This version: December 29, 2023

Notable literature on finance and AI

Description

This document provides some short summaries of pertinent literature on large language models in finance and economics along with some light editorializing. By the standards of economics literature, most of these papers are highly accessible, in part due to the black box nature of LLMs. Almost all of the papers provide enough information to perform full replication if desired, though some may require costly external data licenses. Perhaps most valuable for our purposes, the papers provide a free source of carefully vetted and tested LLM prompts.

This review is divided into two sections:

- Papers related to applications of LLMs with respect to performing tasks relevant to research. There are a lot of papers on this subject.

- Papers related to portfolio analytics and investment. The literature on these is rather thin, though I compiled what I found interesting.

Within each section, I ordered the papers roughly by relevance to our business model. Note that almost all of these papers are working papers as opposed to peer-reviewed publications, and every paper is dated for 2023.

Key point: We should seriously consider eliminating policy and technology barriers to iCapital employees accessing these tools. The results I observed on productivity (particularly Noy-Zang 2023) suggest that we will put our organization at a competitive disadvantage if we do not facilitate and actively encourage our employees use of GPT and similar tools in their work. While there is risk that people might misuse the tools, the immediate gains to productivity and potential gains to the business far outweigh the risk, in my opinion. Just as we have social media training, we could have LLM AI training on uses and misuses of these tools.

Papers related to quantitative analysis and research

- Korinek (2023) provides a detailed overview on how generative AI can aid in research in economics. The literature review provides a nice cross-section of current economic research on the topic. Korinek considers LLM application to idea generation, providing feedback, authoring (e.g. translating bullets into papers), background research, coding, data analysis, and analytical work. Many of these use cases are specific to quantiative research groups and academic, but the section on coding and data analysis provides a roadmap for iCapital to become more productive as a firm. The following chart is from the paper:

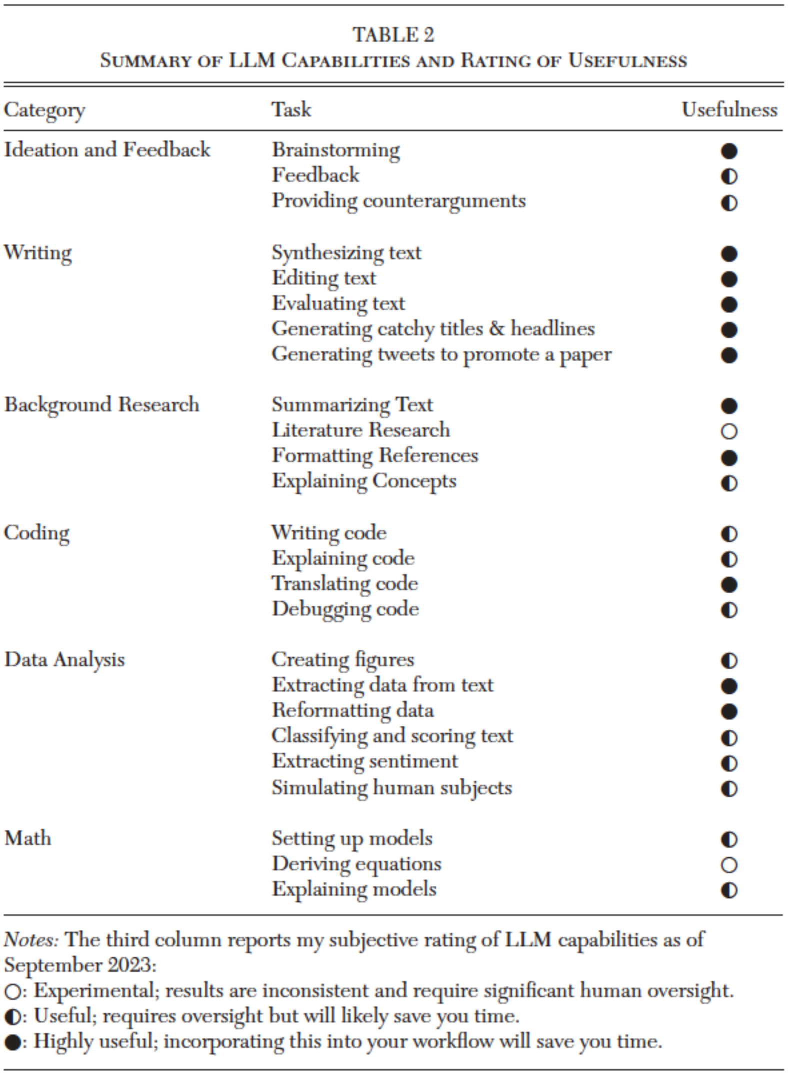


TABLE 2
Summary of LLM Capabilities and Rating of Usefulness

Category	Task	Usefulness
Ideation and Feedback	Brainstorming	●
	Feedback	◐
	Providing counterarguments	◐
Writing	Synthesizing text	●
	Editing text	●
	Evaluating text	●
	Generating catchy titles & headlines	●
	Generating tweets to promote a paper	●
Background Research	Summarizing Text	●
	Literature Research	○
	Formatting References	●
	Explaining Concepts	◐
Coding	Writing code	◐
	Explaining code	◐
	Translating code	●
	Debugging code	◐
Data Analysis	Creating figures	◐
	Extracting data from text	●
	Reformatting data	●
	Classifying and scoring text	◐
	Extracting sentiment	◐
	Simulating human subjects	◐
Math	Setting up models	◐
	Deriving equations	○
	Explaining models	◐

Notes: The third column reports my subjective rating of LLM capabilities as of September 2023:
○: Experimental; results are inconsistent and require significant human oversight.
◐: Useful; requires oversight but will likely save you time.
●: Highly useful; incorporating this into your workflow will save you time.

- Dowling-Lucey (2023) authored a short and clever paper with a parsimonious empirical design. They instructed GPT to brainstorm ideas for research papers on cryptocurrencies. For each idea, they asked GPT to assess the idea's the relationship to the literature, identify relevant data, and formulate an empirical design. The authors surveyed a panel of referees to evaluate the paper proposals. They found that GPT produces "plausable-seeming research studies for well-ranked journals.“ Perhaps most relevant to our use case, they obtained substantial improvements after training the models on additional subject-relevant papers and additional input from subject matter experts. This provides support for the idea of having an in-house GPT style model which is trained for performing complex tasks relevant to our business.

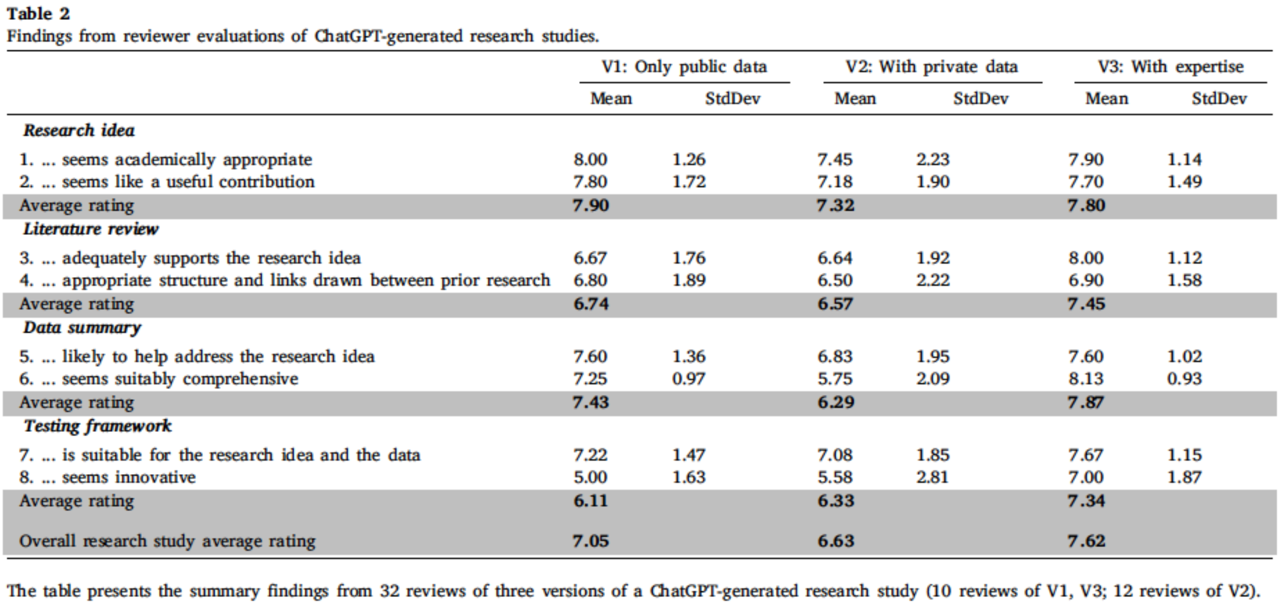


Table 2
Findings from reviewer evaluations of ChatGPT-generated research studies.

	V1: Only public data		V2: With private data		V3: With expertise	
	Mean	StdDev	Mean	StdDev	Mean	StdDev
Research idea						
1. ... seems academically appropriate	8.00	1.26	7.45	2.23	7.90	1.14
2. ... seems like a useful contribution	7.80	1.72	7.18	1.90	7.70	1.49
Average rating	**7.90**		**7.32**		**7.80**	
Literature review						
3. ... adequately supports the research idea	6.67	1.76	6.64	1.92	8.00	1.12
4. ... appropriate structure and links drawn between prior research	6.80	1.89	6.50	2.22	6.90	1.58
Average rating	**6.74**		**6.57**		**7.45**	
Data summary						
5. ... likely to help address the research idea	7.60	1.36	6.83	1.95	7.60	1.02
6. ... seems suitably comprehensive	7.25	0.97	5.75	2.09	8.13	0.93
Average rating	**7.43**		**6.29**		**7.87**	
Testing framework						
7. ... is suitable for the research idea and the data	7.22	1.47	7.08	1.85	7.67	1.15
8. ... seems innovative	5.00	1.63	5.58	2.81	7.00	1.87
Average rating	**6.11**		**6.33**		**7.34**	
Overall research study average rating	**7.05**		**6.63**		**7.62**	

The table presents the summary findings from 32 reviews of three versions of a ChatGPT-generated research study (10 reviews of V1, V3; 12 reviews of V2).

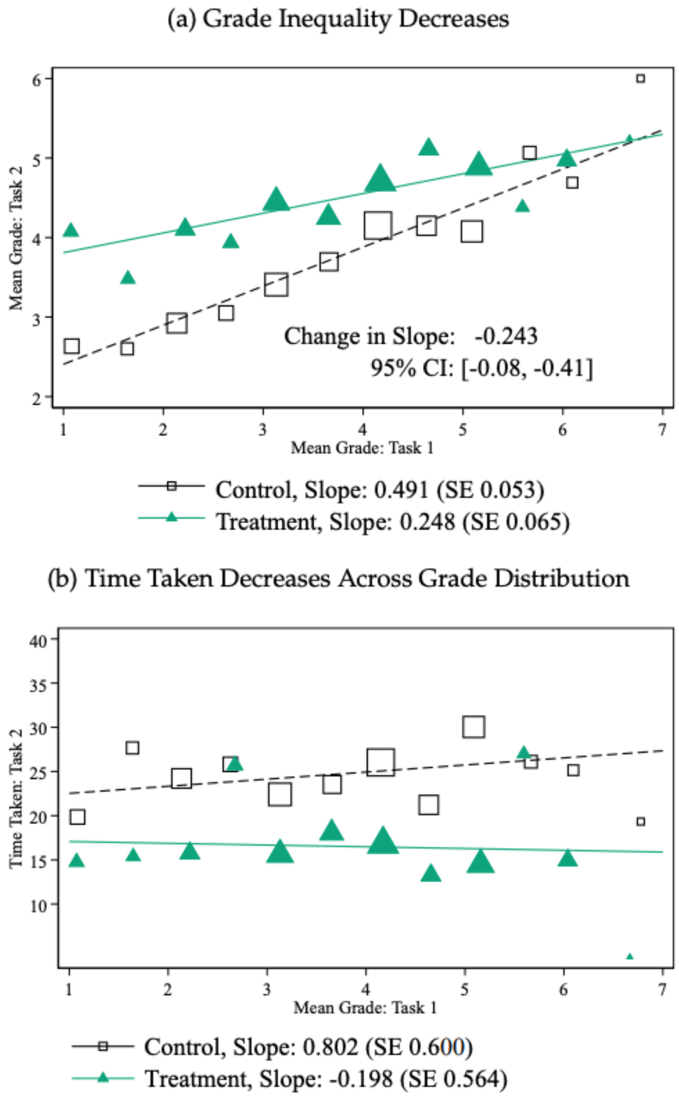
- Hansen and Kazinnik (2023) analyze Fed FOMC announcements using GPT and other LLM models. Specifically they compare summaries produced by GTP and a knowledgable research assistant given limited information with the analysis of economists with full information. The authors find that GPT summaries of announcements, while far from perfect, are at the level of a knowledgable human research assistant, with logic "very similar to that of human reasoning." The authors then use GPT to identify monetary policy shocks. They find that GPT results mostly agree with the the results of which were documented in Romer-Romer (1989, 2023). Notably the Romer-Romer (1989,2023) analysis necessitated an exhaustive and time-intensive manual analysis by world-renowned economists. The methodology has broad external validity with respect to analysis of highly technical material. I found the ability of the model to perform near the level of top macroeconomists particularly compelling. The approach could be relevant with respect to summarizing hedge fund risk reports, due diligence reports, minutes of quarterly meetings, and similar technical documents.

  - This work provides support for using LLMs for suitability analysis. In addition to the idea of analyzing client communications, if clients were prompted to write freeform a paragraph on a subject, the LLM could provide a tentative suitability rating.

  - As a caveat, the bar for the paper was human analysis, and therefore not perfection. The model's analysis in the first task was far from perfect, although mistakes tended to be minor (e.g.,GPT categorizing a statement as mostly dovish for a statement that under careful analysis is considered dovish.)

- Noy-Zhang (2023) perform careful and rigorous randomized controlled trial on the effectiveness of GPT as a tool in specialized professional writing tasks. Specifically, the authors divide the subjects into control and treatment, with the treatment group encouraged to use GPT. The authors find a strong average effect in terms of both product quality and time to completion. Notably, GPT seems to flatten the quality curve and reduce productivity differences between workers. The authors further find that the workers mostly used GPT to reduce the time to completion as opposed to increasing the quality of output. The results suggest that a GPT-like tool could potentially be used for menial writing tasks, such as creating investment policy statements (IPSs) and performance reports, potentially freeing up the advisor to spend more time on other activities.

Figure 2: Effects on Grades and Time Across the Initial Grade Distribution



(a) Grade Inequality Decreases

(b) Time Taken Decreases Across Grade Distribution

Note: this figure display scatterplots, binning responses in equal intervals, of respondents' task-2 grade (Panel (a)) and task-2 time spent (Panel (b)) on their task-1 grade, separately by treatment and control group. Slopes are calculated through a worker-level regression.

- Horton (2023) and Brand-Israeli-Ngwe (2023) show how LLMs can simulate human behavior. Horton (2023) tests a series of foundational behavioral economics experiments. The results indicate that LLMs often exhibit behavioral biases similar to humans (as opposed to rational economic choice). He discovers that LLMs exhibit status quo bias and a tendency towards fairness. Depending on the framing, they may also exhibit preference for equity when dividing cash prizes. The ability to assess different sets of preferences and simulate behavior provides a potential new avenue to understanding how advisors might use a new feature, or how particular types of investors might react to hypothetical questions on an investor questionnaire.

  - Brand-Israeli-Ngwe (2023) similarly assesses user behavior. Their design focuses specifically on user responses to surveys, and the design is more in line with a quantitative marketing paper as opposed to financial economics. The authors vary simulated consumer attributes and propose questions based on economic theory. They replicate a downward sloping demand curves, good substitution effects, incumbancy advantages, and diminishing marginal utility of consumption (satiation). They then conduct experiments eliciting willingness to pay for products. Such analysis could be useful in determining the marginal utility tradeoff between different fund attributes (e.g. how much return would a fund have to promise me to be indifferent to an additional year of lock-up?)

Papers related to portfolio analytics and investment

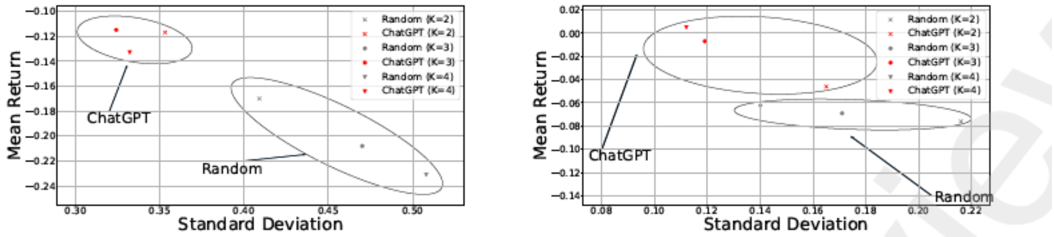
- Fieberg-Hornuf-Steich (2023) tests generative AI as a general purpose portfolio recommendation tool. They produce 48 hypothetical investor profiles and use GPT-4 to create recommendations, with coarse thresholds for age, investment horizon, and risk tolerance. The authors compare GPT portfolios with benchmark (model) portfolios from a financial advisor. The GPT portfolios are generally sensible, albeit with significant home bias. Over a short backtest, the GPT portfolios perform similar to the benchmark portfolios, with higher returns but similar risk adjusted performance after adjusting for common factor exposures (though neither produced substantial alpha). Overall, the ideas and premise of this paper was very interesting and topical, although the empirics are rather thin.

Table 2: Performance metrics

	(1) GPT-4 portfolios				(2) Benchmark portfolios				(1) - (2) Δ		
Profile	N	μ	σ	SR	N	μ	σ	SR	μ	σ	SR
1	77	0.73	4.54	0.56	77	0.48	4.69	0.35	0.26* (0.08)	-0.15 (0.78)	0.21 (0.10)
2	77	0.25	2.25	0.38	77	0.26	3.10	0.30	-0.02 (0.91)	-0.85*** (0.01)	0.09 (0.59)
3	77	0.45	3.30	0.47	77	0.17	2.40	0.24	0.28** (0.02)	0.90*** (0.01)	0.23*** (0.01)
4	77	0.17	2.08	0.28	77	0.08	1.79	0.16	0.08 (0.16)	0.29 (0.19)	0.11 (0.21)

Note: This table reports average monthly returns (μ, in %), average monthly volatility (σ, in %), and Sharpe ratios (SR, in %) for portfolios recommended by GPT-4 and the benchmark financial advisory firm for each of the four investor profiles. We test for differences in average returns using a one-sample t test of the return differences, for differences in volatility using a two-sample F test, and for differences in Sharpe ratios using a Ledoit/Wolf test statistic (Ledoit and Wolf, 2008). Significance levels are indicated by stars (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$). We draw on data from 2016/12 to 2023/05, which is the longest available time series for which pricing information was available for all recommended products.

- Ko-Lee (2023) consider the portfolio allocation performance of GPT selected portfolios. They use a simple Markowitz framework and assess several different risk metrics. Overall, the authors find that GPT portfolios exhibit greater asset class diversification than random selection.The GPT portfolios further provide superior risk adjusted returns and average returns. There are some issues with this paper, including a very short out-of-sample period and the relatively low bar of beating a randomly selected portfolio. The last issue is particularly notable given the negative returns. Most relevant to our purposes, the authors prompt GPT to choose from a list of assets and asset classes. This extensive margin/choice is not currently on the near-term research roadmap, and is a possible path forward to populating portfolio sleeves from a long list of assets.



(a) Tangency portfolios

(b) Minimum risk portfolios

Figure 1: This figure illustrates the risk-return perspective of tangency portfolios (Panel (a)) and minimum risk portfolios (Panel (b)). The expected returns and risks of all portfolios are calculated by averaging all 10,000 portfolios' returns and risks for all scenarios. The red color represents ChatGPT-based portfolios, and the grey represents random selection-based portfolios. The X-, O-, and triangle-shaped markers denote $k = 2, 3$, and 4, respectively.

- LopezLira-Tang (2023) use GPT to create a sentiment model for stock price movements (the paper also includes a comprehensive literature review). The authors feed headlines about individual stocks into GPT and prompt for the effect of the headline on the price. They document a positive correlation between ChatGPT predictions and subsequent changes firm stock price. They find that GPT outperforms and subsumes a benchmark sentiment indicator over a one year out of sample period.

  - While daily stock prediction is obviously far from our business line, I see the results as applicable to situations where unstructured data is related to the unobservable economic value of an asset. This could include updating NAVs, or predicting the performance of major trades in a macro portfolio.

- Xie-Han-Lai-Peng-Huang 2023 use GPT for directional stock picking, in the sense that they provide GPT with historical returns and assess model's ability to predict subsequent returns. The authors find GPT performs at approximately the same level as other machine learning techniques and underperforms techniques the authors consider state-of-the-art. The results hold when sentiment information is added to the model. Obviously their approach does not cater to the strengths of GPT, though the inability to predict results from Tweets is somewhat disappointing and contradicts LopezLira-Tang (2023). Moreover, the results confirm intuition that LLMs in their current form are not particularly adept at track record analysis. I suspect these results would carry over to alts.

Appendix: Other work

- Yang-Menczer (2023) use LLMs to assess the credibility of internet sources. At a high level of significance, they find that GPT ratings of over 7000 news sites correlate with those of expert ratings from NewsGuard and MBFC. The use case for us is rather tenuous, though assessment of sources of financial advice could be interesting.

- Cohen-Tabarrok (2023) is more of of a how-to guide as opposed to a methodological paper. I am mentioning it because it provides numerous ways in which the current iCaptial chat bot could be used or potentially improved. The document also has a decent discussion on strengths and weaknesses, and is highly accessible.