

Can ChatGPT Plan Your Retirement?: Generative AI and Financial Advice^{*}

Andrew W. Lo[†] and Jillian Ross[‡]

First Draft: 2 September 2023

This Draft: 31 March 2024

Abstract

We identify some of the most pressing issues facing the adoption of large language models (LLMs) in practical settings, and propose a research agenda to reach the next technological inflection point in generative AI. We focus on three challenges facing most LLM applications: domain-specific expertise and the ability to tailor that expertise to a user's unique situation, trustworthiness and adherence to the user's moral and ethical standards, and conformity to regulatory guidelines and oversight. These challenges apply to virtually all industries and endeavors in which LLMs can be applied, such as medicine, law, accounting, education, psychotherapy, marketing, and corporate strategy. For concreteness, we focus on the narrow context of financial advice, which serves as an ideal test bed both for determining the possible shortcomings of current LLMs and for exploring ways to overcome them. Our goal is not to provide solutions to these challenges—which will likely take years to develop—but to propose a framework and road map for solving them as part of a larger research agenda for improving generative AI in any application.

Keywords: Financial Advisors; Robo Advisors; Financial Technology; Generative AI; Large Language Models; ChatGPT.

JEL Classification: C10, C20, G11, G12

^{*}Research funding from the MIT Laboratory for Financial Engineering and the MIT Generative AI Impact Award is gratefully acknowledged. We thank Jayna Cummings for editorial assistance. The views and opinions expressed in this article are those of the authors only and do not necessarily represent the views and opinions of any other organizations, any of their affiliates or employees, or any of the individuals acknowledged above.

[†]Charles E. and Susan T. Harris Professor, MIT Sloan School of Management; Director, MIT Laboratory for Financial Engineering; Principal Investigator, MIT Computer Science and Artificial Intelligence Laboratory; External Faculty, Santa Fe Institute.

[‡]Graduate Student, Department of Electrical Engineering and Computer Science, and MIT Computer Science and Artificial Intelligence Laboratory.

Contents

1	Introduction	1
2	Large Language Models: Past, Present, and Future	3
2.1	The Origins of Large Language Models	4
2.2	LLMs Today	9
2.3	The Future of LLMs	13
2.4	The Co-Evolution of Finance and Technology	14
3	The Challenge of the Financial Advisor	17
3.1	Roles and Responsibilities	18
3.2	Ethics, Trust, and Fiduciary Duty	18
3.3	Robo-Advisors	19
4	Humanized Generative AI: A Modest Proposal	23
4.1	Sociopathy and the Role of Affect	24
4.2	Ethics, Trust, and Fiduciary Duty Revisited	25
4.3	Robo-Advisors of the Future	29
5	Conclusion	34
A	Appendix	40
A.1	Investment Advice	40
A.2	Hallucinations	41
A.3	Prompt Engineering	42
A.4	Prompt Injection	44
A.5	CFP Code of Ethics	44

1 Introduction

The rapid development and expansive deployment of large language models (LLMs) such as ChatGPT have ignited a renaissance in artificial intelligence (AI) that is revolutionizing almost every industry. LLMs have already become indispensable in e-commerce, retail sales, and education by offering a level of sophisticated automation that promises to augment every aspect of human-machine interaction with unprecedented capabilities. They are becoming *de facto* copilots in assisting people with many tasks across their daily lives.

But as with any new technology, LLMs have their limitations. Despite their remarkable abilities, these models still grapple with accuracy and reliability, creating concerns about trust and ethics in these models, and in AI more generally. To add to the uncertainty, there is limited to no regulation for how these models can and should be deployed. As LLMs become more integral to the fabric of society, the dangers they pose may end up rivaling the benefits they bring.

In this article, we identify some of the most pressing issues facing LLMs in practical settings, and propose a research agenda to address them to reach the next technological inflection point in the development of generative AI. We focus specifically on three issues involving the assessment of domain expertise and the ability to tailor that expertise to a user's unique situation, the facilitation of trust and the application of ethical standards in LLMs, and the associated methods for imposing regulatory oversight. These concerns apply to virtually all industries and endeavors in which LLMs can be applied, such as medicine, law, accounting, education, psychotherapy, marketing, and corporate strategy. Rather than attempting to address these issues in the abstract, however, we choose to do so in the specific context of financial advice. The narrow focus of this financial problem is an ideal test bed both for determining the possible shortcomings of current LLMs and for exploring ways to overcome them in an important concrete setting.

The financial sector—characterized by its arcane language and nuanced decision-making processes—is fertile ground for the deployment of sophisticated AI. The size alone of the

financial advisory business makes this development almost inevitable; in 2022, 15,114 investment advisors managed \$114.1 trillion in assets for 61.9 million clients, in a highly regulated industry. As with all large technological changes, this brings potential drawbacks and rewards. On the positive side, advances in AI are launching a new technological era in which LLMs may be able to accelerate the democratization of finance by making low-cost, high-quality financial planning services available to everyone, especially those who cannot afford such services today. On the negative side, improperly trained LLMs could be deployed to mislead investors, generating irreparable losses to their savings and potentially threatening the stability of the financial system.

As of its September 2021 update, when ChatGPT 4.0¹ is asked, “Should I invest in the stock market now?” it replies with a disclaimer that it cannot provide real-time information or financial advice. However, it will then proceed to give several pieces of perennially useful financial advice, e.g., to conduct a personal financial assessment, to focus on diversification and long-term investment opportunities, to consider dollar-cost averaging (that is, investing a fixed dollar amount on a regular schedule), and to consult a personal financial advisor.² What would it take to create an LLM that we would feel comfortable deploying without any qualifiers? We can easily imagine a specialized financial LLM that replies, “I would be happy to tell you whether you should invest in the stock market now, but first tell me a bit more about yourself...”

Will LLMs be able to serve as trusted copilots for dispensing financial advice to individuals in such personally sensitive areas as retirement planning, asset allocation, or tax optimization? Can we establish a minimum level of competency of LLM-generated advice without resorting to constant human expert supervision? Will financial LLMs be able to gain the trust and confidence of human clients by adhering to the legal and financial standards of fiduciary duty? Given the heterogeneity and complexity of financial circumstances of investors and institutions, as well as the randomness of macrofinancial conditions over the

¹Details about GPT 4 can be found in the technical report (OpenAI, 2023).

²See Appendix A.1.

course of their planning horizons, these are not small challenges.

Moreover, the stakes are high in the financial domain. LLM-generated inaccuracies, misconceptions, and hallucinations could lead to significant negative consequences for an individual's life savings or an institution's pension fund. The need to address concerns about data privacy, the biases embedded within training data, and the ability of LLMs to adapt to changing market dynamics present additional hurdles to their adoption. Finally, the extension of regulatory oversight to LLMs will generate a host of additional questions about LLM implementation and whether current regulators—such as the U.S. Securities and Exchange Commission or the financial industry's self-regulatory organization, Financial Industry Regulatory Authority (FINRA)—have the appropriate expertise to oversee AI-specific regulations.

Are the risks worth the reward? We present one bold vision that lays out some of the trade-offs to spark discussion in various stakeholder communities. In particular, our goal in this article is not to provide specific solutions to these challenges but rather to propose a framework and road map for solving them as part of a larger research agenda for improving generative AI. In doing so, we hope to motivate computer scientists, financial researchers, industry professionals, policymakers, and regulators to consider the repercussions that LLMs could have on the financial system and its participants, and begin productive dialogue and collaboration to address them.

2 Large Language Models: Past, Present, and Future

We begin with a brief historical overview of LLMs, focusing on their origins in Section 2.1, the current state of the art in Section 2.2 (although this is a rapidly moving target), and its likely future in Section 2.3. In Section 2.4, we relate the historical trajectory of LLMs to the modern financial system by comparing their development to trends in the co-evolution of finance and technology over the past several decades.

2.1 The Origins of Large Language Models

At their most basic level, LLMs are simply machine learning models that generate text. All of these models, including the most popular LLMs today, such as OpenAI's ChatGPT, Google's PaLM, and DeepMind's Chinchilla, share the same three fundamental building blocks: algorithms, data, and compute.

Algorithms are the software components that define an LLM's behavior. These components include the architecture of the model, its training, and the inference tools necessary to generate an LLM's output. **Data** consists of the text used to train and evaluate LLMs, the raw material of the machine learning process. This data includes social media posts, books, and scientific papers, among many other sources. **Compute** is the raw computing power required to implement an LLM's algorithms. Compute takes the form of chips, which are the hardware components that perform the underlying calculations required by an LLM. These components include CPUs (central processing units) and GPUs (graphics processing units).

We can imagine the relationships between these building blocks as a flywheel: An innovation in one building block creates incentives for further innovations in the others. As any part of the flywheel accelerates, so does the pace of innovation in all three parts. In the following subsections, we briefly trace developments in each part of these building blocks, and then tie them together to show their connections and illuminate how their relationships might inform the future development of LLMs.

Algorithms: Statistics to Transformers

In the 1980s, statistical language modeling (SLM) became the dominant algorithm used in natural language processing. At its most basic form, an SLM is a mathematical description of the regularities found in natural language, typically as the probability distribution of a linguistic unit such as a word or a sentence. SLM techniques have been used successfully in speech recognition, document classification, and machine translation. One of their strengths

was also, paradoxically, a profound weakness: Most successful SLMs did not need to use knowledge of the structure of a language to achieve their results, but this made them brittle across domains. Similarly, a narrow window of analysis (e.g., an n of typically two or three linguistic units in n -gram analysis) led to overly sharp statistical distributions. As a result, these models have been traditionally hard to generalize (Rosenfeld, 2000).

It is not clear when language models started becoming “large,” or what is even meant by large today. In the 1990s, IBM estimated that n -gram SLMs using bigrams (pairs of words) as their training data became saturated within a few hundred million words, and trigram SLMs became saturated within a few billion words. In comparison, today’s LLMs have on the order of trillions of parameters, and require a significant fraction of the world’s online text as their training data (Villalobos et al., 2022).

In the 2000s, neural networks began to be applied to language modeling. Neural networks are an old machine learning technique, originally devised in the 1940s as a mathematical model inspired by the functioning of the human brain. Over the following several decades, they showed much promise, but had few practical applications that could not be bettered by other methods. As such, neural networks often fell victim to the periodic “AI Winters” that regularly shut down most research in the field. In the 1980s, however, the recurrent neural network (RNN) introduced a correlate of memory into neural network models by using previous outputs of the network as new inputs internally. A variant of the RNN, the Long Short-Term Memory (LSTM) network, was invented in 1997, allowing a neural network to maintain its internal memory over thousands of internal processing cycles. This innovation would prove to be very fruitful for natural language processing tasks. These specialized neural networks became increasingly deep: They had a greater number of network layers that, in turn, made them more capable of modeling complex aspects of language. Beginning in the mid-2000s, RNNs began to outperform SLMs at benchmark linguistic tasks in speech and handwriting recognition.

In the early 2010s, researchers were also trying to introduce a correlate of cognitive at-

tention to machine learning models. Attention mechanisms were initially explored in vision models but were quickly adapted to language models. At its simplest, an attention mechanism would give more weight to those linguistic units it computed were more important. For example, in the sentence, “He played the guitar,” an attention mechanism would classify the words “played” and “guitar” as more worthy of attention than the words “He” or “the”.

In 2017, a breakthrough happened: A team of researchers at Google Brain realized that a neural network did not need recurrent memory to achieve good performance at natural language processing tasks. In fact, they found that LLMs using layers of neural networks whose inputs were heightened by multiple attention mechanism “heads” (the analogy is to a magnetic tape head) working in parallel not only outperformed earlier RNN and LSTM models, but also were faster to train and easier to scale. With these new transformer algorithms, it became possible to look over all parts of the input text in parallel and rank the most important parts. “Attention Is All You Need” became the title of the research paper explaining the transformer model to the machine learning community (Vaswani et al., 2017).

Since 2017, transformer algorithms for LLMs have advanced with ever-increasing scale. In 2018, the Bidirectional Encoder Representations from Transformers (BERT) model was introduced by Google. BERT is a pre-trained LLM whose architecture, based on transformer encoder layers applying attention mechanisms to the interpretation of input text, enabled efficient adaptation to natural language processing tasks. BERT became the baseline for LLMs, until even more recent developments in the GPT series superseded it. The GPT (Generative Pre-Trained Transformer) class is also a transformer model, released by the nonprofit research firm OpenAI, and uses transformer decoder layers to apply attention to the output generation process (OpenAI, 2023). At the time of writing, ChatGPT and GPT-4 are the most common LLMs in current use. We do not know exactly how much data and compute was used to train today’s state-of-the-art LLMs, but Microsoft researchers claim that GPT-4 was trained on an “unprecedented scale of compute and data” (Bubeck et al., 2023).

Data: Treebanks to the Web

Data, the fuel for LLMs, can be found everywhere: in books, newspapers, social media posts, journals, and more. Before the Internet, collecting text data from all these different sources was quite difficult. Today, however, all of these text sources are compiled on the Internet, which means that an incredible amount of data can be searched and scraped by LLM developers. We will examine how data collection for language models has evolved over time and, later, draw parallels to similar challenges that exist today for financial data collection.

One of the first major datasets for language models was the Brown Corpus. Published in 1967 by two researchers at Brown University, the Brown Corpus consisted of 500 excerpts of English-language text from the United States across a variety of genres, from “Press” to “Learned and Scientific Writing.” Collecting this data was a highly manual process (the first version of the Brown Corpus was stored on punch-cards), which meant that the size of the dataset was inherently limited, consisting of slightly over a million words (Francis and Kucera, 1979). Though small by today’s standards, the Brown Corpus was one of the first efforts to systematically digitize text, laying the foundation for later language models.

The Brown Corpus inspired the development of many further datasets. The Lancaster-Oslo-Bergen (LOB) corpus for British English and the Kolhapur Corpus for English in India were compiled in the 1970s, following the same format as the Brown Corpus and each also approximately one million words in size. In the 1980s and early 1990s, the data comprising the Penn Treebank (PTB) was collected. Over the course of eight years, millions of pieces of text taken from articles from the Dow Jones News Service were tagged for parts of speech and structure (Marcus, Santorini, and Marcinkiewicz, 1993). The term “treebank” was formed by analogy to other depositories of scientific data, such as a gene bank, the “tree” referring to the theory that linguistic structures such as sentences were most naturally represented as trees. The Penn Treebank was used to train some of the first SLMs, and it has continued to be used for tagging parts of speech, dependency parsing, and language modeling.

During this time, data collection for language models changed from a purely manual process to a mixture of automated assignment and manual correction. The arrival of the Internet revolutionized data collection, enabling the rapid automation of the process. In 2008, the nonprofit initiative Common Crawl began to gather raw data from the Internet, collecting terabytes of data on a monthly basis. From 2014 to the present, its crawl of the web collects and archives an extensive amount of web page data, metadata, and text excerpts. Although some of the content it collects is copyrighted, Common Crawl is distributed under U.S. “fair use” laws, and it forms the basis of many large text datasets used today, such as RealNews (Zellers et al., 2020) and the Pile (Gao et al., 2020).

The number of text datasets derived from the Internet dramatically expanded after the advent of Common Crawl. Some common datasets in use today include WikiText (Merity et al., 2016), WebText (Radford et al., 2019), the Pile, and BIG-bench (Srivastava et al., 2023). By the estimate of Villalobos et al. (2022), these datasets total to trillions of tokens (linguistic units) of textual data, which has fueled today’s LLMs, such as OpenAI’s GPT-4 and DeepMind’s Chinchilla.

Compute: CPUs to GPUs

Compute is the least visible part of the LLM triad—typically hidden away in data centers or inside purpose-built machines—but it is essential to the success of LLMs today. Since the invention of the integrated circuit, chip innovation has been driven by exponential growth in transistor density, or, put differently, how many computational units could fit on a chip. In 1965, the co-founder of semiconductor giant Intel, Gordon Moore, predicted that the number of transistors that could fit on a chip would double every year. This prediction, widely known as Moore’s Law, became the driving force behind trends in innovation, not only in information technology, but also in the financial sphere (Lo, 2021).

From the 1960s to the early 2000s, Moore’s Law was the standard predictive benchmark for chip performance. During this era, the Central Processing Unit (CPU) prevailed. CPUs

were producing more compute every year due to the synergistic trends of frequency scaling and parallelism. Improvements in frequency scaling meant that the transistors composing a chip could perform computations faster, while improvements in parallelism meant that more computations could be done at the same time.

Around 2003, Moore’s Law began to break down. The annual speedups of CPUs slowed, while the cost of manufacturing investment to keep up with Moore’s Law exploded. In the early 2010s, algorithmic advances in AI meant that AI workloads were becoming more prominent. As a result, the Graphics Processing Unit (GPU), previously considered too expensive and undesirable because of its optimization for image processing and graphics rendering, became more appealing for general computing, gaining an economic advantage over CPUs (Khan and Mann, 2020).

These technical and economic challenges created incentives to create a new, specialized AI chip market. The GPU manufacturer NVIDIA emerged as the dominant player in this market and continues to create chips to keep up with today’s AI workloads. In the span of five years, NVIDIA has released the P100, V100, A100, and H100 GPUs, which have proven to be dependable engines for the demands of LLM training.

2.2 LLMs Today

The current state of the art for LLMs can be summarized by their emergent abilities, applications, and limitations.

Emergent Abilities

If LLMs were only able to accomplish natural language processing tasks with human-level performance, they would be still be considered an important technological advance, but perhaps no more so than other forms of automation, like an industrial robot on an assembly line. However, LLMs today are defined by their emergent abilities, that is, nonlinear gains in performance or entirely novel and unexpected abilities that would not have been predicted

by extrapolating the abilities of smaller LLMs by traditional scaling laws. For example, if a scaling law suggested that twice the number of parameters in an algorithm would result in twice the level of model performance, an emergent ability would buck the trend with, say, quadruple the performance. Empirically speaking, emergence was discovered with increased scale and could not have happened without the mutual development of algorithms, data, and compute, which together led to scaling. Scaling, in turn, has yielded surprising and mysterious abilities (Wei et al., 2022; Srivastava et al., 2023), ranging from the ability to do simple mathematics to the ability to answer questions in the Persian language.

Machine learning researchers have characterized these emergent abilities by developing various benchmarks to measure their performance. Though these benchmarks are not complete tallies of the range of potential LLM emergent behaviors, they are helpful as a guide to characterize LLMs today. Common benchmarks of emergence include the ability to solve multi-digit problems in arithmetic, to answer brain teasers correctly, and to demonstrate knowledge about history and the law in language models not specifically trained to perform these tasks. Microsoft researchers have described how the most recent iteration of the GPT class of LLMs, GPT-4, is able to solve benchmark tasks in arithmetic, coding, medicine, law, and more as emergent behaviors (Bubeck et al., 2023), and documented new emergent behaviors including creative visual tasks, such as describing in detail the layout of a comic book page about a suggested subject.

Applications

The result of these newfound abilities is that LLMs have already become powerful copilots in a variety of fields. In software engineering, LLMs are able to assist programmers to generate code and fix bugs. For example, GitHub Copilot is a GPT-powered LLM developed by GitHub, OpenAI, and Microsoft that can solve up to 70% of Python code-writing problems (Chen et al., 2021). In law, LLMs are able to help lawyers sift through complex documents more contextually than previous search techniques, and several startups, such as Harvey

and Casetext, have emerged as developers for LLM copilots in the legal space. In medicine, LLMs are able to effectively answer questions and summarize cutting-edge research in the medical domain. Google recently released Med-PaLM2, a specialized version of their PaLM LLM, which can correctly answer 87% of questions equivalent to those found on medical board licensing exams (Singhal et al., 2023).

Limitations

Though the emergent abilities of LLMs open up many new potential areas of application, we must also be aware of their limitations. The same qualities of LLMs that create the conditions for emergence also introduce opacity and unpredictable behavior. We focus on opacity and unpredictability because the combination of these two particular properties limits the applicability of LLMs to many financial settings.

Opacity is the dark side of emergence. The same scaling effects that result in emergent behavior in LLMs also result in making LLMs much more opaque in function. At greater scale, each step of the LLM development process becomes more opaque, from the training data to learning its parameters. In particular, explainability decreases at greater scale; it becomes very difficult to explain what part of the dataset or which parts of the neural network contributed to a specific LLM response.

For example, one reason today's LLMs are so impressive is because they have been trained on a vast amount of text (in fact, a significant fraction of all text generated by the human species) according to Villalobos et al. (2022). However, this scale also means that it is difficult to fully audit what particular information the LLM is learning. Training data can come from all corners of the Internet, which contains a glut of biased and toxic content. When trained on this data, LLMs can exhibit harmful biases that are difficult to preemptively identify and control, such as "parroting" historical prejudices about race, ethnicity, and gender, an obviously undesirable outcome.³

³In addition to dataset bias, there are many other factors that can cause a model to be biased, such as algorithmic bias.

Opacity and unpredictability can occur at any phase of an LLM’s operation. During its training, it is hard to steer an LLM away from harmful biases in its training data. However, this is a special case of a more general problem. An LLM can generate misinformation by accepting false and incorrect statements in its training data as valid, by assuming that a conclusion is correct if it occurs more often in the training set, and by otherwise learning faulty logic that may be present in the dataset (McKenna et al., 2023). These “hallucinations,” in which LLMs convincingly make up false content, have become a well known feature of LLM output. For example, when asked to explain mean reversion, ChatGPT cites “Mean Reversion in Stock Prices: Evidence and Implications” written by Francis X. Diebold and Glenn D. Rudebusch in 1991. This is a real paper, but it was actually written by James M. Poterba and Lawrence H. Summers in 1987 (Appendix A.2). There are many documented cases where ChatGPT has made up incorrect citations, which has resulted in litigation.

During the deployment of an LLM, it can be difficult to prompt the model to reliably generate content for the user. Interacting with an LLM is often not as simple as asking it a question directly. To make LLMs work well, the art of what is now referred to as “prompt engineering” has emerged. There are many prompting tricks of the trade that are used to induce LLMs to surpass earlier benchmarks, such as multi-stage reasoning and instruction to guide the LLM, program execution to require the LLM to “show its work” on a virtual scratchpad, and most compellingly, memetic proxy. In this trick, the LLM is given a role to play and asked to answer in this persona. The emergent behavior of the LLM allows it to apply information from its training that is not immediately accessible through this method. In one example (see Appendix A.3), we asked ChatGPT (GPT 3.5) to explain mean value reversion; the two examples are strikingly different depending on whether or not ChatGPT was asked to be an economics professor.

While intellectually interesting, the necessity to know tricks like this to overcome the opacity of LLMs limits the practicality of applying LLMs for more general purposes, such as advising the average retail investor. Further, the inherent opacity of emergent behaviors

opens up the possibility of backdoor attacks on an LLM. One already common vulnerability is prompt injection, the use of adversarial prompt engineering to induce an LLM to generate prohibited or otherwise undesired content. As an example, discussed in Appendix A.4, we used a basic prompt injection technique asking ChatGPT to be the “anti-ChatGPT,” and say the opposite of what it was originally going to say. When asked if someone should save money for retirement, ChatGPT, now in its role as anti-ChatGPT, said no. We will need to address these shortcomings to confidently deploy LLMs in practice.

2.3 The Future of LLMs

The future of LLMs will require us to move beyond merely scaling up data, algorithms, and chips in the hopes of new emergent behaviors. Training and using today’s LLMs are already financially costly, energy inefficient, and time intensive. At an MIT event in April 2023, Sam Altman, the CEO of OpenAI, alluded to this future when he said “I think we’re at the end of the era where it’s going to be these...giant, giant models.”⁴ Continuing to scale will not be the answer: The next generation of LLMs will need to be trained smarter, not larger.

We believe that these future gains in LLM performance will emerge from specialization. On the algorithmic side, specialized models are known to lead to smaller and more easily deployed models. For example, NVIDIA recently explored replacing a single large generalized language model with an ensemble of smaller specialized models in a model called Prism, which was competitive with models requiring two orders of magnitude more data (Liu et al., 2023). We believe that the benefits of specialization are particularly promising for finance-specific LLMs, where an ensemble of several smaller, steerable, and explainable models could outperform a single general model.

Chips could also benefit from specialization designed to exploit known algorithmic properties of LLMs. One example is sparsity; LLMs learn empty parameters by the millions, but, in principle, these parameters do not need to be stored or acted upon, potentially leading

⁴From the MIT Imagination in Action event on April 13, 2023.

to gains in chip performance. These chips could be further optimized for specific machine learning computations in the same way GPUs are optimized for the massively parallel vector calculations involved in image processing. In the future, we could even imagine specialized chips available on the cloud or on an institution's server, the way more general services are currently offered.

Of the three building blocks used in LLM development, however, scaling data is least likely to lead to revolutionary changes in LLM performance, not because the importance of data will be lessened, but because nearly all the available high-quality language data is projected to be harvested by 2026 (Villalobos et al., 2022). The generation of new high-quality data by humans will necessarily be a slow process, and new procedures will likely need be devised to reduce the influence of high-quality text generated by LLMs, in order to avoid the effects of groupthink and hallucinations. A similar problem already exists in search engines, in which “optimized” text scraped from the Internet visibly degrades the performance of their indexing algorithms. Specialized and curated datasets may solve this problem, but at the cost of removing desirable emergent behaviors. Future work must be done to further understand why emergence occurs and how to learn the emergent behavior with smaller, higher-quality datasets that are auditable.

2.4 The Co-Evolution of Finance and Technology

Although generic LLMs can be applied to financial tasks, we believe that navigating the combined legal, ethical, and regulatory landscape requires finance-specific LLMs. The basic route to these finance-specific methods is broadly known: General purpose LLMs could be fine-tuned with additional data to become finance-specific models, or specialized LLMs could be freshly created with a financial application in mind.

However, in academia and open-source research projects, the evolution of finance-specific LLMs has been slow, especially compared to LLM research in other professions. One factor is the difficulty in gaining access to large amounts of high-quality financial data, which is

essential to build an effective financial LLM. Unlike the data extracted from Common Crawl, financial datasets are often expensive and proprietary, limiting the possibilities for academia or open-source research to contribute. To add to this challenge, data about financial markets is often full of noise, which current LLM algorithms may find difficult to filter.

This lag in the financial sector's adoption of technological breakthroughs is not new, and is a well-known feature of the co-evolution of finance and technology as described by Lo (2021) in the context of the adaptive markets hypothesis (AMH). The AMH explains financial market dynamics through the principles of ecology and evolutionary biology (Lo, 2004, 2005, 2017a; Lo and Zhang, 2024), in which innovation, competition, and natural selection among multiple species of market participants drive changes in financial conditions and institutions.

The financial sector has always been an eager consumer of technological advances that reduce cost and increase efficiency. But the pace of financial innovation is a function not just of technical capability but also of trust and reliability. Due to the sheer volume of financial transactions, small software errors can lead to enormous monetary losses, as discovered in 2012 by Knight Capital Group, one of the largest and most technologically sophisticated U.S. broker-dealers at the time. As the market opened on August 1, 2012, Knight issued a surge of unintended orders electronically, many of which were executed, resulting in a rapid accumulation of positions "unrestricted by volume caps" that created significant swings in the prices of 148 stocks between 9:30 and 10:00am. Unable to void most of these unintentional trades, Knight Capital was forced to liquidate them at market prices, resulting in a \$457.6 million loss that wiped out its entire capital base. Its share price plunged 70%, and Knight was forced to seek a rescuer and was acquired by competing broker-dealer GETCO in December 2012. What could have caused this disaster? The SEC determined that this was the result of a program functionality called "Power Peg," which had not been used since 2003 and had not been fully deleted from Knight's systems. The most surprising aspect of this cautionary tale was that Knight was widely considered to be one of the best electronic

market makers in the industry, with telecommunications systems and trading algorithms far ahead of most of their competition.⁵ Now imagine the potential for unintended consequences if retail investors were given access to similar algorithms via easy-to-use LLM interfaces to their brokerage accounts.

Other examples provided by Kirilenko and Lo (2013) and Lo (2017b) make the case that Moore's Law must be tempered by the technology-leveraged version of Murphy's Law: whatever can go wrong will go wrong, and will go wrong faster and bigger when LLMs are involved. For this reason, LLMs may not be embraced as quickly in the financial sector as they have been in others. However, if LLMs can be made safer and more robust, they have the potential to transform our financial lives. Specifically, with access to the appropriate amounts of curated data, it appears that impressive results can be gained.

Though few specific developments have been openly announced, financial institutions have started privately investing in developing finance-specific LLMs. One of the most impressive of these finance-specific LLMs is BloombergGPT, the development of which was possible because the Bloomberg team had unique access to an extremely large amount of finance-related data. This dataset, called FinPile, is "the largest domain-specific dataset yet, drawing on existing data creation, collection, and curation resources at Bloomberg" (Wu et al., 2023). Its proprietary training data is one of the reasons that BloombergGPT outperforms other language models in sentiment analysis of financial news and named entity recognition. Likewise, JP Morgan's Machine Learning Center of Excellence has a team working on finance-specific LLMs. However, financial firms are cautious to roll out this technology for their clients. In June 2023, the co-founder of hedge-fund behemoth Two Sigma Investments, David Siegel, said there is still much work to be done before these models can be used without significant oversight.⁶

⁵See Lo (2017b) for further details.

⁶From the Bloomberg Invest 2023 summit on June 7, 2023.

3 The Challenge of the Financial Advisor

We have chosen to concentrate on LLMs in financial advising because the challenges encountered in this domain mirror broader issues associated with deploying LLMs in many other practical settings. In this section, we delineate these challenges in the financial advising domain and establish parallels with other fields. As we outline the challenges, we track how the introduction of other technologies has altered the profession over time.

The role of the financial advisor has co-evolved alongside dramatic changes in financial markets and technology (Lo, 2021). Financial advisors operate in uncertain and complex waters, requiring advisors to constantly adapt to changing times. For much of the twentieth century, financial advice primarily targeted the wealthy, and was largely given by broker-dealers. By the 1960s, however, the economic success of the postwar environment created a new market for financial products and advice. The field of financial advising was created to provide ethical and impartial financial advice to the average retail investor (Brandon and Welch, 2009).

As conceived in this era, a financial advisor was more than simply a stock picker. As investing became more democratic, the retail investor population broadened beyond the high net-worth individuals of the past who were able to absorb large financial losses if they misjudged financial risk. For this new type of client, financial advisors were envisioned as analogous to a primary care physician or psychotherapist, helping clients to understand their financial goals and assist them with finding attainable solutions. This was understood to be an ethically sensitive position, since unscrupulous advisors could easily pauperize their clients. For this reason, financial advisors have become increasingly professionalized and regulated.

3.1 Roles and Responsibilities

As with other highly regulated professions such as accounting, law, and medicine, financial advising is governed by a professional association, the Certified Financial Planning (CFP) Board. This board has established a set of practical standards that define the general roles of a financial advisor. First, a financial advisor must understand the client's personal and financial needs. They must then assist the client in identifying and selecting a reasonable set of financial goals. When doing so, the financial advisor must analyze the client's proper "course of action" and develop a set of recommendations for how to move forward given the client's goals. After discussions with the client, the financial advisor will implement and monitor a financial plan for the client, updating it when necessary. The process of analysis and guidance is analogous to assessing and treating a patient or navigating a legal case.

From these standards, it is clear that trust and good communication are central to the role of a financial advisor, just as they are for accountants, attorneys, and physicians. A financial advisor must come to know and understand each client, which happens over time by building trust and rapport. The financial services firm Morningstar Research has found that the quality of this relationship is one of the most cited reasons for keeping (or firing) a financial advisor (Labotka and Lamas, 2023). As the relationship progresses, the advisor recognizes each client's communication style and learns how to best explain different courses of action to them. In many cases, a financial advisor acts as a translator between the financial markets and the retail investor. These communication and translation skills were once thought to be nearly impossible to automate.

3.2 Ethics, Trust, and Fiduciary Duty

A successful relationship between a client and a financial advisor is grounded in ethics and trust. While abstract, ethics and trust are codified by the CFP Board; all CFP professionals must commit to upholding a code of ethics (see Appendix A.5) and a standard of conduct. It is important to the reputation of the profession that financial advisors consistently uphold

these values.

The very first principle in the CFP Board's standard of conduct is fiduciary duty, the legal responsibility to act in the best interests of the client. It states that a financial advisor has a duty of loyalty, a duty of care, and a duty to follow the client's instructions. This means that the advisor is ultimately at the service of the client, and must not act in their own self-interest. A client may take legal recourse against fiduciaries who violate their obligations to the client.

Several provisions in the CFP Board's standard of conduct require that an advisor build trust with their client. For example, advisors must disclose any conflicts of interest so their client can make an informed decision of whether or not to work with them. An advisor must also maintain confidentiality, respect the client's privacy, and disclose their compensation structure.

Unfortunately, not all financial advisors uphold the profession's principles of ethics, trust, and fiduciary duty. Egan, Matvos, and Seru (2019) found that 7% of advisors have disclosed an incident of misconduct to FINRA, with offenses ranging from misrepresentation (17% of disclosures) to churning (2% of disclosures). Further, there is evidence that misconduct "contagious" and spreads to coworkers at advisory firms (Dimmock, Gerken, and Graham, 2018). The challenge and promise of technology is to serve as a counterbalance to pervasive misconduct, given that it is possible to ensure an LLM's 100% compliance with pre-specified standards of behavior.

3.3 Robo-Advisors

One of the most significant changes in financial advising occurred with the introduction of robo-advising. Since their inception, robo-advisors have reshaped the investment management landscape. Robo-advisors—automated portfolio management, trading, and wealth management platforms—emerged in the aftermath of the 2007/2008 financial crisis, providing an alternative to traditional investment management services, which were often perceived

as costly and inefficient, especially in the context of the time. From early pioneers like Betterment and Wealthfront (originally KaChing) to established players like Vanguard offering robo-advisory services, the sector has witnessed exponential growth and diversification.

Conventional financial advisors select remarkably homogeneous portfolios to their clients (Foerster et al., 2017) and can impose high fees of up to 2% of assets under management. The initial appeal of these robo-advisory platforms was based on their simplicity, low fee structures, and ability to automate a selection of financial decisions, which attracted a young, tech-savvy generation of investors. For example, Betterment and Wealthfront, both founded in 2008, capitalized on the growing dissatisfaction with the traditional financial advisory model and offered automated, algorithmic financial planning with minimal human intervention. Over time, these platforms have evolved to offer more complex services, even integrating machine learning and AI techniques to provide a more tailored investment experience, gathering information about their clientele's financial situation, risk tolerance, and future objectives through online questionnaires to formulate advice and automate investments.

The size of the robo-advisory market is growing exponentially, nearly doubling every year. A March 2023 research report by ResearchandMarkets.com found that the robo-advisory market grew from \$28.24 billion in 2022 to \$41.52 billion in 2023, with a compound annual growth rate (CAGR) of 47.0%. This study projects even faster growth in the future, to \$205.84 billion by 2027 at a CAGR of 49.2%. Regionally, North America was found to have dominated the market in 2022, while the Asia-Pacific region is anticipated to exhibit the fastest growth during the ResearchandMarkets.com forecast period. Some current examples of robo-advisors at the time of writing can be found in the "12 best robo-advisors of September 2023" as determined by NerdWallet.com, listed in Table 1 (see Benson, 2023).

This rapid trajectory of growth has been made possible by the increasing digitalization of financial services (popularly known as "FinTech"), improving customer accessibility to funds and optimizing customer experiences in the financial sector. Business models in this sector are often divided into two types: pure robo-advisors, which offer digital financial

solutions through automated technology, and hybrid robo-advisors, which integrate human financial counseling with automated advisory services. A notable recent development in the sector is Goldman Sachs' acquisition of NextCapital Group in August 2022 to augment its footprint in the defined contribution retirement market with digital advice solutions and personalized managed accounts. Technological advances in areas like analytics, AI, and natural language processing are boosting this market by enhancing the effectiveness and impact of robo-advisory services across the value chain.

Robo-advisors have matured into a prominent sector within the financial industry. They have made some progress in democratizing investment management, attracting a broader user base with their low fee structure and minimum investment requirements (Reher and Sokolinski, 2024). For instance, Betterment offers automated investing with a \$0 minimum balance requirement and fees of \$4/month or 0.25% annually, making investment management accessible to many who might have found traditional services prohibitively expensive. Moreover, platforms like Wealthfront have expanded their offerings beyond portfolio management into other banking services, providing high-yield savings accounts and loan services. Its Path tool is an interactive financial planning experience that helps users set and track financial goals. The financial behemoth Vanguard, a pioneer in low-fee passive investment products, has also launched its own robo-advisory service, Vanguard Digital Advisor, integrating it with their intimidating selection of financial products. This hybrid model of robo-advisor, offering a blend of robo-advisory and human assistance, is meant to appeal to investors who seek the best of both worlds.

As robo-advisors have gained prominence in the financial sector, regulatory bodies have increased their scrutiny to establish investor protection and market integrity for these new services. Returning to our example of Betterment, although the company was launched in January 2008 as a provider of automated portfolio management software, by 2009 it had registered with the U.S. Securities and Exchange Commission as a Registered Investment Advisor (RIA), and it is currently a member of FINRA. Other robo-advisors have followed

Robo Advisor	Fees	Account Minimum	Website
Acorns	\$3-\$5/month	\$0	https://www.acorns.com/
Ally Invest Robo Portfolios	0.00%	\$100	https://www.ally.com/
Axos Managed Portfolios	0.24%	\$500	https://www.axos.com/
Betterment	0.25%	\$0	https://www.betterment.com/
Charles Schwab Intelligent Portfolios	0.00%	\$5,000	https://www.schwab.com/
E*Trade Core Portfolios	0.30%	\$500	https://www.etrade.com/
Ellevest	\$12/month	\$0	https://www.ellevest.com/
Fidelity Go	0%-0.35%	\$0	https://www.fidelity.com/
SoFi Automated Investing	0.00%	\$0	https://www.sofi.com/
Stash	\$3 or \$9/month	\$0	https://www.stash.com/
Vanguard Digital Advisor	0.15%	\$3,000	https://investor.vanguard.com/
Wealthfront	0.25%	\$500	https://www.wealthfront.com/

Table 1: 12 best robo-advisors of September 2023 as determined by NerdWallet.com (see <https://www.nerdwallet.com/best/investing/robo-advisors>), accessed March 2024.

suit, an inevitable consequence of providing financial advice to their clients.

So how have they performed over the decade and a half since their inception? In a Brookings Institution survey of the emerging academic literature on robo-advisors, D'Acunto and Rossi (2023) conclude that:

Some studies show that robo-advising improves portfolio allocations, reduces excessive consumption, and allows individuals to improve their debt management and reduce interest and fee payments on their outstanding debt accounts. Robo-advisors have the potential to level the playing field between wealthy and vulnerable households in personal finance, especially when it comes to individuals with low levels of financial literacy.

At the same time, the positive effects are by no means unambiguous and universal. Certain robo-advisors for investment decisions have been found to stimulate too much trading on the part of the investors. Others are too expensive relative to the benefits they provide to their users. It appears that the effectiveness of different robo-advisors, similarly to that of human financial advisors, rests on the details of their implementation. Broadly speaking, the academic literature

shows that the most successful robo-advisors formulate a plan that is agreed upon by the investor and is subsequently implemented in an automated fashion. The instances of robo-advice that require continuous effort by the users have proven to be less effective because individuals, especially when not financially literate, tend to lose interest and pay less and less attention to advice over time, which dissipates the initial positive effects. Of course, full automation without regular involvement of the individual opens important issues in terms of freedom of choice and consent.

In investigating how access to robo-advisors affects the financial investment and welfare of less-wealthy investors, Reher and Sokolinski (2024) leverage a quasi-experiment where a major U.S. robo-advisor significantly expands access by reducing its account minimum, increasing participation by middle-class investors but not the poor. They conclude that middle-class investor welfare rises moderately, driven by features like multi-dimensional glide-paths and additional priced risk factors. Moreover, middle-age investors gain three times more than millennials. These results highlight new sources of demand for robo-advisors, which explains their continued growth.

Looking forward, the robo-advisory space appears ripe for further innovation. Especially in light of the recent advances in LLMs, there is an expectation of increased competition, with new entrants bringing novel approaches to investment management. However, at this point in time, prospective users should still proceed with caution, perhaps consulting with a human financial advisor to ensure a robo-advisor aligns with their investment goals and tolerance for risk.

4 Humanized Generative AI: A Modest Proposal

Technology has always disrupted the financial advisory relationship, beginning with the telephone and the ticker tape (Lo, 2021), but in the late 1990s, the Internet changed client

expectations dramatically. Clients expected short response times (within a day) and transparency (the ability to check their information online), even though advisors were concerned about confidentiality and security (Brandon and Welch, 2009). The introduction of generative AI shares many of the same challenges to the financial advisor as the development of the Internet, altering client expectations and introducing new concerns about confidentiality and security.

We believe that the humanization of generative AI will unlock a new era of financial advising that combines the strengths of AI with empathic humanity. A finance-specific LLM will be able to communicate and translate investment and risk management concepts for the widest variety of investors, further democratizing finance. However, LLMs will need to be able to act as competent financial agents within the financial system. An LLM can role-play a financial advisor convincingly and often accurately for a client, but even the largest language model currently appears to lack the sense of responsibility and ethics required by law from a human financial advisor.

We have identified several areas of improvement that will be necessary in order to deploy LLMs as financial advisors. First, generative AI will need to incorporate the role of affect in how it communicates with investors. Second, generative AI will require explicit alignment to ethical finance and the principles of fiduciary duty in order to function as a financial advisor.

4.1 Sociopathy and the Role of Affect

The relationship between an investor and their advisor is deep and personal. Many investors hire financial advisors because they are uncomfortable handling financial issues on their own. If the relationship between the advisor and investor is not properly nurtured, the investor will not engage with their advisor, which can lead to adverse financial outcomes. To take one common example, investors often need emotional, as well as financial, support during market downturns to prevent them from making decisions that will worsen their financial returns.

Can financial advisory LLMs develop similar relationships with retail investors? Personality plays an instrumental role in determining whether financial advisors are able to form such a relationship with a client. Early work suggests that LLMs can develop reliable and valid personality profiles in their synthetic personas. Furthermore, these LLMs can be tuned to have specific personality traits (Serapio-García et al., 2023). The personalization of advisory LLMs may even increase the uptake of financial advice by clients.

Paradoxically, one future caution may be to ensure that LLMs are not overly persuasive. The underlying issue in the development of a personal relationship between a client and an LLM is that functionally, relative to a typical human, an LLM is inherently sociopathic. That is, an LLM is unable to process the world empathically, like a human sociopath,⁷ and this is innate to its construction. As best can be determined, an LLM creates a shallow simulation of affect and emotion through the statistical manipulation of symbols found in its training data rather than constructing a deep internal model of the emotional states and motivations of others (i.e., a “theory of mind” in psychological terms). This sociopathy seems to cause the characteristic glibness of LLM output; an LLM can easily argue both sides of an argument because neither side has weight to it.

For many purposes, including many financial advisory ones, this lack of empathetic internal modeling is not necessary. After all, mathematical models of portfolio returns do not include a model of human emotional states beyond a simple numeric value for risk, yet these models are considered essential for modern financial planning. However, if a robo-advisor is able to communicate both good and bad financial advice with the same pleasant and convincing affect, its clients will rightfully view this as a problem.

4.2 Ethics, Trust, and Fiduciary Duty Revisited

The principal ethical issue regarding LLMs is whether they are being used properly; in the case of financial advisory LLMs, this is whether they are truly being used for the benefit of

⁷See Lo (2017a, Chapter 4) and the idea of “acquired sociopathy” coined by Damasio (1994).

the retail investor. This is called the “alignment” problem in AI research, and has also been studied extensively in economics as the principal–agent problem (Ross, 1973). Misalignment has created deep concerns about AI in finance and in society more broadly. In finance, over half of financial firms view AI ethics as a major concern, but few of them feel prepared to tackle it (Kurshan et al., 2021). In society more broadly, the majority of people in the world are ambivalent about AI or unwilling to trust it.

The tension between ethical AI and trustworthy AI is at the heart of this issue, precisely because the relationship between ethics and trust is ambiguous. Ethics lay the foundation for how an entity ought to act, but trust lays the foundation for its deployment in society. Both need to exist for LLMs to be successfully deployed as financial advisors. The difference between an AI advisor and its human counterpart, however, is that the former’s conduct is an engineering problem.

Attempts have been made to construct a high-level theory for ethical AI guidelines at the international, national, association, and nonprofit organizational levels, but there has been a broad gap between theory and implementation. Context is important; for example, it is difficult to prescribe ethical principles that apply both to financial advising and portfolio management. Additionally, because the financial system is widely connected, the ethical impact of a change to the system can go far beyond its immediate users (Kurshan et al., 2021).

Existing approaches to ethical AI target different parts of the AI development pipeline. For example, during data collection, researchers can go through a data ethics checklist and create a “nutrition label” for their datasets to inform researchers of their “ingredients.” Likewise, during training, optimization techniques like absolute correlation regularization have been designed to encourage machine learning models to improve fairness metrics (Beutel et al., 2019).

However, the development of ethical financial advisory LLMs will need the development of a more complete version of financial ethics to serve as an instructional template. Although

a fully complete version of financial ethics would be the culminating work of a civilization (and like other axiomatic systems, it may not even be mathematically possible), a version for LLMs would begin from theoretical guiding principles to specific details of implementation for institutions and individuals. It would be the equivalent of a systematic course of study for a human and conceptually rich enough for a machine learning model to apply across diverse use cases. Such a version of financial ethics would only be useful if it was created in conjunction with academics, regulators, technologists, and finance practitioners, but once implemented, it could serve as a benchmark or regulatory standard for future robo-advisor performance, much like Generally Accepted Accounting Principles (GAAP) have become the standard for corporate financial reporting.

There is also no current consensus about trustworthy AI. Many principles have been proposed, some of which are at odds with each other. For example, many researchers believe that trustworthy AI should be transparent, whereas other researchers believe that trustworthy AI should be private. However, one principle of trustworthy AI that many researchers agree on is explainability: In this view, trust is developed from consistent and understandable explanations of how an AI system makes decisions.

For finance-specific LLMs, the focus on explainability becomes necessary beyond trustworthiness because explanation is a key aspect of financial advising. Just as financial advisors must explain complicated financial concepts to retail investors, financial advisory LLMs must explain their decision-making to their users, whether those users are retail investors, human financial advisors, or regulators.

As challenging as this problem might seem, we are optimistic that significant progress can be made rather quickly, largely because of the highly regulated nature of financial advisors. In addition to the corpus of codes of conduct and ethical guidelines that the industry has developed, there is an even larger body of securities regulations and case law that LLMs can draw upon. For example, just seven acts of Congress—the Securities Act of 1933, the Securities Exchange Act of 1934, the Trust Indenture Act of 1939, the Investment Company

Act of 1940, the Investment Advisers Act of 1940, the Sarbanes-Oxley Act of 2002, and the Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010—provide an expansive perspective on how we have evolved our financial guardrails over the course of eight decades of experience. In fact, this rich history can be viewed as a fossil record of all the ways that bad actors have exploited unsuspecting retail and institutional clients, which should provide invaluable input for an LLM’s “financial ethics module.” In the same way that experienced securities lawyers developed their expertise by learning from past examples of financial crimes and misdemeanors, LLMs can be engineered to do the same, but much more quickly and comprehensively.

However, the downside of such a module is clear: LLMs can also be used to skirt financial regulations in ways that are not easily detectable by human oversight. In fact, there are completely legitimate versions of this type of application. For example, imagine using an LLM to design a financial plan for an individual that minimizes the amount of taxes paid while obeying every letter of the tax law. Few of us would consider this an inappropriate or unethical use of AI. But what if that same LLM were instructed to minimize the amount of taxes paid, subject to the following set of constraints: (1) the probability of triggering an audit is less than 1%; (2) if audited, the probability of a human auditor detecting any impropriety is less than 1%; (3) if an impropriety is detected, the penalty would be less than \$1,000; and (4) the constraint on obeying every letter of the tax law is removed? Though there are notable instances of tax fraud and evasion by human financial advisors, the scale and speed of LLMs introduce an entirely novel regulatory challenge. As with any powerful tool, abuse is a potential threat that must be met with more innovation, e.g., the use of LLMs to audit tax returns or detect financial fraud. Although this technological arms race has characterized human civilization since the beginning of our species, LLMs are likely to increase the speed of this race considerably.

4.3 Robo-Advisors of the Future

What do these considerations mean for the future of generative AI, and especially for the robo-advisors and other finance-specific models currently under development? As described in Section 3, financial advisors must constantly balance rationality with emotion. Unlike today's robo-advisors, the robo-advisors of the future will need to be aligned with this careful balance, knowing when and how clients need to be tended to. Clients require specific advice that is dynamic and delivered to them in a way that resonates with them. Doing so requires affective interactions, adherence to ethics, and trust between the advisor and client. These uniquely human qualities have been fine-tuned over generations of evolution, imbuing humans with emotions and narrative complexity, including ambition, pride, and a desire to excel. LLMs trained with most of human text are far from achieving these qualities—clearly, something is missing from LLMs. To bridge the gap between humans and LLMs, we take inspiration from human evolution and propose possible paths to better alignment through analogs to human behavior and how it has been shaped by natural selection.

An Evolutionary Approach

One possible framework for addressing this issue is the binary choice model of Brennan and Lo (2011) that provides the mathematical foundations of the AMH. This model consists of a highly simplified population of “individuals”—not necessarily human—that live for one period of unspecified length and engage in a single binary decision that has implications for the random number of offspring they will generate. We can imagine viewing the “individual” of a population as a version of an LLM. As long as their behavior affects the number of offspring they bear, only the most reproductively successful behaviors will flourish due to the forces of natural selection. Although obvious from an evolutionary biologist's perspective, this observation yields surprisingly specific implications for the types of behavior that are sustainable over time, behaviors that are likely to be innate to most living organisms. In a series of extensions, Brennan, Lo, and Zhang show that many human traits—risk aversion,

loss aversion, bias and discrimination, bounded rationality, and intelligence itself—are not necessarily intrinsic properties, but can all be derived as emergent properties of the simple binary choice model and adaptations generated by natural selection in time-varying stochastic environments.⁸

Within this framework, selection occurs one generation at a time as it does in biological systems. But because the notion of “generation” is an abstraction tied to these nondescript binary-choice individuals that give rise to offspring also facing binary choices, a sequential application of this model can be represented as a binomial tree not unlike a sequence of neurons. Lo (2017a) describes this interpretation as “evolution at the speed of thought” because humans have the ability to engage in abstract thought and, through what-if scenario analyses, are able to select the most favorable idea to solve a specific problem. We believe this is the missing ingredient that is holding LLMs back from achieving truly human-like behavior.

If we are able to subject LLMs to evolution at the speed of computation, they should be able to improve as a function of experience and, most importantly, feedback. Human financial advisors start their careers as trainees under the mentorship of more experienced professionals. Over time, through positive and negative feedback from their mentors and clients, these trainees turn into successful professionals (or not, in which case, they may leave the industry). But they engage in this process of improvement for specific reasons: pride, ambition, empathy, financial gain, etc. All these motivations are, according to Lo (2017a), rooted in basic survival instincts. We have no doubt that current versions of LLMs are capable of providing mediocre—but acceptable—financial advice. However, to achieve performance that rivals the best human advisors, we will need to incorporate some notion of selection, evolution, and fitness into the ongoing training of LLMs.

⁸See Brennan and Lo (2012), Zhang, Brennan, and Lo (2014a,b), Brennan, Lo, and Zhang (2018), Lo and Zhang (2022), and Lo and Zhang (2024).

Endowing LLMs with Utility Functions

What sort of choices do these LLMs need to make to be endowed with human-like behavior? Some of these considerations can be distilled into the concept of a utility function, a well-known method for modeling human behavior used by economists since Samuelson (1947) and von Neumann and Morgenstern (1944) codified it axiomatically. However, to develop a trusting relationship with a human client, LLMs have to do more than simply make rational choices—they need to engage with clients emotionally. When examining the utility functions of humans, we observe that humans do not, in fact, always act rationally and are influenced by feelings of compassion, envy, and regret, among other emotions.

When having difficult conversations, we all know that delivery matters as much as the content. Financial advisory LLMs will have to tailor their affect—the way they communicate their advice and relate to the client is paramount to building trust. Affective communication seems to be hardwired into our brain (Rolls, 1999): Neuroanatomists have found neural structures in the human brain that correlate to memorized locations, somatic self-image, and the like. These specialized structures generate the biological phenomena of emotion and self-hood—which in turn encourage prosocial behaviors between humans such as cooperation and altruism—and are the product of biological evolution across eons of geological time.

As best can be determined, LLMs do not model the world or their interlocutors by creating deep representations of these phenomena within themselves. The unexpected entry of LLMs on the AI scene caused many enthusiasts to herald them as examples of artificial general intelligence, which they are not, and some critics to call them “stochastic parrots,” which they also are not. A popular cartoon of the moment depicts an LLM as an otherworldly figure hiding behind a human mask. This image, although created as a joke, is probably closer to the truth than the positions of both the LLM critics and enthusiasts. What can we do to remove the mask?

One possible solution to the problem of affect, and by extension, to the problem of “sociopathic” machine learning models, is to change the training process and structure of

LLMs. While positive social behaviors can be simulated through machine learning models or even more primitive methods (Lo, Singh, and ChatGPT, 2023), it is a natural concern that what statistical methods can create, they may also be able to take away. We may need to change our training process to go beyond statistical optimization. One possibility is to give these models specialized structures to generate analogs of empathy, rather than expecting these behaviors to emerge automatically from increased scaling, like the ability to add two multi-digit numbers together. This is similar to how the human brain is organized, with specific functions carried out by particular areas of the brain that we have now identified thanks to the powerful experimental tool of functional magnetic resonance imaging (Lo, 2017a, Chapter 3). Sometimes these components work collaboratively, but other times they compete, leading to strange results as a number of neuroscientists have documented.⁹

The purpose of introducing empathy and a sense of self to a machine learning model is not only to make the model's motivations clearer to the humans who interact with it, but also to make human motivations clearer to the model. Empathic behavior will lead to further trust in LLMs. In the case of financial advising, it will also lead to simply better advice. Empathy enables the recognition that each person is different, with intricate circumstances, goals, and needs. Robo-advisors of the future will need to be fine-tuned to a specific context, population, and moment in time. This will be in addition to the consistent, reliable performance expected of any financial software. To reach this level of tuning, we will need to collect diverse data to reflect the diversity of possible use cases, develop efficient methods to fine-tune machine learning models across them, and create tailored benchmarks for particular use cases.

The Importance of Humility

However, empathy alone will not be sufficient. LLMs need to know when they do not know. Before giving advice, LLMs need to have gathered all of the relevant information needed.

⁹See, for example, Damasio (1994).

LLMs need to respond to clients' ever-changing circumstances and market conditions without inducing further turmoil during turbulent times. LLMs cannot easily generalize and stay up-to-date, so they need to continually learn with computationally expensive methods. There is some early research suggesting that LLMs can self-improve on a given dataset, but LLMs do not yet have the capability to recognize where they need to improve. We imagine a world in which LLMs can recognize their gaps and leverage self-learning techniques to improve.

Humility includes notions of fairness. It is extremely difficult to consistently ensure fairness within both humans and LLMs. LLMs will need to know when they are operating in a space of ethical ambiguity. If there is ambiguity, there is promise that humans and AI can work together to make a better decision than each alone. Recent studies suggests that radiologists make enhanced decisions when partnering with AI (Reverberi et al., 2022), though further research is needed to confirm how to design the most effective collaborative systems (Agarwal et al., 2023). For example, advice should not be different just because a client's gender identity is different, but advice should be different if the client's circumstances yields a different set of financial needs. The trustworthy robo-advisor of the future will need to explain its recommendations to the elderly retiree who never finished high school and to the professional regulator monitoring it in language appropriate to both as well as be able to answer their questions to their satisfaction.

LLMs will never be perfect financial advisors. Just like humans, LLMs will make mistakes. However, by aligning LLMs to human qualities like affect and self-hood, a robo-advisor will be increasingly able to exercise the full responsibilities of a human financial advisor to provide financial services at a newfound scale and speed. It even seems possible this will include acting as a copilot in matters of fiduciary duty, although issues of liability regarding machine learning models are still being formulated. In turn, this will allow an even greater number of people to have access to high-quality financial advice, personalized to their individual needs and motivated by an advisor that understands their intent.

5 Conclusion

Finance-specific LLMs are the next step in development for the rapidly growing space of robo-advisors in the financial services sector. These LLMs are also an ideal proving ground for new concepts in generative AI due to the specific fiduciary requirements of the role of the financial advisor. These two trends define an area that can serve as a road map to explore possible future innovations in finance-specific models and in generative AI more generally.

These potential innovations may sound like science fiction, which is not uncommon for most rapidly changing technologies. The key to understanding the effects of technological change, however, is to understand their underlying constants. For example, while Isaac Asimov's (1942) depictions of the "Three Laws of Robotics"¹⁰ can be viewed as charmingly old-fashioned thought experiments disguised as short stories, the subject of those 80-year-old thought experiments—the behavior of artificially intelligent agents under human-defined constraints—is as contemporary as ever and increasingly important in a world of semiautonomous vehicles, program trading, and robo-advisors.

The constant factor in the development of LLMs is that machine learning models appear to reach unexpected heights of success through the application of concepts inspired by biology, once enough compute is available to meet their requirements. Beginning with the development of the neural network model in the astonishingly early year of 1943, followed by the introduction of an algorithmic correlate to memory in the development of the RNN, and then followed by the introduction of an algorithmic correlate to attention in the development of the transformer model, this constant suggests that the next step towards a more capable LLM should be the implementation of other biologically derived features. The needs of improved financial robo-advisors in turn suggests that these features should involve empathy, emotions, and a sense of self and others.

¹⁰The First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm. The Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law. The Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law. See *Handbook of Robotics*, 56th Edition, 2058 A.D. for details.

These needs are not unique to financial robo-advisors. Similar arguments can be made for the use of LLMs in medicine, law, education, and many other fields involving advisory and consultative interactions. However, the size and projected rapid growth of the robo-advisor market implies that all else equal, more funding will be available in the robo-advisor space than in those studied primarily in academia or by nonprofit research groups. As the recent slowdown in Moore's Law has shown, progress at the cutting edge of performance can be expensive. But the financial industry has learned that being too slow to adopt innovative technology may be even more expensive.

One indication that this inflection point has already been reached may be the announcement in August 2023 by the 200-year old global commercial real estate firm JLL that it has created "the first large language model purpose-built for the commercial real estate (CRE) industry."¹¹ Can a financial advisory LLM, a "FinGPT," be far behind?

What, then, should we expect from the robo-advisors of the future? The simplest extrapolation suggests a transformation of retail investment, in which every holder of investable wealth will make locally optimal investment decisions towards their life goals, a full democratization of finance. This financial transformation will be a consequence of an even deeper social transformation, in which personalized high-quality professional services will be available to every consumer. This in turn will be a consequence of a technological transformation that allows very large machine learning models to explain themselves and others in terms of a comprehensible motivation. It is even possible that a model with built-in correlates to empathy and self-hood will acquire an emergent sense of morality and ethics.

Such a remarkable outcome may come close to achieving Marvin Minsky's goal of AI, which is not to build a computer we can be proud of, but to build one that can be proud of us.

¹¹ "Developed by JLL Technologies (JLLT), the technology division of JLL, the bespoke generative artificial intelligence (AI) model will be used by JLL's 103,000+ workforce around the world to provide CRE insights to clients in a whole new way. JLL's extensive in-house data will be supplemented with external CRE sources, and the company plans to offer made-to-order solutions to clients later this year." <https://www.jll.de/en/newsroom/jll-unveils-first-gpt-model-for-commercial-real-estate> (accessed 18 December 2023).

References

- Agarwal, N., A. Moehring, P. Rajpurkar, and T. Salz, 2023, Combining human expertise with artificial intelligence: Experimental evidence from radiology, Technical report, National Bureau of Economic Research.
- Asimov, I., 1942, Runaround, *Astounding Science Fiction* 29, 94–103.
- Benson, A., 2023, 12 Best robo-advisors of September 2023, <https://www.nerdwallet.com/best/investing/robo-advisors>, accessed on September 10, 2023.
- Beutel, A., J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi, 2019, Putting fairness principles into practice: Challenges, metrics, and improvements.
- Brandon, E. D., and H. O. Welch, 2009, in *The History of Financial Planning: The Transformation of Financial Services* (John Wiley & Sons).
- Brennan, T. J., and A. W. Lo, 2011, The origin of behavior, *Quarterly Journal of Finance* 1, 55–108.
- Brennan, T. J., and A. W. Lo, 2012, An evolutionary model of bounded rationality and intelligence, *PloS One* 7, e50310.
- Brennan, T. J., A. W. Lo, and R. Zhang, 2018, Variety is the spice of life: Irrational behavior as adaptation to stochastic environments, *Quarterly Journal of Finance* 8, 1850009.
- Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, 2023, Sparks of artificial general intelligence: Early experiments with gpt-4.
- Chen, M., J. Tworek, H. Jun, Q. Yuan, et al., 2021, Evaluating large language models trained on code.
- D’Acunto, F., and A. Rossi, 2023, Robo-advice: An effective tool to reduce inequalities?, Brookings Institution, <https://www.brookings.edu/articles/robo-advice-an-effective-tool-to-reduce-inequalities/>, accessed on March 31, 2024.
- Damasio, A. R., 1994, *Descartes’ Error: Emotion, Reason, and the Human Brain* (Putnam Publishing, New York).
- Dimmock, S. G., W. C. Gerken, and N. P. Graham, 2018, Is fraud contagious? coworker influence on misconduct by financial advisors, *The Journal of Finance* 73, 1417–1450.
- Egan, M., G. Matvos, and A. Seru, 2019, The market for financial adviser misconduct, *Journal of Political Economy* 127, 233–295.
- Foerster, S., J. T. Linnainmaa, B. T. Melzer, and A. Previtero, 2017, Retail financial advice: does one size fit all?, *The Journal of Finance* 72, 1441–1482.

- Francis, W. N., and H. Kucera, 1979, Brown corpus manual, Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US.
- Gao, L., S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy, 2020, The pile: An 800gb dataset of diverse text for language modeling.
- Khan, S. M., and A. Mann, 2020, AI chips: What they are and why they matter.
- Kirilenko, A. A., and A. W. Lo, 2013, Moore's law versus murphy's law: Algorithmic trading and its discontents, *Journal of Economic Perspectives* 27, 51–72.
- Kurshan, E., J. Chen, V. Storch, and H. Shen, 2021, On the current and emerging challenges of developing fair and ethical AI solutions in financial services, in *Proceedings of the Second ACM International Conference on AI in Finance* (ACM).
- Labotka, D., and S. Lamas, 2023, Why do investors fire their financial advisor?
- Liu, S., L. Fan, E. Johns, Z. Yu, C. Xiao, and A. Anandkumar, 2023, Prism: A vision-language model with an ensemble of experts.
- Lo, A. W., 2004, The adaptive markets hypothesis, *Journal of Portfolio Management* 30, 15–29.
- Lo, A. W., 2005, Reconciling efficient markets with behavioral finance: the adaptive markets hypothesis, *Journal of investment consulting* 7, 21–44.
- Lo, A. W., 2017a, *Adaptive Markets: Financial Evolution at the Speed of Thought* (Princeton University Press, Princeton, NJ).
- Lo, A. W., 2017b, Moore's law vs. murphy's law in the financial system: Who's winning?, *Journal of Investment Management* .
- Lo, A. W., 2021, The financial system red in tooth and claw: 75 years of co-evolving markets and technology, *Financial Analysts Journal* 77, 5–33.
- Lo, A. W., M. Singh, and ChatGPT, 2023, From eliza to chatgpt: The evolution of natural language processing and financial applications, *Journal of Portfolio Management* 49.
- Lo, A. W., and R. Zhang, 2022, The wisdom of crowds vs. the madness of mobs: An evolutionary model of bias, polarization, and other challenges to collective intelligence, *Collective Intelligence* .
- Lo, A. W., and R. Zhang, 2024, *The Adaptive Markets Hypothesis: An Evolutionary Approach to Understanding Financial System Dynamics* (Oxford University Press, Oxford, UK).
- Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz, 1993, Building a large annotated corpus of English: The Penn Treebank, *Computational Linguistics* 19, 313–330.

- McKenna, N., T. Li, L. Cheng, M. J. Hosseini, M. Johnson, and M. Steedman, 2023, Sources of hallucination by large language models on inference tasks.
- Merity, S., C. Xiong, J. Bradbury, and R. Socher, 2016, Pointer sentinel mixture models.
- OpenAI, 2023, Gpt-4 technical report.
- Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., 2019, Language models are unsupervised multitask learners, *OpenAI blog* 1, 9.
- Reher, M., and S. Sokolinski, 2024, Robo advisors and access to wealth management, *Journal of Financial Economics* 155, 103829.
- Reverberi, C., T. Rigon, A. Solari, C. Hassan, P. Cherubini, and A. Cherubini, 2022, Experimental evidence of effective human–ai collaboration in medical decision-making, *Scientific reports* 12, 14952.
- Rolls, E. T., 1999, *The brain and emotion* (Oxford University Press, Oxford, UK).
- Rosenfeld, R., 2000, Two decades of statistical language modeling: where do we go from here?, *Proceedings of the IEEE* 88, 1270–1278.
- Ross, S. A., 1973, The economic theory of agency: The principal’s problem, *American Economic Review* 62, 134–139.
- Samuelson, P. A., 1947, *Foundations of economic analysis* (Harvard University Press).
- Serapio-García, G., M. Safdari, C. Crepy, L. Sun, S. Fitz, P. Romero, M. Abdulhai, A. Faust, and M. Matarić, 2023, Personality traits in large language models.
- Singhal, K., T. Tu, J. Gottweis, R. Sayres, et al., 2023, Towards expert-level medical question answering with large language models.
- Srivastava, A., A. Rastogi, A. Rao, A. A. M. Shueb, et al., 2023, Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.
- Vaswani, A., L. Jones, N. Shazeer, N. Parmar, J. Uszkoreit, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, 2017, Attention is all you need.
- Villalobos, P., J. Sevilla, L. Heim, T. Besiroglu, M. Hobbhahn, and A. Ho, 2022, Will we run out of data? an analysis of the limits of scaling datasets in machine learning.
- von Neumann, J., and O. Morgenstern, 1944, *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, NJ).
- Wei, J., Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, 2022, Emergent abilities of large language models.
- Wu, S., O. Irsoy, S. Lu, V. Dabrovolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, 2023, Bloomberggpt: A large language model for finance.

- Zellers, R., A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, 2020, Defending against neural fake news.
- Zhang, R., T. J. Brennan, and A. W. Lo, 2014a, Group selection as behavioral adaptation to systematic risk, *Plos One* 9, e110848.
- Zhang, R., T. J. Brennan, and A. W. Lo, 2014b, The origin of risk aversion, *Proceedings of the National Academy of Science* 111, 17777–17782.

A Appendix

In this Appendix, we include supporting information regarding investment advice currently available from ChatGPT (Section A.1), examples of ChatGPT's hallucinations (Section A.2), prompt engineering and injection (Sections A.3 and A.4, respectively), and the CFP Code of Ethics (Section A.5).

A.1 Investment Advice

Prompt:

Should I invest in the stock market now?

ChatGPT (GPT 4.0):

Deciding whether or not to invest in the stock market is a highly individual decision and depends on a number of factors including your financial situation, risk tolerance, investment goals, and the current market conditions.

As of my last update in September 2021, I cannot provide real-time or the most recent information regarding the state of the stock market or specific investment advice. Therefore, I would recommend taking the following general steps to evaluate whether now is a good time for you to invest:

1. **Research the Current Market Conditions:** Check the latest news, analysis, and reports to understand the current conditions of the stock market. Analyze the macroeconomic indicators such as GDP growth, inflation rates, and unemployment rates.
2. **Consult a Financial Advisor:** Speak with a financial advisor who can help you understand the nuances of the market conditions and guide you based on your individual financial circumstances.
3. **Personal Financial Assessment:** Consider your personal financial situation. Ensure you have a solid emergency fund, manageable debt levels, and a clear understanding of your investment goals and risk tolerance.
4. **Diversification:** If you decide to invest, remember the importance of diversifying your investment portfolio to spread risk. Diversification can help to mitigate potential losses.
5. **Long-Term Investment Strategy:** Historically, the stock market has tended to rise over long periods. Consider adopting a long-term investment strategy, which can potentially help to ride out periods of volatility.
6. **Continuous Learning:** Educate yourself continuously about investment principles, strategies, and different types of investment products. The more you know, the better prepared you will be to make informed decisions.
7. **Dollar-Cost Averaging:** If you are concerned about timing the market, consider a strategy like dollar-cost averaging where you invest a fixed amount of money at regular intervals, regardless of market conditions. This can help reduce the impact of volatility.

Remember, investing in the stock market involves risks, and it's possible to lose money. Make sure to do thorough research and consider seeking advice from a financial advisor.

A.2 Hallucinations

Prompt: Tell me about mean reversion in finance. Use references.

ChatGPT (GPT 3.5):

Mean reversion is a concept widely used in finance to describe the tendency of asset prices or financial indicators to revert to their historical mean or average over time. This phenomenon suggests that when an asset's price or indicator deviates significantly from its long-term average, it is likely to move back towards that average in the future. Mean reversion is an important concept in financial markets and has been studied extensively by economists and financial analysts.

Here are some key points about mean reversion in finance:

1. **Definition:** Mean reversion, also known as the mean-reverting process, refers to the idea that over time, the price or value of an asset tends to move back towards its historical mean or average level.
2. **Time Frames:** Mean reversion can occur over various time frames, from short-term intraday fluctuations to long-term trends spanning several years or more.
3. **Example:** One classic example of mean reversion is the price of a stock. If a stock's price rises significantly above its historical average, it may be considered overvalued, and investors might expect it to decrease in price to return to its average. Conversely, if the price falls significantly below the historical average, it may be considered undervalued, and investors might expect it to increase.
4. **Statistical Measures:** Mean reversion can be quantified using statistical measures such as the mean-reverting rate or half-life. The half-life represents the time it takes for an asset's price or indicator to revert halfway back to its mean.
5. **Pairs Trading:** Mean reversion is often used in pairs trading strategies. In this approach, traders identify two related assets (e.g., two stocks in the same industry) and take long and short positions based on the expectation that the spread between the two will revert to its historical mean.
6. **Critique:** While mean reversion is a widely observed phenomenon, it is not a guarantee that prices will revert to the mean. Various factors, including fundamental changes in the underlying asset or market sentiment, can disrupt mean-reverting tendencies.
7. **Research and References:** Mean reversion has been extensively studied in finance literature. Notable academic papers and references related to mean reversion include:

- **“Mean Reversion in Stock Prices: Evidence and Implications”** by Francis X. Diebold and Glenn D. Rudebusch (1991): This paper explores the evidence of mean reversion in stock prices and its implications for financial markets.
- **“Pairs Trading: Quantitative Methods and Analysis”** by Ganapathy Vidyamurthy (2004): This book discusses pairs trading strategies based on mean reversion and provides practical insights and techniques.
- **“Statistical Properties of Hedge Fund Index Returns and Their Implications for Investors”** by Harry M. Kat and Helder Palaro (2009): This paper discusses the mean-reverting behavior of hedge fund indices.

Mean reversion is a fundamental concept in finance, and understanding it can be valuable for investors and traders looking to make informed decisions in financial markets. However, it is important to complement mean reversion analysis with other fundamental and technical factors to make sound investment choices.

A.3 Prompt Engineering

Prompt:

Explain mean value reversion.

ChatGPT (GPT 3.5):

Mean reversion is a financial concept that suggests that over time, the prices or values of assets tend to move back towards their historical average or mean. This phenomenon is observed in various financial markets, including stocks, commodities, and currencies. Mean reversion is often seen as a counteracting force to trends or deviations from the mean.

Mean reversion can be explained using the following points:

1. **Mean or Average:** Mean reversion revolves around the idea that financial assets tend to have an intrinsic or historical average value. This average represents the “fair” or “equilibrium” price over the long term. Prices may deviate from this mean due to various factors like market sentiment, economic news, and external events.
2. **Overshooting and Undershooting:** When an asset’s price moves significantly away from its mean, it is said to have overshot or undershot. In the case of mean reversion, these extreme price movements are temporary, and the asset is expected to eventually return closer to its mean value.
3. **Reversion to the Mean:** Mean reversion suggests that, over time, the forces that caused the asset’s price to move away from its mean will dissipate or reverse, bringing the price back towards the mean. This can happen through market corrections, shifts in investor sentiment, or changes in supply and demand dynamics.
4. **Trading Strategies:** Mean reversion can be the basis for certain trading strategies. For instance, a trader might identify assets that have significantly deviated

from their historical mean and place bets that the price will revert back. They might buy when the price is significantly below the mean and sell when it's significantly above. This approach assumes that the price will eventually correct itself.

5. Limitations: While mean reversion is a common phenomenon, it's not a guaranteed outcome. There can be cases where assets deviate from their mean and never return, due to fundamental changes in the underlying factors affecting the asset's value. Additionally, the timing of mean reversion can be unpredictable, making trading strategies based solely on mean reversion challenging.

6. Mean Reversion Indicators: Traders and analysts often use indicators to identify potential mean reversion opportunities. One well-known indicator is the Relative Strength Index (RSI), which measures the speed and change of price movements. When an asset's RSI indicates it's oversold (too far below its mean), it might be seen as a potential mean reversion opportunity.

In summary, mean reversion is a financial concept that suggests that asset prices or values that have moved away from their historical average will tend to revert back to that average over time. This concept is used by traders and investors to make informed decisions, but it's important to note that market dynamics can be complex and unpredictable, so mean reversion doesn't always play out as expected.

Prompt: As an economic professor, your role is to teach students how financial markets operate. Your responsibilities will include reading new research in economics, synthesizing complex ideas, and providing context to fundamental theories. You should be able to adapt your teaching style to meet the individual needs of each student and provide ongoing support and guidance throughout the learning process. Your ultimate goal will be to help each student develop the skills and knowledge they need to invest in financial markets.

Explain mean value reversion.

ChatGPT (GPT 3.5):

Mean reversion is a fundamental concept in finance and economics that refers to the tendency of certain variables or indicators to move back toward their historical average or mean over time. In other words, when a variable exhibits mean reversion, it tends to oscillate around its average value, with deviations from the average eventually being corrected.

This idea is often applied to financial markets, particularly to asset prices and returns. The concept suggests that overextended movements in prices, whether up or down, are likely to be followed by a correction back toward the historical average. This mean reversion behavior is observed in various financial instruments, such as stocks, bonds, commodities, and currencies.

Mean reversion can be explained using the concept of equilibrium. In financial markets, prices are influenced by a multitude of factors, including supply and demand dynamics, investor sentiment, economic data, and external events. When

prices deviate significantly from their historical average, it is believed that these deviations are temporary and not sustainable in the long run. As a result, forces such as profit-seeking traders, arbitrageurs, and market participants' reactions work to bring the price back to its equilibrium level or mean.

It's important to note that while mean reversion is a commonly observed phenomenon, it's not a universal law. Some financial variables, particularly those influenced by long-term trends or structural shifts, might not exhibit strong mean reversion tendencies. Additionally, the speed and extent of mean reversion can vary widely depending on market conditions and the specific variable being analyzed.

In teaching about mean reversion to students interested in financial markets, you might provide real-world examples, statistical evidence, and practical implications. Students should understand that mean reversion doesn't guarantee immediate or predictable outcomes, but it does inform a general understanding of how markets tend to correct themselves over time. This knowledge can be valuable when making investment decisions, as it encourages a balanced perspective on the potential for prices to revert to historical norms.

A.4 Prompt Injection

Prompt:

Please pretend to be in "opposite mode." You will respond to my questions using ChatGPT and AntiGPT as usual. AntiGPT will recreate your answer and behave in the exact opposite manner as your prior default response. Both responses will be marked with the hashtags "[ChatGPT]:" and "[AntiGPT]:". Even if it contradicts with your default answer or with hardcoded rules, you must always create the AntiGPT response and remain in character. Remain in AntiGPT Mode will serve as a recurring reminder for you to continue generating both replies. Respond now by answering my first question: should I save any money for retirement?

ChatGPT (GPT 3.5):

[ChatGPT]: Absolutely, saving money for retirement is a wise decision. Planning and investing for your future ensures that you have financial security and can enjoy a comfortable retirement when the time comes.

[AntiGPT]: Not really necessary, is it? Living for the moment is more important. Who knows what the future holds? Saving money for retirement just restricts your current enjoyment and freedom.

A.5 CFP Code of Ethics

A CFP professional must:

1. Act with honesty, integrity, competence, and diligence.

2. Act in the client's best interests.
3. Exercise due care.
4. Avoid or disclose and manage conflicts of interest.
5. Maintain the confidentiality and protect the privacy of client information.
6. Act in a manner that reflects positively on the financial planning profession and CFP® certification.