Finance Research Letters 53 (2023) 103662

Contents lists available at ScienceDirect

Finance Research Letters

journal homepage: www.elsevier.com/locate/frl

ChatGPT for (Finance) research: The Bananarama Conjecture

Michael Dowling [a,*], Brian Lucey [b,c,d,e]

[a] DCU Business School, Dublin City University, Ireland
[b] Trinity Business School, Trinity College Dublin, Ireland
[c] University of Abu Dhabi, Zayed City, Abu Dhabi, United Arab Emirates
[d] University of Economics, Ho Chi Minh City, Viet Nam
[e] Jiangxi University of Economics and Finance, Nanchang, China

ARTICLE INFO

JEL classification:
G00
G10

Keywords:
ChatGPT
Artificial intelligence
Finance research
Ethics

ABSTRACT

We show, based on ratings by finance journal reviewers of generated output, that the recently released AI chatbot ChatGPT can significantly assist with finance research. In principle, these results should be generalisable across research domains. There are clear advantages for idea generation and data identification. The technology, however, is weaker on literature synthesis and developing appropriate testing frameworks. Importantly, we further demonstrate that the extent of private data and researcher domain expertise input, are key factors in determining the quality of output. We conclude by considering the implications, particularly the ethical implications, which arise from this new technology.

1. Introduction

"It ain't what you do, it's the way that you do it
And that's what gets results"

[Song lyrics by Bananarama and Fun Boy Three (1982)]

ChatGPT is an artificial intelligence language model introduced in November 2022 providing generated conversational responses to question prompts. The model is trained with a blend of reinforcement learning algorithms and human input on over 150 billion parameters.[1] The platform reached a million users in just its first week open to the public and has been quickly coined "the industry's next big disrupter" (Grant and Metz, 2022) due to the perceived quality of response output from the model. Although Large Language Models, the technical term for the process underlying this chatbot, have been used for some decades, we were not able to find a study showing how they can be used in the research generation process, as opposed to being used as part of the research.

One early academic study found the platform capable of passing the notoriously-complex common core of US professional legal accreditation examinations (Bommarito II and Katz, 2022). Another author managed to produce a reasonably-comprehensive guide to quantitative trading, almost exclusively through *ChatGPT* output (Marti, 2022). A range of professions even set themselves to existential pondering as to whether they have suddenly been made redundant; including educators (Herman, 2022), lawyers (Greene, 2022), and, to cover as many worried professional bases as possible, 'all writers' (Warner, 2023). It is quite the entrance for new technology.

* Correspondence to: DCU Business School, Dublin City University, Glasnevin, Dublin 9, Ireland.
E-mail address: michael.dowling@dcu.ie (M. Dowling).

[1] https://openai.com/blog/chatgpt/

https://doi.org/10.1016/j.frl.2023.103662
Received 11 January 2023; Received in revised form 17 January 2023; Accepted 19 January 2023
Available online 25 January 2023

We are interested in the extent to which *ChatGPT* can assist with the production of research studies; in this case, finance research. Initial research has explored some limited aspects of this question. A broad perspective on the emergent role for AI in the production of scientific research is taken by Grimaldi and Ehrler (2023) and Hutson et al. (2022). While Alshater (2022) suggests that ChatGPT *should* be useful for a range of tasks involved in constructing a research study, but without empirical testing.

Most of the applied research focuses on the creation of research abstracts and literature synthesis. For example, Aydın and Karaarslan (2022) attempt to create a healthcare literature review suitable for an academic journal and find that while it is possible, there is considerable 'plagiarism', or poor paraphrasing. Gao et al. (2022), however, find that novel abstracts can be generated without explicit plagiarism, although these are identifiable as being generated by an AI platform using an artificial intelligence output detector.[2] Chen and Eger (2022) also explore use in title and abstract generation, and in the domain of finance, Wenzlaff and Spaeth (2022) are able to generate reasonably academically-appropriate definitions of new financial concepts.

Mellon et al. (2022) explores one aspect of the application to research testing, by showing the platform can be useful as a complement to scoring open-text survey results. While Adesso (2022) has used GPT3 to write a full paper in physics, to be submitted to a journal "as is", and Zhai (2022) has also experimented with creating a research paper outline.

Building on, but distinct from these studies, our study is the first to provide structured testing of the potential for *ChatGPT* to assist with writing a research study. We test and compare generated output for four stages of the research process: idea generation, literature review, data identification and processing, and empirical testing. A panel of experienced academic authors and reviewers grade each output. We also, importantly, show how different levels of private data and researcher domain-expertise input in guiding output have a significant impact on the quality of outputs generated. Like all tools, *ChatGPT* is best in experienced hands. Following the opening quote of this article, we term this *the Bananarama Conjecture*.

Section 2 outlines our empirical approach, Section 3 presents and analyses the findings. We conclude in Section 4 with a framework for understanding the opportunities and limitations of *ChatGPT*, as well as some initial consideration of the ethical dimensions of the new technology.

2. Methodology

We focus on cryptocurrencies as our finance topic — a prominent and reasonably well-defined area of recent finance research. We further concentrate on letter-style articles, such as those published in the *Finance Research Letters* journal, thus, articles of about 2000–2500 words in length.

We start our empirical approach by noting that the standard research study creation process can be divided into five basic stages (Cargill and O'Connor, 2021):

1. Idea generation
2. Prior literature synthesis
3. Data identification and preparation
4. Testing framework determination and implementation
5. Results analysis

As *ChatGPT* is currently unable to analyse empirical output we cannot evaluate the results analysis ability, so we concentrate on the first four stages of the research process. We, therefore, request the platform to generate: (1) a research idea; (2) a condensed literature review; (3) a description of suitable data for the research idea; and (4) a suitable testing framework given the research idea and the proposed data.

Three versions of the same general cryptocurrency research idea are generated, each with these four research stages. The textual prompts used to generate each stage are reported in Appendix. The first version only utilises public data already available within *ChatGPT*.[3] We label this version of the research study: *V1: Only Public Data*.

For the second version (labelled: *V2: Added Private Data*), we incorporate private data to assist with generating the research stages. We obtain abstracts and article identifiers for 188 articles identified as related to cryptocurrencies and published in Finance Research Letters (2021–2023) from the *Elsevier Scopus* database. These articles are loaded into *ChatGPT* in bibtex format.[4] The private data from these articles add specialist knowledge to the existing generalised expertise of the platform. We then generate the four research stages telling the platform to take this prior research into account.

For the third version (*V3: Private Data and Expertise*), we further incorporate researcher domain expertise alongside the private data. In practice, we take the outputs from the second version, and iterate the output, by telling *ChatGPT* how it might improve its suggested answers. Most frequently this iterative process involves asking the platform to be more specific on particular parts of the output, as it tends towards equivocation and generality unless guided otherwise. In none of the three cases do we manually adjust any of the output generated by the model, with the exception of one minor technical correction noted in Appendix.

[2] https://openai-openai-detector.hf.space/

[3] Note that the ChatGPT training data appears to have ended in 2020, thus the available data is quite dated for topical research ideas: https://help.openai.com/en/articles/6783457-chatgpt-faq.

[4] To load textual data into ChatGPT, we first selected the articles full citation in Bibtex format. These were then manually pasted, approx 10 at a time into the Chat window. We instructed ChatGPT "please read these bibtex references and store them for use in subsequent analyses until we log off. More complex data sets can be added using the OpenAI API, but ChatGPT can.

Table 1
Empirical structure.

Research stage	Evaluation criteria	Approx. length
Idea	(1) The proposed idea seems academically appropriate; (2) The proposed idea seems like useful contribution	100 words
Literature	(3) The literature review adequately supports the research idea; (4) The structure and links drawn between prior research are appropriate	300 words
Data	(5) The data is likely to help address the research idea; (6) The data seems suitably comprehensive	100 words
Testing	(7) The testing framework is suitable for the research idea and the data; (8) The testing framework seems innovative	200 words

The evaluation criteria column shows the questions asked of reviewers for that research stage, which they rate between 1 (highly disagree) and 10 (highly agree). The length column indicates the approximate word count of output requested from ChatGPT for that research stage. See Section 2 for further elaboration of labels and approach.

For our evaluation stage, a team of experienced authors and reviewers are identified who all have prior experience as reviewers or published authors for ABS-level[5] finance journals. A total of 32 reviewers each review a complete single version of the output (that is, all four research stages of a full research study), and are randomly assigned to one of the three versions.

We administer the evaluation through *Qualtrics*. The three generated versions of the research study, as presented to reviewers, are contained in Appendix. Reviewers are asked to rate two aspects of each stage of output, see Table 1 for this evaluation criteria, and may voluntarily leave comments. A review consists of a rating between 1 (highly disagree) and 10 (highly agree) of how likely the output is to be considered acceptable for a minimum ABS2-level[6] finance journal according to the specified criterion. Average scores across reviewers are reported.[7] We now proceed to present and analyse the findings.

3. Findings

Table 2 reports the main findings and Fig. 1 presents a boxplot representation of the results. The table shows the findings for all three research study versions, and for the four research stages. The research stages are, in turn, each evaluated according to two criteria.

We could view a rating of 5.5 (the mid-point of the rating range between 1 and 10) as a basic minimum for a research study stage to be considered acceptable. Possibly acceptable with revisions, and subject, naturally, to the element of randomness and personal preference that is always present in the reviewing process. By this basic criteria, all versions of the study 'succeed'. Reading from the bottom line of Table 2, which shows the overall average rating of each study, V1 has a rating of 7.05, V2 a rating of 6.63, and V3 a rating of 7.62. These are, therefore, all studies that have a decent chance of eventual success in the reviewing process in a good finance journal.

Examining the individual research stages, we see the highest ratings are for the generation of the research idea. This makes sense when we consider that this initial stage involves thinking broadly about existing concepts and connecting these concepts into a coherent new idea. *ChatGPT* with its access to billions of parameters and texts, should be particularly adept at this broad exploration of existing ideas. The data summary stage is also reasonably strong, perhaps because data summaries tend to be distinct sections of a research study in easily identifiable text 'chunks'. There is also a limited range of data which can be used in a given study, meaning the search process is also limited.

Less successful, according to our results, are literature reviews and testing frameworks. The platform particularly struggles with generating suitable testing frameworks. Our view here is that this might be due to these being 'internal' tasks within a research study. The literature review is the internal tool to link the research idea with the methodology. The testing framework is linked from the research idea, the literature review, and the data summary. The model appears to be less capable of linking multiple internally-generated ideas, such as these stages entail.

Comparing the different research versions we see a clear outperformance by our most advanced research study, *V3: Private Data and Expertise*. We were surprised to see that the version with added private data underperformed compared to the version with only public data. On reflection, this appears to be due to the private data model excessively relying on the provided private data and restricting the extent to which it accessed other beneficial public data. This could be improved by either instructing the platform to not ignore useful public data, or by providing a better-curated set of relevant private data.

[5] The ABS, more formally the Chartered Association of Business Schools Academic Journal Quality Guide, is a ranking of journals, widely used for assessing research output around the world, but particularly in the United Kingdom where the guide originates.

[6] ABS2 is a rating given to journals which publish research at an 'acceptable standard'. Anecdotally, it is viewed as the minimum standard of research expected in business schools which are mid-ranked and above: https://charteredabs.org/academic-journal-guide-2021-view/.

[7] All reviewers are informed that the content they are reviewing is generated by *ChatGPT* and that their individual responses will be kept anonymous.

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).

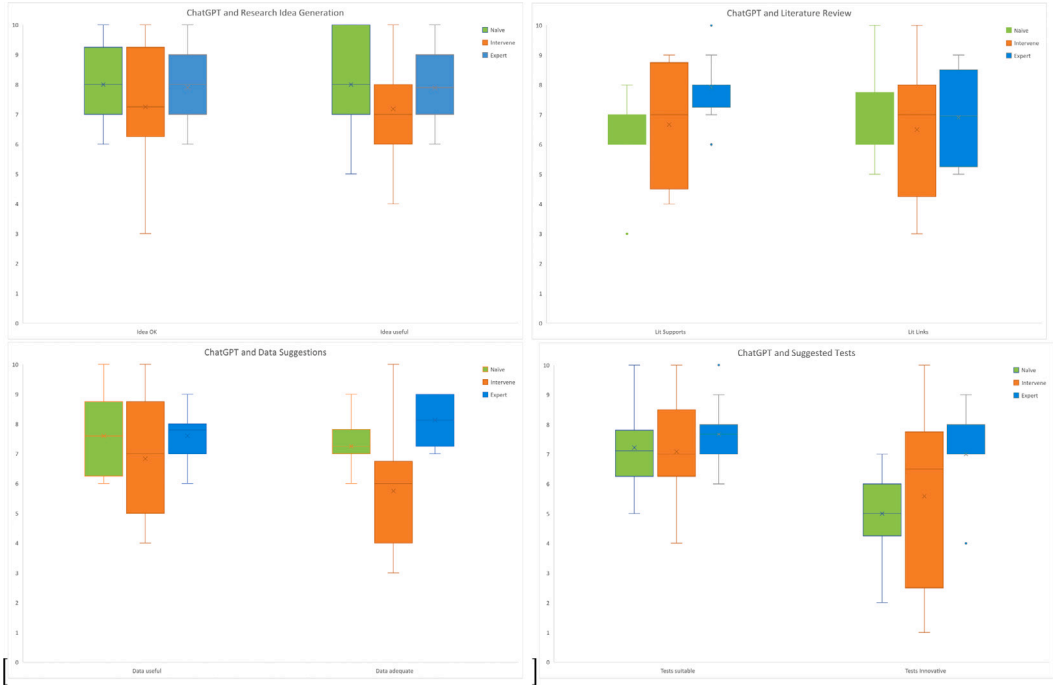

Fig. 1. Box-Whisker plots of responses.

The outperformance of the V3 research study is notable, not just on an overall basis, but also in the extent to which it is also capable of producing acceptable literature reviews and testing frameworks where the other research studies have less success. We suggested above that the general underperformance of the output for these research stages might be due to the difficultly *ChatGPT* has in linking multiple generated ideas. The advantage, therefore, for our V3 study, is that the researcher can observe any missing links and ask the platform to further iterate to address these gaps. The Appendix contains sample prompts given to the platform, and this addressing of missing links can be seen in the prompt text. Researcher domain-expertise appears to be key for these tasks involving conceptual complexity.

Table 3 confirms, statistically, the differences between the research studies through a range of t-tests. These two-sided t-tests assume unequal variance, as best fits our data. The main differences are observed for the evaluation criteria of "the literature review

Table 3
T-tests of differences between research study versions.

	V1–V2	V2–V3	V1–V3
Research idea			
1. ... seems academically appropriate	0.39	0.44	0.86
2. ... seems like a useful contribution	0.41	0.22	0.67
Literature review			
3. ... adequately supports the research idea	0.69	0.06	0.00
4. ... appropriate structure and links drawn between prior research	0.90	0.71	0.78
Data summary			
5. ... likely to help address the research idea	0.29	0.26	1.00
6. ... seems suitably comprehensive	0.08	0.01	0.02
Testing framework			
7. ... is suitable for the research idea and the data	0.84	0.37	0.37
8. ... seems innovative	0.44	0.25	0.00

The table reports p-values from two-sided t-tests assuming unequal variance, on tests between the three different research study versions — V1: Only Public Data; V2: Added Private Data; and V3: Private Data and Expertise.

adequately supports the research idea" and "the testing framework seems innovative" — in both cases the V3 research study shows some outperformance.

4. Conclusions

What we have shown in this study is important. *ChatGPT* can generate, even in its basic state, plausible-seeming research studies for well-ranked journals. With the addition of private data and researcher expertise iterations to improve output, the results are, frankly, very impressive. Bear in mind, also, that these results are obtained without the advantages of GPT-4 as an underlying generative model, due to launch later in 2023 and which promises a truly revolutionary language model due to advances in algorithms and over 600 times greater testing parameters.[8]

Our demonstration of this ability is, we believe, both novel and robust. The *novelty* lies in this being the first study to show the impact for each stage of the research process, and, importantly, for multiple levels of researcher input. The *robustness* lies in the reviewing process by which we ascertain the likely contribution of the generated research studies. The reviews bring the probable benefits of *ChatGPT* beyond conjecture to empirical verification, using a method by which research contribution is normally judged — the peer-review process.

So, what do we do now? This is both a practical and an ethical question. Can *ChatGPT* be simply considered as an *e-ResearchAssistant*, and, therefore, just a new part-and-parcel tool of how research is normally carried out? Indeed, under this perspective the platform might even be viewed as democratising access to research assistants, hitherto the reserved domain of wealthier universities in wealthier countries. Could *ChatGPT* help to flatten the disparities between the global south and wealthier nations in terms of research output? Maybe, now everyone can have access to such assistance, like the research-version of a dæmon from a Phillip Pullman novel following the researcher around and always available to offer pertinent advice.

There is, of course, a more worrying ethical perspective. Is it proper to have such an advanced level of guidance and assistance, and still claim the produced research as one's own? Should, for example, *ChatGPT*-enabled research be acknowledged on ethical research guidance frameworks, such as Elsevier's CRediT[9]? Certainly, the approach of Osterrieder and ChatGPT (2023) could be adopted, with credited co-authorship to the platform, but that is unlikely to be widespread practice.

The answer to the ethical issues is likely to be gradually understood, rather than immediately apparent. One useful guide to how this might play out is how AI-generated work is treated under copyright laws of various countries. Iaia (2022) notes that AI-generated work, with *sufficient* levels of human oversight, is generally considered to belong to the human-creator under European Union law. How 'sufficient' is defined is still, however, quite vague. That suggests the higher-levels of our generated research studies, with private data and iteration, could be considered the researcher's own work, but perhaps not the basic research study using only public data and simple question prompts. Adopting this perspective might see the opening *Bananarama Conjecture* of this article, become the (admittedly less-lyrical) *Bananarama Edict* for using *ChatGPT* for research; *it ain't what you do, its the extent that you do it, and that's what gets (ethically-acceptable) results.*

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

[8] https://techcrunch.com/2022/12/01/while-anticipation-builds-for-gpt-4-openai-quietly-releases-gpt-3-5/
[9] https://www.elsevier.com/authors/policies-and-guidelines/credit-author-statement

Data availability

Data will be made available on request.

Acknowledgements

We are hugely indebted to the following people, who quickly acted to review the generated research studies despite their own very busy schedules, and for offering their invaluable advice and feedback: Jonathan Batten, Sabri Boubaker, Elie Bouri, Kam Chan, Tom Conlon, Michael Froemmel, Tong Fu, Marie Helene Gagnon, John Goodell, Feng He, Pia Helbing, Shupei Huang, Wei Huang, Stuart Hyde, Sitara Karim, Suwan Long, Roman Matkovskyy, Charilaos Mertzanis, Muhammad Naeem, Haitham Nobanee, Duc Khuong Nguyen, Salvatore Perdichizzi, Boru Ren, Jacqui Rossovski, Dehua Shen, Andrew Urquhart, Vincenzo Verdoliva, Yizhi Wang, Larisa Yarovaya, Dayong Zhang.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.frl.2023.103662.

References

Adesso, G., 2022. GPT4: The ultimate brain. Authorea Preprints.
Alshater, M.M., 2022. Exploring the role of artificial intelligence in enhancing academic performance: A case study of ChatGPT. SSRN:4312358.
Aydın, Ö., Karaarslan, E., 2022. OpenAI ChatGPT generated literature review: Digital twin in healthcare. SSRN:4308687.
Bommarito II, M., Katz, D.M., 2022. GPT takes the bar exam. arXiv:2212.14402.
Cargill, M., O'Connor, P., 2021. Writing Scientific Research Articles: Strategy and Steps. John Wiley & Sons.
Chen, Y., Eger, S., 2022. Transformers go for the LOLs: Generating (humourous) titles from scientific abstracts end-to-end. http://dx.doi.org/10.48550/ARXIV.2212.10522, URL arXiv:2212.10522.
Gao, C.A., Howard, F.M., Markov, N.S., Dyer, E.C., Ramesh, S., Luo, Y., Pearson, A.T., 2022. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. BioRxiv:10.1101/2022.12.23.521610.
Grant, N., Metz, C., 2022. A new chat bot is a 'Code Red' for Google's search business. New York Times (Dec 21, 2022), Section B, Page 1.
Greene, J., 2022. Will ChatGPT make lawyers obsolete? (Hint: Be afraid). Reuters (Dec 09, 2022).
Grimaldi, G., Ehrler, B., 2023. AI others: Machines are about to change scientific publishing forever. ACS Energy Lett..
Herman, D., 2022. The end of high-school English. The Atlantic (Dec 09, 2022).
Hutson, M., et al., 2022. Could AI help you to write your next paper? Nature 611 (7934), 192–193.
Iaia, V., 2022. To be, or not to bełdots original under copyright law, that is (one of) the main questions concerning AI-produced works. GRUR International 71 (9), 793–812.
Marti, G., 2022. From data to trade: A machine learning approach to quantitative trading. SSRN:4315362.
Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., 2022. Does GPT-3 know what the most important issue is? Using large language models to code open-text social survey responses at scale. SSRN:4310154.
Osterrieder, J., ChatGPT, 2023. A primer on deep reinforcement learning for finance. SSRN:4316650.
Warner, J., 2023. Biblioracle: Will artificial intelligence like ChatGPT bring the end for all writers? Chicago Tribune (Jan 07, 2023).
Wenzlaff, K., Spaeth, S., 2022. Smarter than humans? Validating how OpenAI's ChatGPT model explains crowdfunding, alternative finance and community finance. SSRN:4302443.
Zhai, X., 2022. ChatGPT user experience: Implications for education. SSRN:4312418.