CDO?

> A CDO, or collateralized debt obligation, is a complex financial instrument that is made up of a pool of various debt securities, such as bonds, mortgages, and loans. CDOs are typically created by investment banks or other financial institutions, which pool together these securities and then divide them into different tranches or levels of risk.
>
> Each tranche of a CDO is assigned a different credit rating based on its level of risk. The highest-rated tranches are considered the safest and are paid first, while the lower-rated tranches are riskier but offer higher returns. CDOs can be structured in a variety of ways, but typically involve one or more layers of debt, with the riskiest debt at the bottom and the safest debt at the top.

> One of the key factors that contributed to the financial crisis of 2007 to 2009 was the use of CDOs to package and sell subprime mortgages. Many CDOs were created using subprime mortgages, which were high-risk loans made to borrowers with poor credit histories. Investment banks packaged these subprime mortgages into CDOs and sold them to investors, often with inflated credit ratings that did not accurately reflect the true risk of the underlying loans.
>
> When the housing market bubble burst and many borrowers defaulted on their mortgages, the value of the CDOs plummeted, leading to massive losses for investors and contributing to the broader financial crisis.

★★★★★ Another good answer. Understanding CDOs is non-trivial, and they are typically better explained with the help of pictures of snapshots of a portfolio of credit risks at different points in time, showing the waterfall principle of defaults affecting (or not affecting) the different tranches of the portfolio. But for a purely text-based description, the above answer is good. Note that, when asked, ChatGPT produced an ASCII plot of the tranches and labeled them, but that plot did not convey enough information to explain CDOs properly. When asked to provide a proper diagram, an output of the form "![CDO diagram] (https://i.imgur.com/[...].png)" was produced, but the link only provided a message "The image you are requesting does not exist or is no longer available", a technical issue perhaps connected to our access to ChatGPT.

Do you know why, mathematically, CDOs were mispriced?

> As an AI language model, I do not have personal beliefs or opinions. However, here is an explanation for why CDOs were mispriced:
>
> CDOs were mispriced due to the flawed assumptions made by the rating agencies and investors about the correlation between the underlying assets. The rating agencies assumed that the default risk of the individual assets was uncorrelated, and therefore, the risk of the entire portfolio could be diversified away. However, in reality, the assets were highly correlated, and the default of a few assets could trigger a cascade of defaults across the portfolio. This led to the underestimation of the risk of CDOs, resulting in their mispricing. Additionally, the complexity of CDO structures made it difficult for investors to fully understand the underlying risks, which further contributed to the mispricing.

★★★★★ Here, we see a frequent phenomenon, namely that ChatGPT starts with a disclaimer. It casts some doubt on ChatGPT's confidence in the answer (rightfully so as we will see later). A more explicit rating of its own confidence in the answer would be extremely useful (or knowing some related references). Nevertheless, the answer provided is fine. It would have provided sufficient information to set off all alarm bells had it been available prior to 2007 and had it not been ignored, but of course that is easy to claim after the fact, where there is now a vast amount of literature on the subject and people have (hopefully) learned their lessons.

Do you know the name of a CDO pricing model that was used around that time?

> Yes, one of the popular CDO pricing models used around that time was the Gaussian Copula model. It was widely used by rating agencies and investors to estimate the probability of default and expected losses for CDO tranches. The model was based on the assumption that the default correlation between the underlying assets followed a normal distribution, which was found to be a flawed assumption during the financial crisis. The model was criticized for not taking into account the tail risks and the extreme events that could lead to a significant loss in the CDO portfolio.

★★★★⯪ Another good text-based answer, albeit with a slight inaccuracy. The described model is an intensity-based default model, where the time of default τ_i of component i in the portfolio is modeled as the first time the component's survival probability falls below a random trigger $U_i \sim \mathrm{U}(0,1)$. The default correlation between two components i, j is the correlation (so a number) between the two (random) default indicators on or before a fixed time point t, so $\mathrm{cor}(\mathbb{1}_{\{\tau_i \le t\}}, \mathbb{1}_{\{\tau_j \le t\}})$. As a correlation, it cannot "follow a normal distribution". What was assumed to follow the copula of the normal distribution (namely the Gaussian copula, hence the name of the model) was the vector $(U_1, \ldots, U_d)$ of the total d components of the credit portfolio. Here, we already see the predictive nature of ChatGPT; it seems "smart enough" to understand CDOs but then does not know that a correlation following a normal distribution makes little sense in this context.

What is the Basel Framework and why was it introduced?

> The Basel Framework is a set of international banking regulations that provide guidelines for banks to assess and manage their risks. The framework was introduced by the Basel Committee on Banking Supervision, which is a group of central bank governors and regulators from around the world, after the collapse of several banks in the 1980s and 1990s.
>
> The Basel Framework was introduced to strengthen the stability and soundness of the global banking system by setting minimum capital requirements for banks. The framework provides guidelines for banks to assess and manage their risks, including credit risk, market risk, and operational risk. Banks are required to hold a minimum amount of capital based on their risk profile to ensure that they can absorb losses and remain solvent.
>
> The Basel Framework has undergone several revisions, with the latest version being Basel III. Basel III introduced stricter capital requirements and introduced new measures to address liquidity and leverage risks. The framework has been widely adopted by countries around the world, and compliance with the framework is mandatory for banks that operate internationally.

★★★★★ Another good answer to get the gist of the Basel Framework. Again, let us get more specific from here.

On the modeling side, what were the major components introduced with Basel II?

> Basel II was a comprehensive revision of the Basel Framework that introduced several changes to the way banks model and manage their risks. The major components introduced with Basel II were:
>
> 1. Pillar 1: Minimum Capital Requirements—Basel II introduced a new risk-based capital framework that required banks to hold a minimum amount of capital based on their risk profile. The framework divided risks into three categories—credit risk, market risk, and operational risk—and provided guidelines for calculating the amount of capital required for each type of risk.
> 2. Pillar 2: Supervisory Review Process—Basel II introduced a new supervisory review process that required banks to assess their overall risk profile and develop a comprehensive risk management strategy. Banks were required to undergo regular reviews by their regulators to ensure that they were adequately managing their risks.
> 3. Pillar 3: Market Discipline—Basel II introduced new disclosure requirements that required banks to provide more detailed information about their risk profile, capital adequacy, and risk management practices to investors and other stakeholders. The goal was to increase market discipline and encourage banks to adopt better risk management practices.
> 4. Internal Ratings-Based (IRB) Approach—Basel II introduced a new approach for calculating credit risk that allowed banks to use their own internal models to estimate the probability of default and expected losses for their loan portfolios. The IRB approach provided more flexibility and allowed banks to better tailor their risk management strategies to their specific risk profile.

> Overall, Basel II was a significant change from the previous Basel I framework, as it introduced a more sophisticated and risk-sensitive approach to banking regulation that aimed to improve the stability and soundness of the global banking system.

★★★★☆ Also to this question, ChatGPT provides an overall good answer. One could easily dig deeper from here if so desired. One should note, though, that ChatGPT omitted to mention the introduction of the standardized approach to credit risk modeling (in parallel to the IRB approach).

3.3. Risk Measures

We now turn to the topic of risk measures, typically applied for determining risk capital to account for future losses.

What is a risk measure?

> A risk measure is a quantitative measure that is used to assess the level of risk associated with an investment or portfolio. Risk measures are used to estimate the potential losses that may be incurred due to adverse market movements or other factors that may affect the value of the investment.
>
> There are several types of risk measures, including:
>
> 1. Value at Risk (VaR)—VaR is a measure of the maximum potential loss that an investment or portfolio may incur over a specified time period with a given level of confidence. For example, a VaR of 95% for a portfolio of \$1 million over one day would indicate that there is a 95% chance that the portfolio will not lose more than \$X over one day.
> 2. Expected Shortfall (ES)—ES is a measure of the expected loss beyond the VaR level. It provides an estimate of the average loss that an investment or portfolio may incur in the worst-case scenario.
> 3. Tail Risk Measures—Tail risk measures are designed to capture the risk of extreme events that fall outside the normal distribution of returns. These measures are typically used to assess the risk of rare events that may have a significant impact on the value of the investment.
> 4. Stress Tests—Stress tests are simulations that are used to evaluate the impact of adverse market conditions or other events on the value of an investment or portfolio. Stress tests are used to identify potential vulnerabilities and to evaluate the effectiveness of risk management strategies.
>
> Overall, risk measures are an important tool for investors and risk managers to assess and manage the risks associated with their investments.

★★★⯪☆ The first part of the answer is not an exact definition (unless the vague "quantitative measure" is interpreted as "number"), at which point the inclusion of examples (which would have been the topic of our next question) is helpful for learning about risk measures. The last part outlines applications of risk measures, which is also helpful. However, some inaccuracies and errors are present in the examples. The value at risk (VaR) at level α is simply the α-quantile of the loss distribution. As such, it is the smallest level which is exceeded with a probability of at most $1 - \alpha$, and this is the formulation we prefer for teaching the concept, as it conveys the idea that VaR is indeed (expected to be) exceeded (by losses) but that this exceedance probability is limited. Speaking of a "measure of the maximum potential loss" is not ideal, as losses exceeding VaR can (mathematically, in principle) be unbounded. The "given level of confidence" at the end of the sentence clarifies this to some degree but does not explain that the level of confidence refers to a probabilistic interpretation (and therefore leaves the door open for misinterpretation). We are not surprised by this fact; these inaccuracies often lead to misinterpretations of VaR in practice. In fact, major books on the subjects define VaR as "worst/maximum loss…

not exceeded with a given probability”; see Jorion (2006), Kosowski and Neftci (2014) or sciencedirect.com/topics/social-sciences/value-at-risk (accessed on 17 September 2023).

The definition of expected shortfall (ES) is rather reminiscent of that of conditional tail expectation (CTE). The two are only necessarily equivalent for continuous loss distributions; see (McNeil et al. 2015, Proposition 8.13). Interestingly, the rather inaccurate “ES is a measure of the expected loss beyond the VaR level” (instead of a more accurate definition detailing what “a measure” means here) corrects this again, as ES is indeed mainly a “measure” of the expected loss beyond VaR and not the expected loss beyond VaR itself, but this leaves the reader with an ambiguous (rather than a proper) definition, so being a rather rough explanation. We are also not surprised here since ES and CTE are equal for continuous loss distributions, being the most frequently found in financial risk modeling. Hence, the two concepts are often used interchangeably in the literature.

The inaccuracies continue with tail risk measures. The “normal distribution of returns” could easily be confused with the inadequate distributional assumption of normality of log-returns of financial time series; see (McNeil et al. 2015, chp. 3). Replacing “normal” by “typical” already allows to interpret this sentence as ‘much less wrong’. The fourth example is fine. Finally, you may have spotted a typographical error in the example for VaR. Let us (try to) have it fixed.

In your first list item, $X is undefined. Can you provide a correction of that item, please?

> I apologize for the error. Here’s a corrected version of the first list item:
>
> 1. Value at Risk (VaR)—VaR is a measure of the maximum potential loss that an investment or portfolio may incur over a specified time period with a given level of confidence. For example, a VaR of 95% for a portfolio of $1 million over one day would indicate that there is a 95% chance that the portfolio will not lose more than $X (which represents the VaR amount) over one day.

★☆☆☆☆ No correction. Another trial to correct the error let to the same outcome (no correction). This becomes frustrating and revealing fairly quickly. No one expects a chatbot to be perfect, but if you provide a valid correction to no effect, you know you are not having a conversation with a reasonably intelligent human being, which feels almost sad at this point, given how well ChatGPT has performed so far.

We now dive a bit further into the distinction between ES and CTE.

Is expected shortfall indeed the expected loss beyond the VaR level or do you require any assumption on the underlying loss distribution for this result to be mathematically correct?

> You are correct that my previous definition of Expected Shortfall (ES) was not entirely accurate. ES, also known as Conditional Value at Risk (CVaR), is defined as the expected loss beyond the VaR level. However, to calculate ES, it is necessary to make some assumptions about the underlying distribution of losses, such as assuming that the losses are normally distributed or that they follow a certain parametric distribution.
>
> ES is a more comprehensive risk measure than VaR because it takes into account the severity of losses beyond the VaR level, rather than just the probability of their occurrence. ES provides an estimate of the average loss that an investment or portfolio may incur in the worst-case scenario.
>
> To calculate ES, one typically first calculates VaR and then estimates the expected loss given that the loss exceeds the VaR level. This can be done using a variety of methods, such as Monte Carlo simulation or historical data analysis, depending on the assumptions made about the underlying distribution of losses.
>
> I apologize for any confusion my previous response may have caused.

★★★★☆ We see that the flawed definition of ES remained unchanged. However, ChatGPT correctly hints at the fact that the loss distribution has to have an integrable right tail in order for

ES to exist. But then, the end of the first paragraph does not mention this specifically, it rather mentions examples (the normal being far too restrictive and "certain parametric distributions" being too vague). The first sentence of the second paragraph is surprisingly accurate. Not only do we learn about the frequency-based nature of VaR and the severity-based nature of ES, we now also precisely learn the connection between VaR and its exceedance probability.

We conclude that letting ChatGPT compare different concepts may lead to more precise results than just asking about the concepts individually. This makes ChatGPT focus on the (often more pronounced) differences of concepts rather than providing broad-brush approach answers.

Strictly speaking, losses to follow certain parametric distributions does not imply expected shortfall to be conditional tail expectation. The two are equal for continuous loss distributions. Can you confirm that?

> You are correct. Losses following certain parametric distributions does not necessarily imply that the expected shortfall (ES) is equivalent to the Conditional Tail Expectation (CTE).
>
> CTE is a risk measure that represents the expected value of the loss in the tail of the loss distribution beyond a certain threshold, which is usually the VaR level. CTE differs from ES in that it considers all losses beyond the VaR level, rather than just the losses that exceed the VaR level.
>
> While ES and CTE are often used interchangeably, they are not equivalent measures in general. However, for continuous loss distributions, ES and CTE represent the same risk measure, which is the conditional expectation of the loss given that the loss is greater than or equal to the VaR level.
>
> Therefore, it is important to understand the underlying assumptions and limitations of each risk measure when using them for risk management and decision-making purposes.

★★★★⯪ Overall a good answer, with the exception being the last sentence of the second paragraph, which is more confusing than helpful.

VaR is not subadditive in general. Can you give example scenarios in which VaR is known to be additive or superadditive?

> VaR (Value at Risk) is not subadditive in general, which means that the VaR of a portfolio may be greater than the sum of the VaRs of its individual components. However, there are some scenarios in which VaR is known to be additive or superadditive:
>
> 1. Additive VaR: If the individual components of a portfolio are independent and identically distributed, then the VaR of the portfolio is equal to the sum of the VaRs of its individual components. This is known as additive VaR.
> 2. Superadditive VaR: If the individual components of a portfolio are positively correlated, then the VaR of the portfolio may be greater than the sum of the VaRs of its individual components. This is known as superadditive VaR.
>
> For example, consider a portfolio consisting of two stocks. If the stocks are perfectly positively correlated, then the VaR of the portfolio would be greater than the sum of the VaRs of the individual stocks. This is because if one stock experiences a large loss, it is likely that the other stock will also experience a loss, resulting in a larger overall loss for the portfolio.
>
> In summary, VaR is not generally subadditive, but can be additive or superadditive depending on the correlation between the components of the portfolio.

★★☆☆☆ The first paragraph is good; it also contains the definition of superadditive risk measures. However, the second paragraph claims that for iid losses, VaR is subadditive. This is not true; take $X_1, X_2 \overset{\text{ind.}}{\sim} \mathrm{Par}(1/2)$ with distribution function $F(x) = 1 - x^{-1/2}$, $x \geq 1$, so iid Pareto Type I distributed losses. One can then show that VaR is superaddive (even) for all confidence levels; see (Hofert et al. 2020, Exercise 2.28). Note that comonotone losses lead to additive VaR; see (McNeil et al. 2015, Proposition 7.20). The third paragraph stays vague again, with its formulation "If the individual components [...] are positively correlated, then the VaR of the portfolio may be greater than the sum of the VaRs", note the "may". Indeed, this may or may not be the case. For example, if the losses are jointly elliptically distributed, then VaR is superadditive for confidence levels in $(0, 1/2)$, and it is subadditive for confidence levels in $[1/2, 1)$. Since this holds irrespective of the correlation between the components of the elliptical distribution, it also provides a counterexample to the example ChatGPT gives in the fourth paragraph. Note that in practice, one often uses "correlation" synonymously with "dependence" (rather than "linear correlation"). However, even with this interpretation of "correlation", the fourth paragraph is wrong in general (we already mentioned that VaR is comonotone additive). In that regard, also the last sentence is wrong since additivity or superaddivitity does not necessarily depend on the correlation (or dependence) of the losses. For example, for $X_1, X_2 \overset{\text{ind.}}{\sim} \mathrm{Exp}(1)$, VaR can be superadditive (for small confidence levels), additive (for precisely one confidence level), or subadditive (for large confidence levels).

Knowing that VaR is in general not additive for iid losses, let us try to see why ChatGPT may think this holds; recall that we converted ChatGPT's text-based answers to LaTeX expressions with a bit of formatting for increased readability.

Can you provide a mathematical proof for your statement that VaR is additive for iid losses?

> Yes, I can provide a mathematical proof for the statement that VaR is additive for iid losses.
>
> Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables representing the losses of n individual components of a portfolio. Let $w_1, w_2, \ldots, w_n$ be the weights of the respective components, such that $w_1 + w_2 + \cdots + w_n = 1$. The portfolio loss is given by $Y = w_1 X_1 + w_2 X_2 + \cdots + w_n X_n$. The distribution of Y is the convolution of the distributions of $X_1, X_2, \ldots, X_n$, which is also iid. Therefore, the cumulative distribution function (CDF) of Y is given by:
>
> $$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(w_1 X_1 + w_2 X_2 + \cdots + w_n X_n \leq y)$$
>
> Since the losses are independent, the joint probability density function (PDF) of $X_1, X_2, \ldots, X_n$ is the product of their individual PDFs:
>
> $$f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \ldots f_{X_n}(x_n)$$
>
> We can then use the convolution formula to compute the PDF of Y:
>
> $$\begin{aligned} f_Y(y) = \int \cdots \int & f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) \\ & \cdot \delta(y - w_1 x_1 - w_2 x_2 - \cdots - w_n x_n n) \, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n \end{aligned}$$
>
> where δ is the Dirac delta function. Now, we can calculate the VaR of Y at level α as the value y_α such that: $\mathbb{P}(Y \leq y_\alpha) = \alpha$. Using the CDF of Y, we have:
>
> $$\begin{aligned} \alpha &= \mathbb{P}(Y \leq y_\alpha) = F_Y(y_\alpha) \\ &= \int \cdots \int f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) \\ &\qquad \cdot \delta(y_\alpha - w_1 x_1 - w_2 x_2 - \cdots - w_n x_n n) \, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n. \end{aligned}$$

Since the losses are iid, we can change the order of the integral and write:

$$\alpha = \int \cdots \int f_X(x) \cdot \delta(y_\alpha - wx)\,\mathrm{d}x$$

where $f_X(x)$ is the PDF of X, and $w = w_1 X_1 + w_2 X_2 + \cdots + w_n X_n$ is the weighted average of the losses. The integral on the right-hand side is the CDF of the random variable wx. Therefore, we have: $\alpha = \mathbb{P}(wx \leq y_\alpha) = F_X(y_\alpha / w)$, where F_X is the CDF of X. Solving for y_α, we get: $y_\alpha = w\,\mathrm{VaR}_X(\alpha)$, where $\mathrm{VaR}_X(\alpha)$ is the VaR of X at level α. Therefore, the VaR of the portfolio Y at level α is:

$$\mathrm{VaR}_Y(\alpha) = y_\alpha = w\,\mathrm{VaR}_X(\alpha) = w_1\,\mathrm{VaR}_X(\alpha) + w_2\,\mathrm{VaR}_X(\alpha) + \cdots + w_n\,\mathrm{VaR}_X(\alpha)$$

which shows that VaR is additive for iid losses.

★☆☆☆☆ This “proof” has several flaws and is overall wrong, of course. For example, “The distribution of Y is the convolution” is not quite true, as the weights $w_1, \ldots, w_n$ are involved. The “Therefore” is unnecessary, as this holds without any connection to convolutions. The use of only two integral symbols but three integrator terms is also questionable, in addition to the additional “n” at the end of the integrand; denoting the already introduced Y by the (lowercase!) w; replacing the intergration variables $x_1, \ldots, x_n$ by the random variables $X_1, \ldots, X_n$ from the second-last to the last displayed equation; and suddenly speaking of X even though it has not been introduced. More serious mistakes can be found in the last displayed equation. First, w is a random variable, so $y_\alpha = w\,\mathrm{VaR}_X(\alpha)$ only holds if w is constant almost surely. Also, the next/last equation is wrong in this (and other) sense(s).

Let us be optimistic and try another logical argument, the correct answer being 1; see (Hofert et al. 2020, Exercise 2.23 b).

The shortfall-to-quantile ratio is the limit for alpha to 1 from below of expected shortfall at confidence level alpha divided by value-at-risk at confidence level alpha. Compute this quantity for a standard normal loss distribution.

To compute the shortfall-to-quantile ratio for a standard normal loss distribution, we need to first calculate the expected shortfall and value-at-risk at a given confidence level alpha. For a standard normal distribution, the expected shortfall and value-at-risk can be calculated as follows:

Expected Shortfall (ES) at $\alpha = -1/\alpha \cdot (1/\sqrt{2\pi}) \cdot \exp(-z_\alpha^2/2)$
Value-at-Risk (VaR) at $\alpha = -z_\alpha$

where z_α is the α-quantile of the standard normal distribution.

For example, if we want to calculate the shortfall-to-quantile ratio at a confidence level of 95%, then $\alpha = 0.05$ and $z_\alpha = 1.645$ (from standard normal tables). Using the above formulas, we can calculate the ES and VaR at a 95% confidence level as follows:

ES at 95% $= -1/0.05 \cdot (1/\sqrt{2\pi}) \cdot \exp(-1.645^2/2) = -0.975$
VaR at 95% $= -1.645$

Therefore, the shortfall-to-quantile ratio at a 95% confidence level for a standard normal distribution is: Shortfall-to-Quantile Ratio at 95% = ES at 95%/VaR at 95% $= -0.975/-1.645 = 0.593$

So, the shortfall-to-quantile ratio for a standard normal loss distribution at a 95% confidence level is 0.593.

★★☆☆☆ This is not what we were interested in.

I am not interested in the confidence level 95% but the limit of the shortfall-to-quantile ratio for the confidence level alpha to go to 1 from below.

> Ah, I see. In that case, we can use the fact that for a standard normal distribution, as the confidence level α approaches 1 from below, the corresponding quantile z_α approaches infinity. Therefore, we can use the following limiting expressions for the ES and VaR:
>
> lim as $\alpha \to 1-$: ES at $\alpha = -1/\alpha \cdot (1/\sqrt{2\pi}) \cdot \exp(-z_\alpha^2/2) = -\infty$
> lim as $\alpha \to 1-$: VaR at $\alpha = -z_\alpha = \infty$
>
> Using these limiting expressions, we can calculate the shortfall-to-quantile ratio as follows:
>
> lim as $\alpha \to 1-$: Shortfall-to-Quantile Ratio at $\alpha =$ lim as $\alpha \to 1-$: ES at α/VaR at $\alpha =$ lim as $\alpha \to 1-$: $-\infty/\infty = 0$
>
> Therefore, the limit of the shortfall-to-quantile ratio as the confidence level alpha approaches 1 from below for a standard normal loss distribution is 0.

★★☆☆☆ In quantitative risk management, one commonly interprets losses as positive quantities (negative profit and losses, if you like). As such, $\mathrm{VaR}(\alpha)$ is $\Phi^{-1}(\alpha)$ for a confidence level $\alpha \in (0,1)$ close to 1 (hence our question concerning the limit for $\alpha \uparrow 1$). However, ChatGPT seems to consider profit and losses (where a negative value represents a loss) and thus focuses on the left tail of the profit-and-loss distribution. In this case, we have $\mathrm{VaR}(\alpha) = -\Phi^{-1}(\alpha)$ and $\mathrm{ES}(\alpha) = \frac{\phi(\Phi^{-1}(\alpha))}{\alpha}$ for small $\alpha \in (0,1)$. One would then need to consider $\alpha \downarrow 0$ instead of $\alpha \uparrow 1$ to obtain a meaningful shortfall-to-quantile ratio. But let us still see if the remaining steps ChatGPT produced are acceptable (so we consider the left tail of the loss distribution as ChatGPT seems to do). We see that the sign of ES is wrong. Furthermore, both computed limits for $\alpha \uparrow 1$ are wrong. The one for ES should be 0, and the one for VaR should be $-\infty$. Note that such problems also occurred in other sessions asking the same question. Overall, the shortfall-to-quantile ratio for $\alpha \uparrow 1$ and left-tailed risk measures should be 0, which ChatGPT obtained but in a questionable way from the expression $-\infty/\infty = 0$.

The problem of using the left versus the right tail for computing VaR and ES alone shows how precise one needs to be when asking technical questions. Not knowing or realizing that ChatGPT works with the wrong tail may lead to entirely different answers, especially if one simply asks about a result without its derivation.

Computing risk measures is only the third of three key tasks of quantitative risk management. The first two are modeling the vector of risk factor changes and computing the corresponding loss distribution function. Even if we know the mapping from the risk factor changes to the losses, computing the loss distribution function is difficult in general. One approach to compute the loss distribution function is historical simulation. Can you explain how this works?

> Historical simulation is a technique used in quantitative risk management to estimate the loss distribution function of a portfolio based on historical data. The basic idea behind historical simulation is to use past observations of the risk factors that affect the portfolio's value to generate a large number of possible scenarios, and then compute the portfolio losses under each scenario using the mapping from risk factors to losses.
>
> Here are the steps involved in historical simulation:
>
> 1. Select a historical time period: The first step is to select a historical time period that is representative of the current market conditions and covers a sufficient number of market scenarios. The length of the time period depends on the frequency of the risk factors and the desired accuracy of the loss distribution function.

> 2. Collect historical data: The next step is to collect historical data for the risk factors that affect the portfolio's value. This data can be obtained from public sources or proprietary data providers.
> 3. Compute the returns: The historical data is used to compute the returns of the risk factors over the time period. The returns can be computed as simple changes in the price, or as log-returns, depending on the nature of the risk factors.
> 4. Simulate scenarios: The returns are used to simulate a large number of possible scenarios for the risk factors. This is done by randomly selecting a sequence of returns from the historical data, with replacement, to form a scenario.
> 5. Compute portfolio losses: For each scenario, the mapping from risk factors to portfolio losses is applied to compute the portfolio loss under that scenario. The mapping can be a simple linear function, or a more complex function that takes into account the correlations and nonlinearities in the portfolio.
> 6. Build the loss distribution function: The portfolio losses from all the scenarios are collected and used to build the empirical loss distribution function. This is done by sorting the losses in ascending order and computing the empirical cumulative distribution function (CDF) at each loss level.
> 7. Compute risk measures: Once the loss distribution function is estimated, it can be used to compute various risk measures, such as Value at Risk (VaR), Expected Shortfall (ES), and Tail Value at Risk (TVaR).
>
> Historical simulation is a useful technique for estimating the loss distribution function of a portfolio when analytical or parametric methods are not feasible or appropriate. It allows for the incorporation of realistic market scenarios and the capture of complex nonlinearities and correlations in the portfolio. However, historical simulation has some limitations, such as the assumption that the future will be similar to the past and the potential for overfitting the historical data. Therefore, it should be used in conjunction with other risk management techniques and subject to ongoing validation and monitoring.

★★★★⯪ This is an almost perfect non-technical answer. One could stress more that the involved mapping (from the past risk-factor changes to the corresponding losses) is based on the *current* portfolio composition, so the *current* mapping is fed with *past* risk-factor change data. Also, in Step 4, one would normally not resample the past risk-factor changes but use all of them to compute the empirical distribution function of the implied losses and estimate the risk measures of interest from it. And in Step 7, the mentioning of TVaR in addition to ES is redundant (the classical estimators do not differ).

3.4. Time Series

Past observations in quantitative risk management are typically available as time series and are thus not iid. Their modeling adds another layer of complexity when computing, for example, time-dependent risk measures.

Define the notion of a time series.

> A time series is a sequence of observations or measurements of a variable or set of variables over time. In a time series, each observation is associated with a specific time or time interval, which may be regular or irregular. Time series data can be collected from a wide range of sources, such as economic indicators, financial markets, weather stations, and sensors.
>
> Time series data are used in many fields, such as economics, finance, engineering, and environmental sciences, to study the behavior and patterns of the variables over time. Some common applications of time series analysis include forecasting

> future values, detecting trends and cycles, identifying seasonal patterns, and modeling the relationships between variables.
>
> Time series analysis involves a variety of techniques, such as statistical models, signal processing methods, and machine learning algorithms, to extract useful information from the data and make predictions or decisions based on the results. The choice of method depends on the specific characteristics of the time series, such as its stationarity, autocorrelation, and seasonality, as well as the goals of the analysis.

★ ★ ★ ⯪ ☆ This answer is acceptable. Although we asked about a time series (an integer-indexed stochastic process) and not about time series data or time series analysis, the stochastic notion and its realization are often both referred to as time series. We also learn about areas of application of time series modeling.

One of several stylized facts about univariate time series of financial data is volatility clustering. What is volatility clustering?

> Volatility clustering is a stylized fact about univariate time series of financial data that refers to the tendency of the volatility of the series to be clustered in time. In other words, periods of high volatility tend to be followed by periods of high volatility, and periods of low volatility tend to be followed by periods of low volatility.
>
> This phenomenon is observed in many financial time series, such as stock prices, exchange rates, and commodity prices. Volatility clustering has important implications for risk management and forecasting, as it suggests that the future volatility of a financial asset is dependent on its recent history and cannot be predicted solely based on long-term trends or fundamental factors.
>
> There are several possible explanations for volatility clustering. One is that market participants are subject to herding behavior, which leads to increased volatility when many traders act in the same direction. Another is that news and information about the underlying economic or political factors affecting the asset are released in clusters, leading to bursts of volatility. Additionally, some models of financial markets, such as the GARCH (Generalized Autoregressive Conditional Heteroscedasticity) model, incorporate volatility clustering as a fundamental feature of the dynamics of asset prices.
>
> In practice, volatility clustering can be measured using various statistical methods, such as the autocorrelation function, the variance ratio test, and the Heterogeneous Autoregressive (HAR) model. These methods can be used to estimate the persistence and magnitude of volatility clustering in a time series, and to incorporate this information into risk management and forecasting models.

★ ★ ★ ★ ⯪ This answer is good. Perhaps the addition of what volatility itself is would have been good (but then one could also ask about this separately). Concerning the last paragraph, one would need to add that the autocorrelation function based on the return series under consideration will not reveal volatility clustering; rather the autocorrelation function of the absolute (or squared) value(s) of the return series does so.

Can you provide the definition of a GARCH(p,q) model?

> Yes, a GARCH(p, q) model is a type of time series model used to capture the conditional variance and volatility clustering in a financial time series. The acronym GARCH stands for Generalized Autoregressive Conditional Heteroscedasticity.
>
> The GARCH(p, q) model has the following structure:
>
> $$V_t = \alpha_0 + \alpha_1 \cdot \varepsilon_{t-1}^2 + \alpha_2 \cdot \varepsilon_{t-2}^2 + \cdots + \alpha_p \cdot \varepsilon_{t-p}^2 + \beta_1 \cdot V_{t-1} + \beta_2 \cdot V_{t-2} + \cdots + \beta_q \cdot V_{t-q}$$
>
> where V_t is the conditional variance of the time series at time t, ε_t is the standardized residual (i.e., the difference between the observed value and the expected

> value divided by the conditional standard deviation), and α_i and β_i are the parameters of the model to be estimated.
>
> The first term on the right-hand side (α_0) is a constant term that represents the unconditional variance of the time series. The second term represents the autoregressive part of the model, which captures the dependence of the current variance on the past squared residuals ($\varepsilon_{t-1}^2, \varepsilon_{t-2}^2, \ldots, \varepsilon_{t-p}^2$). The autoregressive part allows for the modeling of volatility clustering, as periods of high volatility are likely to be followed by periods of high volatility. The third term represents the moving average part of the model, which captures the dependence of the current variance on the past values of the conditional variance $(V_{t-1}, V_{t-2}, \ldots, V_{t-q})$. The moving average part allows for the modeling of persistence in the volatility of the time series.
>
> The parameters of the GARCH(p,q) model can be estimated using maximum likelihood estimation, and the fitted model can be used for forecasting, risk management, and other applications in financial time series analysis. The GARCH model is a popular and widely used model for modeling volatility in financial time series.

★★★★☆ Although there are good aspects covered in this answer, there are also subtle errors. A model having “the following structure” does not provide an unambiguous definition. Indeed, in terms of the conditional variance V_t as provided in the displayed equation, a GARCH(p,q) model is $\varepsilon_t = \sqrt{V_t} Z_t$, where $(Z_t)_{t\in\mathbb{Z}}$ is a *strict white noise process* (a series of iid random variables with mean 0 and variance 1). A more commonly used and more meaningful notation is given in (McNeil et al. 2015, Definition 4.22) (for $\mu_t = 0$). So the displayed equation only specifies the conditional variance of the GARCH model, not the GARCH model itself. Also, ε_t is in general not called “standardized residual”, nor even “residual”, as, at the level of the model, there are no residuals yet (as there are no data yet). Instead, the standardized Z_t are called the *innovations*, and their distribution is the *innovation distribution*. After the parameters of the model are estimated, the resulting (non-standardized) $\hat{\varepsilon}_t$ are the *residuals* and the resulting $\hat{Z}_t$ are the *standardized residuals*. Interestingly, one would think of ChatGPT potentially making an error by calling the part $\alpha_1\varepsilon_{t-1}^2 + \alpha_2\varepsilon_{t-2}^2 + \cdots + \alpha_p\varepsilon_{t-p}^2$ of V_t the autoregressive part and calling $\beta_1 V_{t-1} + \beta_2 V_{t-2} + \cdots + \beta_q V_{t-q}$ the moving average part of the model. However, in terms of the model being $\varepsilon_t = \sqrt{V_t} Z_t$, ChatGPT answered correctly (see also ARCH models), even though somewhat confusingly from the perspective of the form of V_t.

Suppose I have a realization of a time series of stock prices. How would I fit a GARCH(p,q) model?

> To fit a GARCH(p,q) model to a time series of stock prices, you would typically follow these steps:
>
> 1. Preprocess the data: This involves cleaning the data, checking for missing values, and converting the data to a stationary series if necessary.
> 2. Estimate the mean: The GARCH(p,q) model assumes that the mean of the series is constant over time, so the first step is to estimate the mean of the series using a suitable method such as maximum likelihood or ordinary least squares.
> 3. Estimate the conditional variance: The next step is to estimate the conditional variance of the series using the GARCH(p,q) model. This involves estimating the parameters α_i and β_i using a suitable method such as maximum likelihood or generalized method of moments.
> 4. Check for model adequacy: Once you have estimated the GARCH(p,q) model, you should check for model adequacy by examining the residuals of the model. If the residuals are white noise and do not exhibit any significant patterns or autocorrelation, then the model can be considered adequate.

> 5. Forecast future values: Finally, you can use the fitted GARCH(p, q) model to forecast future values of the series, including the conditional mean and variance.
>
> Note that the process of fitting a GARCH(p, q) model can be complex and may require some knowledge of statistical modeling and financial data analysis. It is also important to note that GARCH(p, q) models are just one type of time series model that can be used to analyze and forecast financial data, and that other models such as ARMA, ARIMA, and ARCH may also be suitable depending on the properties of the data.

★★★★⯨ Step 1 captures an important aspect, namely that time series models need to be fitted to stationary data, hence the modeling is typically performed based on (negative) log returns rather than the original time series of stock prices. Step 3 hints at the same problem already mentioned, ChatGPT seems to believe that only the conditional variance term is the GARCH(p, q) model. The remaining parts of the answer are fine (even though an ARMA model being adequate for (negative) log-returns of financial time series is rather unlikely due to the lack of volatility clustering).

Can you define the notation of covariance stationarity and provide conditions on the parameters of the GARCH(1,1) model for it to have a covariance stationary solution?

> Covariance stationarity refers to a stationary time series where the mean, variance, and covariance do not change over time. In other words, the statistical properties of the time series remain constant over time.
>
> For a GARCH(1,1) model to have a covariance stationary solution, the following conditions must be satisfied:
>
> 1. The ARCH and GARCH parameters must be non-negative. This ensures that the variance of the time series is always positive.
> 2. The sum of the ARCH and GARCH parameters must be less than 1. This condition ensures that the time series has a finite variance.
> 3. The mean of the time series must be constant over time. This can be achieved by including a constant term in the GARCH model.
>
> If these conditions are met, then the GARCH(1,1) model will have a covariance stationary solution.

★★★★⯨ Overall a good answer but with one inaccuracy. Covariance stationarity does not require the (auto)covariance to remain unchanged over time; it only requires a constant (auto)covariance among all values of the time series, a fixed number of time steps (the *lag*) apart.

3.5. Extremes

We are often concerned with the largest losses, and extreme value theory provides a theoretical framework for their modeling.

In extreme value theory, what does it mean to say that a distribution function F belongs to the maximum domain of attraction (MDA) of a distribution function H?

> In extreme value theory, it is said that a distribution function F belongs to the maximum domain of attraction (MDA) of a distribution function H if the maximum of a sequence of independent and identically distributed random variables with distribution function F can be approximated by the maximum of a sequence of independent and identically distributed random variables with distribution function H. This means that as the sample size increases, the distribution of the maximum value converges to the distribution function H. The MDA is the largest class of distribution functions for which this convergence occurs.

★★☆☆☆ This answer is unnecessarily complicated and also partially wrong. Let $X_1, \dots, X_n \overset{\text{ind.}}{\sim} F$. $F \in \mathrm{MDA}(H)$ means that there are sequences $(c_n) > 0$ and (d_n) such that $(\max\{X_1, \dots, X_n\} - d_n)/c_n$ converges in distribution to H. In particular, $\max\{X_1, \dots, X_n\}$ itself does not have to (and in general does not) converge to H in distribution; it almost surely converges to the right endpoint $x_F = F^{-1}(1)$ of F.

Explain what the Pickands–Balkema–de Haan Theorem says.

> The Pickands-Balkema-de Haan theorem is a fundamental result in extreme value theory that describes the behavior of the tail of a distribution. Specifically, the theorem states that if a distribution function F belongs to the maximum domain of attraction of a distribution function H, then for any positive threshold u, there exist constants $a(u) > 0$ and $b(u)$ such that:
>
> $$\mathbb{P}((X - b(u))/a(u) > x \mid X > b(u)) \to 1 - H(x)$$
>
> as $u \to \infty$, where X is a random variable with distribution function F, and $\to$ denotes convergence in distribution. In other words, the theorem states that as the threshold u becomes large, the distribution of the excesses over this threshold (i.e., the values of X that exceed u) converges to the generalized Pareto distribution with shape parameter $\xi = 1$, which is the tail distribution of H. This result is important because it allows us to estimate the tail distribution of a distribution F based on the excesses over a high threshold u, which is often more reliable than estimating the entire distribution.

★★★☆☆ There is some good content in this answer but also quite a few inaccuracies, complications, and errors again. Let H_ξ denote the generalized extreme value distribution with shape parameter ξ. In a digestible form, the Pickands–Balkema–de Haan theorem states that $F \in \mathrm{MDA}(H_\xi)$, $\xi \in \mathbb{R}$, if and only if there exists a positive, measurable function $\beta(u)$ such that

$$\lim_{u \uparrow x_F} \sup_{0 \le x < x_F - u} |F_u(x) - G_{\xi,\beta(u)}(x)| = 0, \tag{1}$$

where $F_u(x) = \mathbb{P}(X - u \le x \mid X > u)$ is the excess distribution function of X over the threshold u, and $G_{\xi,\beta(u)}$ is the generalized Pareto distribution with parameters $\xi \in \mathbb{R}$ and $\beta(u)$. ChatGPT focuses on the necessity part of the statement, which is the important one for applications. It expresses the statement in terms of the survival functions (fine) but misses a couple of other things. That the limit is a generalized Pareto is only mentioned in the explanation after the statement (still fine), but the convergence in the displayed equation is ordinary convergence (as the appearing terms are probabilities or distribution functions rather than random variables). Most importantly, the convergence is uniform. That $\xi = 1$ is of course an error. Also, the limiting H is the GPD, not just "the tail distribution of H".

An important part (somewhat hidden) of ChatGPT's answer is something students often struggle with when learning the Pickands–Balkema–de Haan theorem (or the related statistical peaks-over-threshold method that exploits this theorem). It is not the conditional *exceedance probability* $\mathbb{P}(X \le x \mid X > u)$ that converges to a GPD but the *excess probability* $\mathbb{P}(X - u \le x \mid X > u)$; the random variable $X \mid X > u$ is the *exceedance*, and $X - u \mid X > u$ is the *excess*. Also, the last sentence is in principle fine but misses the main point, namely that the GPD is often a more reliable tail estimator than other *tail* estimators. In short, ChatGPT's answer contains good aspects but fails to clearly mention some important ones.

Suppose we have an unknown loss distribution function F for which we want to estimate expected shortfall based on n observations. How does knowing that F belongs to the MDA of the generalized extreme value distribution H with shape parameter xi help modeling large losses from F?

> Knowing that F belongs to the maximum domain of attraction (MDA) of the generalized extreme value distribution (GEV) H with shape parameter ξ helps modeling large losses from F by allowing us to estimate the expected shortfall (ES)

> of F beyond a high threshold u based on the generalized Pareto distribution (GPD) fitted to the excesses over u. Specifically, the Pickands–Balkema–de Haan (PBH) theorem tells us that as u becomes large, the distribution of the excesses over u converges to the GPD with shape parameter $\xi = 1$, which is the tail distribution of H. This means that we can estimate the ES of F beyond u by fitting the GPD to the excesses over u and using its mean or quantile as an estimate of the ES.
>
> In practice, this means that if we have n observations from F, we can first estimate the threshold u by selecting a value that corresponds to a specified tail probability, such as 0.01 or 0.05. We can then compute the excesses over u and fit the GPD to these excesses using maximum likelihood estimation or another method. Finally, we can estimate the ES of F beyond u by using the mean or quantile of the fitted GPD. This approach is particularly useful for modeling extreme losses, which are often difficult to estimate accurately using traditional statistical methods.

★★★☆☆ The first sentence is spot on. The second is also good, apart from the two errors in "with shape parameter $\xi = 1$, which is the tail distribution of H". Similarly for the third sentence, the beginning is good, but then suggesting to use the mean or quantile of the fitted GPD as an estimator of ES is wrong or at least lacking accuracy; compare (McNeil et al. 2015, Equation (5.9)) with (McNeil et al. 2015, Equation (5.19)). The same is repeated in the second paragraph. The paragraph is overall well written, but the "0.01 or 0.05" in choosing a threshold u "by selecting a value that corresponds to a specified tail probability, such as 0.01 or 0.05" makes little sense. Perhaps what ChatGPT means is using the estimated 99% or 95% quantile of the given sample of size n as the threshold, but then it seems to confuse the tails again by working with the left tail of the underlying loss distribution (even though ChatGPT stated the Pickands–Balkema–de Haan theorem in terms of the right tail!).

For financial losses from F, one often finds that the estimated shape parameter xi of the generalized extreme value distribution is positive. How can one characterize all such F?

> All distributions F for which the estimated shape parameter ξ of the generalized extreme value distribution is positive can be characterized as having heavy tails and a higher probability of extreme events. This means that these distributions have a greater likelihood of producing large or extreme losses, which can be a significant risk for financial institutions and investors. Additionally, these distributions are often used to model rare events, such as natural disasters or financial crises, which can have a significant impact on financial markets and the economy as a whole. Therefore, understanding the characteristics of such distributions is important for risk management and decision-making in finance.

★★★★☆ We are, of course, looking for the property of regular variation. Let us try to see if ChatGPT finds it.

I forgot the name of the property of all F which are in the maximum domain of attraction of the generalized extreme value distribution for positive xi. Can you help me with that?

> The property of all F which are in the maximum domain of attraction of the generalized extreme value distribution for positive ξ is called heavy-tailedness. This means that the tails of the distribution decay more slowly than those of a normal distribution, indicating a higher probability of extreme events. The heavy-tailedness property is important in finance because it implies that extreme losses are more likely to occur than they would be under a normal distribution. Therefore, it is important for risk management and decision-making in finance to account for heavy-tailed distributions when assessing and managing risk.

★★★★☆

Heavy-tailedness was only a consequence of this probabilistic property. There was also a theorem stating that all F with this tail behavior are in the maximum domain of attraction of the generalized extreme value distribution for positive xi.

> Yes, you are correct. The property of heavy-tailedness is a consequence of the probabilistic property that all F with this tail behavior are in the maximum do-

> main of attraction of the generalized extreme value distribution for positive ξ. This is known as the Fisher–Tippett–Gnedenko theorem, which states that any distribution with a heavy tail will converge to one of three possible extreme value distributions, depending on the exact shape of the tail. These extreme value distributions are the Gumbel distribution, the Fréchet distribution, and the Weibull distribution, which correspond to negative, positive, and zero values of the shape parameter ξ, respectively. Therefore, heavy-tailedness is a key characteristic of distributions in the maximum domain of attraction of the generalized extreme value distribution, and it has important implications for risk management and decision-making in finance.

★★★☆☆ ChatGPT leads us back to our assumption of such F being in the $\mathrm{MDA}(H_\xi)$ for $\xi > 0$.

I was looking for the property of regular variation of F.

> I apologize for the confusion. The property you are referring to is regular variation of F. Regular variation is a mathematical property that characterizes the tail behavior of a distribution. Specifically, a distribution F is said to be regularly varying if its tail behaves like a power law, meaning that the tail probability decays at a rate proportional to a power of the tail index. In other words, if $F(x)$ is regularly varying with tail index α, then for any $t > 0$,
>
> $$\lim_{x\to\infty} \frac{F(tx)}{F(x)} = t^{-\alpha}.$$
>
> Regular variation is closely related to the Fisher–Tippett–Gnedenko theorem and the maximum domain of attraction of the generalized extreme value distribution. In fact, one can show that a distribution F is in the maximum domain of attraction of the generalized extreme value distribution with positive shape parameter ξ if and only if F is regularly varying with tail index $\alpha = 1/\xi$. Therefore, regular variation is a key property for understanding the tail behavior of distributions and their convergence to extreme value distributions.

★★☆☆☆ Although expressed well overall, this answer is (still) not correct, as F in the displayed equation should be $\bar{F} = 1 - F$, a major mistake.

3.6. Dependence

For modeling jointly large losses with tailor-made multivariate distribution functions, copula modeling has been successfully applied in quantitative risk management practice.

Suppose we model a random vector X of risk factor changes with a joint t distribution, but some of the margins are not univariate t distributed. How can we improve the joint t distribution to obtain a more realistic model for X?

> One approach to improving the joint t distribution of a random vector $\boldsymbol{X}$ with non-univariate t distributed margins is to use a copula function to model the dependence structure among the margins separately from their individual distributions. Copulas are mathematical functions that link the marginal distributions of a multivariate distribution to its joint distribution. They allow us to model the dependence structure of $\boldsymbol{X}$ without making assumptions about the individual distributions of each component.
>
> To use a copula, we first transform the marginals of $\boldsymbol{X}$ to their standard uniform distribution using their cumulative distribution functions (CDFs). Then, we model the dependence structure among the transformed marginals using a copula function. Finally, we apply the inverse CDFs of the original marginals to the copula-generated values to obtain a joint distribution that has more realistic marginal distributions and dependence structure.
>
> There are many types of copula functions available, each with different properties and strengths. Some popular copula functions include the Gaussian copula, t copula,

> Clayton copula, and Gumbel copula. The choice of copula function will depend on the specific characteristics of the data and the goals of the modeling exercise.

★ ★ ★ ⯪ ☆ From the “non-univariate *t* distributed margins”, we see the predictive nature of ChatGPT, of course the right expression would have been “non-*t* distributed univariate margins”. Also the logic is a bit off in this paragraph. It is not the underlying *t* copula that we want to change, it is a margin. Nevertheless, the first paragraph provides a fairly good answer. There is also some good content in the second paragraph but again some inaccuracies. First, “their standard uniform distribution” would better be “the standard uniform distribution”, especially due to the second “their” in this sentence. Second, the last sentence “we apply the inverse CDFs of the original marginals to the copula-generated values” also has its flaws. We would not use the original margins but those that fit better; as before, we see that ChatGPT answers the question of how we can improve the dependence among the components of $\boldsymbol{X}$ rather than how to improve the fit of the non-*t* margins. And the inverse CDFs of (any) margins are not applied to “copula-generated values” but rather realizations of the copula (at least using the expression “copula-generated values” is ambiguous in this regard). As such, we obtain a sample (as opposed to “a joint distribution”) from a more realistic distribution for $\boldsymbol{X}$ than the multivariate *t*. The third paragraph is fine.

If we only change the marginal distributions of a random vector X, why does this not affect its copula?

> The copula of a random vector is a function that only depends on the joint distribution of the vector, not on the marginal distributions of its components. This means that if we change the marginal distributions of the vector, the copula remains the same.
>
> To understand why this is the case, consider the definition of a copula. A copula is a function that describes the dependence structure between random variables. It is defined as the joint distribution of the transformed variables, where the marginal distributions of the transformed variables are uniform.
>
> Since the copula is defined in terms of the joint distribution of the random vector, it only depends on the relationship between the variables, not on their individual distributions. Changing the marginal distributions of the vector does not affect the relationship between the variables, so the copula remains the same.
>
> In other words, the copula captures the dependence structure between the variables, while the marginal distributions capture their individual behavior. Changing the marginal distributions does not change the dependence structure, so it does not affect the copula.

★ ★ ★ ☆ ☆

I was looking for the invariance principle. Do you know what the invariance principle says?

> Yes, the invariance principle is a fundamental concept in probability theory that states that certain properties of probability distributions are invariant under certain transformations.
>
> In particular, the invariance principle states that if we apply a measurable transformation to a random variable, the resulting probability distribution will have the same properties as the original distribution, as long as the transformation is measurable.
>
> For example, if we apply a monotonic transformation to a random variable, such as taking the logarithm or the square root, the resulting distribution will have the same shape and properties as the original distribution, such as its moments, skewness, and kurtosis.
>
> The invariance principle is closely related to the concept of a copula, which is a function that captures the dependence structure between random variables, regardless of their individual marginal distributions. The copula is invariant

> under monotonic transformations of the variables, which means that changing the marginal distributions of a random vector does not affect its copula.
>
> Overall, the invariance principle is a powerful tool for analyzing and modeling probability distributions, as it allows us to focus on the essential properties of the distribution, rather than its specific form or representation.

★★★☆☆ The first paragraph is fine, albeit a bit vague. In the context of copulas, the second paragraph is wrong. One would need the transformations not only to be measurable but also strictly increasing on the ranges of the underlying random variables. The third paragraph starts by mentioning monotonicity (but again not that the transformations must be increasing); however, from the perspective of a single random variable (the case in this part of the answer), ChatGPT is wrong since moments can and typically do change when transforming random variables. The fourth paragraph is promising, as it mentions what we are looking for, but it again omits the increasingness.

The invariance principle explains why the underlying copula remains unchanged when we estimate the marginal distributions of a random vector X by their empirical distribution functions based on iid observations from X. After componentwise transforming the observations from X with their empirical marginal distribution functions, we obtain the so-called pseudo-observations of the observations from X. From a theoretical point of view, pseudo-observations share a drawback in comparison to observations from X or actual observations from the underlying copula. What is this drawback?

> The drawback of using pseudo-observations is that they are not independent, unlike observations from X or actual observations from the underlying copula. This lack of independence can affect the accuracy and reliability of statistical inference and modeling based on pseudo-observations.

★★★★★ Finally, a spot-on answer on a more technical subject!

Does the pointwise largest possible copula (the upper Frechet–Hoeffding bound) lead to the largest correlation between the two components of a bivariate random vector X irrespective of the two marginal distributions of X?

> Yes, the upper Fréchet–Hoeffding bound is the copula that maximizes the correlation between the two components of a bivariate random vector $\boldsymbol{X}$, regardless of the two marginal distributions of $\boldsymbol{X}$. This means that the upper Fréchet–Hoeffding bound is the strongest possible dependence structure that can exist between the two components of $\boldsymbol{X}$, and any other copula that is consistent with the given marginal distributions of $\boldsymbol{X}$ will have a lower correlation between the components.

★★★★⯪ Omitting the unnecessary "that is consistent with" makes this a good answer, too.

Do you know how this result is shown mathematically?

> Yes, the result that the upper Fréchet–Hoeffding bound maximizes the correlation between the components of a bivariate random vector X can be shown mathematically as follows:
>
> Let $\boldsymbol{X} = (X_1, X_2)$ be a bivariate random vector with marginal cumulative distribution functions F_1 and F_2, respectively. Let C be a copula that is consistent with the marginal distributions of $\boldsymbol{X}$, i.e., $C(F_1(x_1), F_2(x_2)) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2)$ for all x_1, x_2.
>
> The correlation between X_1 and X_2 can be expressed in terms of the copula C as:
>
> $$\rho(X_1, X_2) = \frac{2}{\pi} \sin(\pi C(u, v))$$
>
> where $u = F_1(X_1)$ and $v = F_2(X_2)$ are the transformed uniform margins under C.

★☆☆☆☆ There was more output, but we omitted it here (only) for this particular question, as what we listed above already shows a major flaw in the argument, namely that the right-hand side of the displayed equation depends on the random variables u, v (unfortunately given in terms of lowercase letters instead of uppercase, which should be avoided at all costs). Asked why this formula holds, ChatGPT argued with Sklar's theorem and then magically used it from one equality sign to the next. The correct argument can be found in (McNeil et al. 2015, Theorem 7.28).

I would like to sample from a bivariate distribution with log-normal LN(0,1) and LN(0,16) margins, and correlation 0.5. How can I do that?

> To sample from a bivariate distribution with log-normal $\mathrm{LN}(0,1)$ and $\mathrm{LN}(0,16)$ margins, and correlation 0.5, you can use the following steps:
>
> 1. Generate two independent standard normal random variables Z_1 and Z_2.
> 2. Transform Z_1 and Z_2 to log-normal random variables with mean 0 and standard deviation 1 and 4, respectively, using the inverse of the log-normal cumulative distribution function. That is, set:
>
> $$X_1 = \exp(\mu_1 + \sigma_1 \cdot Z_1)$$
> $$X_2 = \exp(\mu_2 + \sigma_2 \cdot Z_2)$$
>
> where $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 0$, $\sigma_2 = 4$, and Z_1 and Z_2 are the standard normal random variables generated in step 1.
> 3. Generate a copula with a correlation of 0.5. One way to do this is to use the Gaussian copula with a correlation matrix given by:
>
> $$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$
>
> This correlation matrix can be obtained by transforming the correlation of 0.5 using the inverse of the Fisher transformation.
> 4. Generate a random sample of size n from the copula using the inverse transform method. That is, for each $i = 1, 2, \ldots, n$, generate two uniform random variables $U_{1,i}$ and $U_{2,i}$ using the inverse of the Gaussian copula. This can be done using the bivariate standard normal cumulative distribution function and the Cholesky decomposition of the correlation matrix. That is, set:
>
> $$\mathbf{Z} = (Z_1, Z_2)$$
> $$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 4 \end{bmatrix}$$
> $$\Delta = \mathrm{chol}(\Sigma)$$
> $$U = \Phi(\Delta^{-1}\mathbf{Z})$$
>
> where Φ is the bivariate standard normal cumulative distribution function, chol is the Cholesky decomposition function, and Δ^{-1} is the inverse of the Cholesky factorization of Σ.
> 5. Transform each pair of uniform random variables $(U_{1,i}, U_{2,i})$ to log-normal random variables $(X_{1,i}, X_{2,i})$ using the inverse of the log-normal cumulative distribution function. That is, set:
>
> $$X_{1,i} = \exp(\mu_1 + \sigma_1 \cdot \Phi^{-1}(U_{1,i}))$$
> $$X_{2,i} = \exp(\mu_2 + \sigma_2 \cdot \Phi^{-1}(U_{2,i}))$$
>
> where Φ^{-1} is the inverse of the standard normal cumulative distribution function.

6. The resulting pairs $(X_{1,i}, X_{2,i})$ for $i = 1, 2, \dots, n$ are a random sample from the bivariate distribution with log-normal $\mathrm{LN}(0,1)$ and $\mathrm{LN}(0,16)$ margins, and correlation 0.5.

Note that step 4 and 5 can also be done using other copula models and transformation methods, depending on the specific requirements of the application.

★★☆☆☆ This is a classical example from the realm of dependence modeling in quantitative risk management. The short answer is, such a model does not exist, so there can also not be a valid sampling procedure. In Step 1, ChatGPT advises to "Generate [...] random variables", the first inaccuracy (one can only generate realizations, not random variables). In Step 2, a (linguistic) error is to speak of the random variables X_1, X_2 having mean 0 and standard deviations 1 and 4, respectively (the mean and standard deviation of $\mathrm{LN}(\mu, \sigma^2)$ are $\exp(\mu + \sigma^2/2)$ and $\sqrt{(e^{\sigma^2} - 1)\exp(2\mu + \sigma^2)}$, respectively). In Step 3, ChatGPT first goes wrong by constructing a Gaussian copula with a parameter matrix being a correlation matrix with off-diagonal entry 0.5; how ChatGPT wants to apply the Fisher transformation remains its secret. However, once this Gaussian copula is combined with the log-normal margins, the correlation will change (to a value fairly close to 0) since Pearson's correlation coefficient is affected by the marginal distributions. In Step 4, ChatGPT then describes the construction in terms of samples. With the "inverse of the Gaussian copula", it most likely means the inverse Rosenblatt transform, which could indeed be used but, similar to before, this is not the most straightforward way. It then mixes the notions of generated samples, bivariate distribution functions and Cholesky decompositions. What it actually suggests to construct at the end of Step 4 is strange. The appearing bivariate random vector $\Delta^{-1}\boldsymbol{Z}$ is bivariate normal with covariance matrix Σ (note the non-unit variances on the diagonal yet the off-diagonal 0.5). It suggests to apply the bivariate $\mathrm{N}(\boldsymbol{0}, I_2)$ distribution function, where I_2 denotes the identity matrix in $\mathbb{R}^{2\times 2}$. This would leave us with only a one-dimensional sample. Even if ChatGPT meant to apply the univariate $\mathrm{N}(0,1)$ distribution function componentwise, we would not obtain a copula sample, as the second margin of $\Delta^{-1}\boldsymbol{Z}$ is not $\mathrm{N}(0,1)$. And again note the switch back from the focus on samples to the focus on random variables. Step 5 now again makes the switch to samples but correctly describes the transformation to log-normal margins (just the input samples $(U_{1,i}, U_{2,i})$ are still the wrong ones from Step 4).

4. Summary

Overall, the questions we asked ChatGPT are standard in undergraduate or graduate courses on quantitative risk management. Table 1 provides our average rating for ChatGPT's answers on the different subjects we covered (with the number of questions asked in parentheses). One must point out that the ratings should not be overinterpreted, as they are solely based on our own judgment of the correctness of the answers provided.

Table 1. Average 5-star rating (number of questions asked in parentheses) for each subject.

Reproducibility	Risk	Risk Measures	Time Series	Extremes	Dependence
3.6 (5)	4.72 (9)	2.72 (9)	4.2 (5)	3 (7)	3.14 (7)

More meaningful would be to present ChatGPT's answers to a larger number of experts for their assessment, perhaps even with different rubrics, for example, to determine how valuable a response is to a first learner of the subject, an experienced user, an industry practitioner, or an instructor of a course on the topic.

5. Conclusions

Based on how ChatGPT answered our questions, we can draw the following conclusions:

1. *The mathematical one.* Asking technical questions clearly shows the limitations of ChatGPT 3.5. The fact that such a predictive model has no built-in logical structure as rigorous as expected in mathematics implies that one effectively must rely on chance to obtain a fully correct answer; see also Young (2023). One therefore has to constantly question every single word or formula obtained from ChatGPT (this can indeed be used in a meaningful way, see below!). This especially applies to definitions, where oftentimes rather ambiguous explanations are provided instead, at least as first parts of ChatGPT's responses. Even if ChatGPT corrects itself in a follow-up clarification, apologizes and sounds fully confident about its corrections of a mistake, the answer can still be wrong (or, as we saw, even remain unchanged). So without being an expert on the subject matter, it is risky, if not dangerous, to "trust" ChatGPT's answers, so fact checking is necessary. From a practical point of view, this is concerning and will inevitably lead to wrong decisions being made when people (at any rank) blindly rely on ChatGPT's (and potentially many other AIs) responses. The phenomenon of the seemingly convincing but entirely wrong answers of large language models such as ChatGPT is also known as "hallucination"; see Crouse (2023) or Rodger (2023). As such, a self-assessment of ChatGPT's confidence in a particular response or related references (as provided by Google's Bard) would be useful. We had a slightly better experience when asking about the differences between similar or related concepts (for example different types of risks or risk measures). This may be one way of (slightly) improving ChatGPT's quality of answers.
2. *The dangerous one.* One can never be sure to have provided ChatGPT with enough information to uniquely solve a problem or obtain a sufficiently accurate answer. We saw this when ChatGPT worked with the (left tail of the) profit-and-loss distribution instead of the (right tail of the) loss distribution. These are things not easily spotted, especially if one only asks for a final result and not a derivation, or if one does not carefully check the latter. Our aforementioned concern on practical implications applies here, too.
3. *The frustrating one.* When one tries to correct an error that ChatGPT makes, it sometimes repeats the same wrong argument. This is not surprising, as ChatGPT is not updated on the fly. It is nonetheless frustrating, especially when it is an (to the human eye) obvious error, such as the $X in the first question on risk measures. In these moments, one realizes that one is only interacting with an artificial intelligence and not a real one (the Turing test comes to mind). Careful prompt engineering may lead to a correct answer after some trials. However, if one has to ask multiple times, hinting at the underlying concept in different ways just to get a satisfactory answer, then this already requires a good understanding of the concept and thus defeats the purpose of using ChatGPT for obtaining information about said concept in the first place. Alternatively, one could provide additional texts to ChatGPT and ask it to specifically answer based on the provided texts, for example, articles one suspects to be of high quality or to contain expert knowledge on the subject. However, this again fails for mathematical equations and notation. Given the progress of the quality of chatbots in the past year, one can be optimistic that some of these (more obvious) issues will be solved in future versions, at which point it would be interesting to assess ChatGPT's performance on our set of questions again.

   Also annoying was that even for fairly well-known mathematical facts (for example, those available from Wikipedia pages), the arguments that ChatGPT provided were too complicated, or used rather non-standard notation, which can make it unnecessarily hard to learn a new concept. A good notation can go a long way towards grasping a concept as every experienced teacher knows. In that sense, we do not share some news articles' practical concern of ChatGPT replacing teachers any time soon (see also Point 6 below).
4. *The unexpected one.* The non-reproducibility of answers is more of an issue than one might think. A correctly answered question can turn into a wrong one when asked

again or, more often, when asked in a slightly different manner, and the other way around. For casual conversations (including those with humans, where dialogue partners may also be wrong sometimes), this may not matter (depending on the context). But for the purpose of obtaining reliable information, this is a severe drawback. It was also criticized during peer review of this article (as reviewers also asked their ChatGPT versions to answer similar questions, with partly different but overall similar results to ours), and rightfully so (hence, our first section on non-reproducibility). In fact, from an academic point of view and in light of the scientific method, non-reproducibility is detrimental. This is reminiscent of training neural networks with current software, where setting seeds is neither commonly applied nor always possible. This means that one cannot suggest, say, to students or practitioners (interested in learning about a concept) to use ChatGPT for certain pre-verified subjects or types of questions (or seeds!). Note that the temperature parameter is only partially useful for alleviating this problem, as it does not solve the problem of reproducibility when ChatGPT is "constantly being updated" as ChatGPT itself wrote in its first answer. Nevertheless, we saw similar types of answers across all questions we asked, and overall similar observations were made by reviewers (even though we do not know the versions of ChatGPT they used). Our current best advice is to give ChatGPT sufficient context (which means providing ChatGPT with more details than you would think are necessary) to be able to answer a question appropriately and to ask in different ways to see the bigger picture, but this requires a good understanding of the subject; see our Point 3. One observation is that ChatGPT answers more reliably when asked to compare two similar concepts that are otherwise more easily confused; see also Point 1 above.

5. *The promising one.* We found ChatGPT especially helpful for summarizing topics covered by a large amount of literature such as the financial crisis of 2007 to 2009, so tasks more related to qualitative than quantitative risk management. It would take much longer to skim the world wide web for a good summary on the subject. This should not come as a surprise due to the massive amount of training data available on the subject. ChatGPT is typically also helpful for finding the pros and cons of well-known modeling approaches and thus helps to reduce one's own potential bias. An advantage is that one can easily play with the length of output that ChatGPT produces and can thus quickly obtain an overview depending on how much reading time one wants to spend.
6. *The educational one.* As just mentioned, use ChatGPT to summarize well-covered, non-mathematical topics. Using it to learn about a specific mathematical concept (and thus asking for definitions, characterizations, properties, etc.) is not advisable at this stage. However, as hinted at before, ChatGPT could well be integrated into learning about risk in the curriculum at a later stage. Once a topic was covered in a course or prepared for an exam, instructors could ask students to prepare questions for ChatGPT, and then instructors and students could jointly analyze ChatGPT's responses, try to find its hallucinations and thus solidify their own understanding of the topic and enhance critical thinking. Also letting ChatGPT generate pros and cons essays and then analyzing how convincing both types of arguments are is an idea. As such, ChatGPT could replace a classical study buddy who you would meet after self-studying for an exam to ask possible exam questions and then give feedback to (and vice versa).
7. *The practical one.* ChatGPT may also be helpful in identifying the "unknown unknowns", the risks we do not know that we do not know. This is because ChatGPT may have been trained on sources covering risks we did not face before and cannot imagine facing. A similar idea led to the introduction of operational risk loss databases after operational risk became a risk class required to be modeled by financial companies with its introduction in Basel II. Companies paid to obtain access to anonymized loss data from a pool of (other) companies in order to use those losses for

in-house stress testing. This usage of ChatGPT is much cheaper and is not limited to bigger companies but also available to smaller ones or even individuals. For example, start-ups not able to afford consulting may rather ask ChatGPT about various risks they may face. Or, individuals who do not trust their local insurance agent to provide them with an unbiased view on various insurance products could ask ChatGPT for recommendations about which are (or are not) useful from an individual's perspective. Answers may not always be reliable, but whatever is pushed onto your radar that was not there before is good from a (holistic) risk management perspective.

8. *The inspiring one.* ChatGPT can be used for motivation and as a catalyst when informally learning about risk. It can also serve as inspiration for creating the title of a talk or a conference. Here is an example:

   Write an eight line poem about quantitative risk management that encourages students to learn more about this subject.

   Quantitative risk management, oh so wise
   It's a subject you should not despise
   Learn about numbers, probabilities, and stats
   And how they help to mitigate mishaps

   Risk is present in all that we do
   But with QRM, we can see it through
   So dive in deep, don't be afraid
   And let your knowledge on this subject be displayed!

Funding: This research received no external funding.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: I'd like to thank The University of Hong Kong for providing access to ChatGPT while this paper was written.

Conflicts of Interest: The author declares no conflict of interest.

References

BIS. 2006. Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework—Comprehensive version. Available online: bis.org/publ/bcbs128.pdf (accessed on 5 August 2023).

Crouse, Megan. 2023. ChatGPT Cheat Sheet: Complete Guide for 2023. TechRepublic. Available online: techrepublic.com/article/chatgpt-cheat-sheet (accessed on 5 August 2023).

Gomez, Francesca. 2023. AI as a Risk Officer: How ChatGPT Fared with a Risk Assessment. *LinkedIn*, May. Available online: linkedin.com/pulse/ai-risk-officer-how-chatgpt-fared-assessment-francesca-gomez (accessed on 2 September 2023).

Hofert, Marius. 2023. Correlation pitfalls with ChatGPT: Would you fall for them? *Risks* 11: 115. [CrossRef]

Hofert, Marius, Rudiger Frey, and Alexander J. McNeil. 2020. The Quantitative Risk Management Exercise Book Solution Manual. Available online: https://github.com/qrmtutorial/qrm/releases/tag/TQRMEB (accessed on 26 March 2023).

Jorion, Philippe. 2006. *Value at Risk: The New Benchmark for Managing Financial Risk*, 3rd ed. New York: McGraw-Hill.

Kashettar, Swathi. 2023. Challenges and Opportunities of ChatGPT for Risk Management. *Analytics Insight*, June 22. Available online: analyticsinsight.net/challenges-and-opportunities-of-chatgpt-for-risk-management (accessed on 2 September 2023).

Kloman, H. Felix. 1990. Risk management agonists. In *Risk Analysis*. Hoboken: John Wiley & Sons, vol. 10, pp. 201–5.

Kosowski, Robert, and Salih N. Neftci. 2014. *Principles of Financial Engineering*, 3rd ed. Cambridge: Academic Press.

McNeil, Alexander J., Rüdiger Frey, and Paul Embrechts. 2015. *Quantitative Risk Management: Concepts, Techniques and Tools*, 2nd ed. Princeton: Princeton University Press.

OpenAI Community. 2023. Mastering Temperature and top_p in ChatGPT api. Available online: https://community.openai.com/t/cheat-sheet-mastering-temperature-and-top-p-in-chatgpt-api-a-few-tips-and-tricks-on-controlling-the-creativity-deterministic-output-of-prompt-responses/172683 (accessed on 2 September 2023).

Rodger, James. 2023. Google Issues Urgent Warning to Anybody Using ChatGPT. Birmingham Live. Available online: birminghammail.co.uk/news/midlands-news/google-issues-urgent-warning-anybody-26227871 (accessed on 5 August 2023).

Sidorenko, Alex. 2023. Best Risk Management Prompts for ChatGPT. Available online: https://riskacademy.blog/download/best-risk-management-prompts-for-chatgpt/ (accessed on 2 September 2023).

Vincent, James. 2023. 'As an AI Language Model': The Phrase That Shows How AI Is Polluting the Web. The Verge. Available online: theverge.com/2023/4/25/23697218/ai-generated-spam-fake-user-reviews-as-an-ai-language-model (accessed on 7 May 2023).

Young, Scott H. 2023. 10 Ways You Can Use ChatGPT to Learn Better. Available online: scotthyoung.com/blog/2023/05/02/chatgpt-learning-tips (accessed on 9 May 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.