GPT-4 Calibration as a Research Assistant in Economics

Abstract

This paper defines the Plugin Forest, a best-practice prompt engineering strategy for research. We then leverage crowd feedback to comparatively calibrate GPT-4 summary literature reviews against human-authored documents related to a variety of research topics within the field of economics. We find that reviews constructed with a Plugin Forest obtain an average quality rating equal to doctorate-level researchers in the field. Further, the high-quality documents generated by GPT-4 exhibited smaller quality variation compared to human-authored documents. Notably, GPT-4 using the Plugin Forest technique generated literature reviews with zero hallucinated citations. We find that GPT-4 quality is sensitive to paper topic, with lower performance for lesser-studied topics. Further, graders were reliably unable to determine when a document was authored by GPT-4, even when the grader had a graduate education in the field. Graders did not identify GPT-4 authored material as written by a doctorate, however, indicating a notable difference in style even while quality remains matched. We conclude that authoring research using GPT-4 and the Plugin Forest reliably enhances researcher productivity with no loss in quality for a variety of tasks.

Authors: John, Josh

Keywords:

- GPT-4
- ChatGPT
- Economic Literature Review
- AI in Academic Research
- AI Augmentation
- Multimodal Models
- Research Productivity
- Prompt Engineering
- Tree of Thoughts
- In-Context Learning
- Mixture of experts

JEL Codes:

- C88 - Other Computer Software
- A11 - Role of Economics; Role of Economists; Market for Economists
- C80 - General (Data Collection and Data Estimation Methodology; Computer Programs)
- B41 - Economic Methodology
- I23 - Higher Education and Research Institutions
- A10, B41, J23, O3 [used by Korinek]

Background and Introduction

Economists are told that generative artificial intelligence tools including ChatGPT are on the one hand increasingly prolific and useful tool, while on the other hand there are critical voices calling such tools risky on quality or even legally problematic[Bilal, Buchanan et al]. There is a lack of empirical work on productivity and output quality, a lack of standard usage guidance, and a mixing up of large language models, ChatGPT, and other tools of artificial intelligence that have distinct usage patterns and productivity implications.

This paper is the first to provide quantitative data calibrating GPT-4 against doctorate-produced research. We provide a level of utility and recognize a level of nuance by task that is absent in many discussions of research productivity by looking at a variety of topics within the field of economics. This paper describes a new best practice and achieves results that represent the cutting edge of generative artificial intelligence through the use of GPT-4 with a novel prompt engineering technique called a Plugin Forest.

We find that GPT-4 with a Plugin Forest produces literature review summaries of comparable point-estimated quality to a doctorate-level researcher while retaining lower variation in quality compared to a human author. Of note, zero hallucinated citations were generated following this process, in contrast to results frequently attributed to ChatGPT, but in fact generally attributable to GPT-3.5, a substantially inferior and architecturally distinct model compared to GPT-4.

We describe nuances with these results, such as weaker performance for lesser-published topics, we describe novel use cases for GPT-4 that add to known large language model use cases for research, and we make use of GPT-4 plugins with systematic and reprodicible plugin selection, a topic for which there is currently no published empirical work. We conclude with a discussion on expected future productivity trends and open research areas, providing an evidence-based case that productivity over time in the space of generative artificial intelligence is expected to grow at a modest pace.

- TODO: novel tasks include reading and generating images like diagrams, executing, interpreting, and generating code, and translating documents between technical formats like latex, html, and markdown, and knowledge retrieval
  - GPT-4 provides unique productivity opportunities for researchers due to its ability to generate, improve, and execute code for data analysis and both read and create figures and tables. These tasks are not possible for pure large language models.
- TODO: more discussion on what a plugin forest is and why it's awesome
  - more than half a dozen best practices built in, such as tree of thoughts, roles, mixture of experts, and chain of thought prompting, plus making systematic and reprodcible use of gpt-4 plugins
- we know prompt strategy drives productivity, but research on GPT-4 that does exist fails to incorporate best-practice prompt techniques and

leverage GPT-4 capabilities like plugins. Plugins importantly provide access to academic papers, and we know from research like "textbooks are all you need" that access to academic material is a very important driver of producing high-quality academic-level results.

- GPT-4 performance has notably varied over time. Chen et al notes significant performance decreases for some tasks, and this observation has been independently replicated by the ChatBot Arena Leaderboard Project. Importantly, this model shift took place during the observation period for the present study, and as a result we report a coefficient for performance shift relevant to our particular task in economic research.
  - Our observation period was May-July and the ChatGPT UI at that time used the March GPT-4 API, then switched to the June GPT-4 api during the observation period.

Methodology

Overview

This study adopts a comparative approach to evaluate the efficacy of GPT-4 as a tool for conducting economic literature reviews, juxtaposed against human researchers with varying academic credentials in economics. The methodology encompasses participant engagement, GPT-4 interaction, review process, and statistical analysis.

Participant Recruitment and Task Distribution

Selection Criteria

- Participants were selected based on their academic background in economics, comprising one bachelor's degree holder, one master's degree holder, and two doctoral degree holders.

Task Description

- Each participant was assigned the task of drafting a single-page summary literature review. The scope and format of these reviews were intended to parallel the background section of a standard scholarly article.

Integration of GPT-4

ChatGPT Interface Use

- The principal investigator engaged with GPT-4 using the ChatGPT web interface to generate additional literature reviews corresponding to the research questions assigned to human participants.

Application of Plugin Forest Technique

- A novel 'Plugin Forest' approach was employed, which entailed the configuration of multiple plugin collections and the execution of a 'tree of thoughts' prompting strategy. The resultant data from each collection was subsequently synthesized.

Blind Review and Scoring Mechanism

Blind Assessment

- To ensure impartiality, participants were unaware of the authors of the papers they reviewed, except for their own contributions. The principal investigator, however, was privy to the authorship details of all submissions.

Scoring Parameters

- Participants rated the papers on a scale of 1 to 10, focusing on two aspects: perceived quality of the content and likelihood of the paper being generated by GPT-4. This dual-scale assessment aimed to minimize any bias in quality perception influenced by the assumed origin of the paper.

Editing Process and Selection for Analysis

Randomized Paper Selection

- A random selection algorithm was used to pick one GPT-4 authored review per research question for further analysis. These selected pieces underwent editing by the principal investigator.

Editing Workflow Evaluation

- The study considered both the unedited (naive) AI-authored submissions and the potential benefits of AI-assisted editing in refining the final output.

Analytical Approach

Statistical Comparison

- The analysis involved a statistical comparison of the ratings assigned by participants, accounting for potential biases, including those inherent to individual participants and the principal investigator.

Methodological Integrity

- The study adhered to rigorous standards of academic research, ensuring the reliability and validity of the findings through methodical data collection and analysis.

Ethical Compliance

Informed Consent

- All participants were provided with a comprehensive informed consent form, elucidating the academic and commercial use of their input in an anonymized format.

Voluntariness of Participation

- Participation was voluntary, with the provision for participants to withdraw at any stage of the study. However, data collected prior to withdrawal were retained for analysis as per the study's protocol.

The methodology of this research was meticulously designed to ensure a thorough and unbiased evaluation of GPT-4's capabilities in comparison with human expertise in economic research writing. The study's structured approach aimed at delivering insightful conclusions on the role and effectiveness of AI in academic literature review processes.

Results

Case Study Results

Generalized Case Study

- People couldn't tell it was GPT-4
- Even field doctorates couldn't tell

Regression Results

- 3 model table
- Regression of assessed education level on GPT Authorship

Conclusion

The findings of this study strongly confirm the utility of GPT-4 and the Plugin Forest for use in general research and publication-level work in the field of economics. The demonstrated ability of GPT-4 to generate literature reviews of quality comparable to that of doctorate-level researchers, while demonstrating reduced variation in quality and zero hallucinated citations removes the controversy around such tools and pushes the academy to better understand the value of these tools and the concrete techniques by which they are optimally utilized. This study not only sheds light on the efficacy of GPT-4 but also pioneers the systematic and reproducible use of its plugins, setting clear best practices for the state of usage today, and also marking clear opportunities for future research.

Key Takeaways

1. **Equivalency in Quality**: GPT-4's output, calibrated with the Plugin Forest technique, matches the quality of doctorate-level research in economics, providing a reliable approach for literature review composition.
2. **Consistency and Reliability**: The consistency of GPT-4's performance, exhibiting less variance compared to human-authored documents, underscores its reliability as a research tool.
3. **Style Distinction**: Graders, including those with graduate education, could not consistently distinguish between GPT-4 and human-authored documents, indicating the model's advanced capability. However, the distinct style of GPT-4's outputs, while maintaining quality, suggests room for stylistic refinement to better align with traditional academic writing.
4. **Topic Sensitivity and Limitations**: The study acknowledges GPT-4's lower performance in lesser-studied topics, indicating a need for further development in handling niche academic subjects.

Related Trends

This paper discussed quality and performance for research tasks using GPT-4 and the Plugin Forest. Two related trends of note include tehnical multimodal improvements and growth in the GPT-4 plugin store.

Technical Improvements in Multimodal Generative AI Analysis of GPT-4 over a generic reference to ChatGPT moves the academy forward along the axis of model architecture from the analysis of large language models to an analysis of multimodal models. In April of 2023, Sam Altman, the CEO of Open AI, commented that "we're at the end of the era where it's gonna be these giant models, and we'll make them better in other ways," in an interview with TechCrunch.

Less than a week later, an impressive an open-source multimodal model was released called Llava. Later in 2023, Google unveiled Gemini, a competitor in the native-multimodal space. See Yin et al for a survey of multimodal developments.

GPT-3 was released in 2020 and GPT-3.5 was released in 2022. The halving between the time to release GPT-3, GPT-3.5, and GPT-4 might seem like cause for concern on the velocity of generative artificial intelligence accelaration, but fundamentally GPT-3.5 follows the same language model architecture of GPT-3.

Mamba and Heyena Hierarchy are two examples of fundamentally different model architectures, although the practical improvement from these architectures is yet to be seen.

In general, a trend toward multimodal models with improved technical capabilities is clear, but practical improvements in year-over-year productivity remain nonuniformly increasing and modest. A clear empirical comparison is found

between GPT-3.5 and GPT-4. Zheng et al provide a useful and robust approach to comparative quality analysis with a crowdsourced Chatbot Arena. Table 3 below reproduces the top 3 leaderboard entries from the arena as observed on January 5, 2024, along with the release date for each model.

Table 3

Model	Arena Elo rating	MT-bench (score)	Release Date
GPT-4-Turbo	1243	9.32	11/6/2023
GPT-4-0314	1192	8.96	3/14/2023
GPT-4-0613	1158	9.18	6/13/2023

The arena data show the dominance of GPT-4 over more than fifty models compared. Using the ratings provided by users, we see an improvement in the Arena Elo Rating for the top-performing model of about 4.02 percent over the course of about eight months. If we use the MT benchmark scores, we observe a slightly higher 4.28 percent increase over the same period, or an approximate improvement of 6.42 percent annualized. This seems decidedly modest compared to much of the popular rhetoric or academic writing on runaway technology or the dangers of fast takeoff AI.

GPT-4 Plugin Growth Over Time ChatGPT Plugins were opened to the public on March 23, 2023. There was an initial catalogue of 13 plugins. There were 83 plugins available in May, and at the time of writing on January 3, 2024, there are now 1039 plugins listed in the directory. Figure 3 illustrates the growth of these plugins over time.

Over this ten-month span, we see an approximate eighty-fold increase in the count of available plugins, or about one hundred new plugins each month. We expect this upward trend to correlate with improved access to academic literature for GPT-4, although the the magnitude of research quality and productivity improvements that may result are an open question.

Future Research

The study suggests several avenues for future research.

1. Larger Sample Size A larger sample size with better representation of more varied graduate degree holders and more varied topics under assessment, would provide a more comprehensive understanding of GPT-4's capabilities across different educational backgrounds.

2. Beyond Literature Reviews Exploring GPT-4's utility in other sections of academic papers, such as methodology and data analysis, would further expand its applicability.

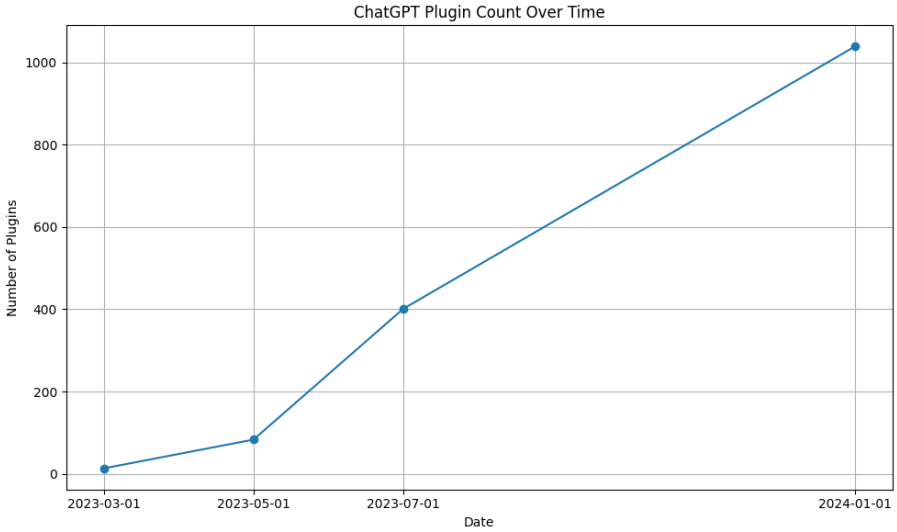

Figure 1: ChatGPT Plugins Over Time

3. Technical Model and Prompt Improvements GPT-4 Turbo is a new model of GPT-4 with a higher token limit and improved benchmark results. Token limit expansion will generally facilitate multishot prompting, which is known to independently improve results. Periodic reporting on state-of-the-art model performance will significantly improve on the rough estimation of productivity increases over time that were presented in the conclusion of this paper.

Prompt engineering represents an opportunity for further independent technical improvement. This paper took an AI-driven approach to role selection in the mixture of experts for the Plugin Forest, but perhaps specific roles can be identified that perform better on average compared to AI-selected roles.

4. Journal-Targeted Models Textbooks Are All You Need showed the power of manipulating the training data on a model, and this seems a likely route to solve for the weak academic style adoption demonstrated in this paper. Compared to training on textbooks, publication chances might be further improved by directly training on material for a given journal. Predicting publication odds and constructing research techniques that optimize directly on that outcome would add significantly to the explanatory approach on quality taken in this paper.

5. Tool Selection, Reputation, Price Signals, and the GPT-4 Store This paper used an AI-driven plugin selection approach. This approach was selected in part due to the absence of alternatives. Identification of product quality is often made on the basis of rating aggregation, price, and other market

signals. To date, there is no central repository of plugin ratings which would make this approach feasible.

In 2024, Open AI has announced the release of a GPT-4 Store. This store would operate as a marketplace for some kinds of tools, potentially including plugins. A useful contribution to the literature would be to identify generative artificial intelligence plugins and tools that lead the market based on these or other economic signals.

6. Ethical Considerations Addressing open ethical considerations remains an important concern for generative AI.

7. AI-Driven Research This study compared the quality of human-authored summary literature reviews to those produced by GPT-4 with a Plugin Forest. The drafts produced by GPT-4 are high in quality. This is more important than a mere tool addition to the toolset of the researcher. We should take a step back and begin to consider more strongly the broader concept of AI-driven research.

This study directly justifies a systematic literature review implemented by artificial intelligence, but such a process is still fundamentally supervised and therefore constrained. A truly AI-driven approach would begin by asking a multimodal model, or other advanced model, about which research topics are most valuable in the first place.

Appendix A: Questionnaire

Title:

GPT-4 Survey

Initial Message to Participants:

The purpose of this survey is to determine the academic writing ability of GPT-4 compared to humans. This survey will present 18 articles, each about a page in length, written by a mix of humans at a variety of educational levels and GPT-4. Please read each article and rate the quality. This survey is expected to take 30-90 minutes, varying mainly by reading speed.

Question 1

On a scale of 1-10, with one being the least attention and 10 being the most attention, please indicate how much attention you applied while completing this study

Question 2

Enter your email or Participant ID to receive a participation reward.

Question 3

What is your highest level of education?

Responses:

1. High School or Less
2. Some College
3. An Undergraduate Degree
4. A Graduate Degree
5. A Ph.D.

Question 4

Do you have a postsecondary degree in Economics? (Y/N)

Question 5-23

This question is repeated for a DOCUMENT_ID ranging from 1 to 18

For Document ID [DOCUMENT_ID], please answer the following three questions using a comma-separated format.

1. What education level does the writer appear to have? Use ""u"" for undergraduate or lower, ""m"" for the master's level, or ""p"" for Ph.D. or higher.

2. Rate the article quality on a scale from 1-10.
3. Rate the likelihood that the article is written by GPT-4 on a scale from 1-10.

An example answer would be ““u,1,1””

Appendix B: Documents and ChatGPT Source Threads

1. Review of Macroeconomic Indicators
    1. Plugin Forest Identification
    2. Journalist Persona
    3. Professor Persona
    4. Data Scientist Persona
    5. Ph.D. Student in Economics Persona
    6. Synthesis Draft
2. Review of Gender Effects in the Post-Pandemic Labor Market
    1. Plugin Forest Identification
    2. Policy Analyst Persona
    3. Professor Persona
    4. Author Persona
    5. Synthesis Draft
3. Review of LLM Best Practices*
    1. Plugin Forest Identification
    2. Journalist Persona
    3. Professor Persona
    4. Data Scientist Persona
    5. Policy Analyst Persona
    6. Synthesis Draft*
4. Comparative Review of the Austrian and Neoclassical Schools*
    1. Plugin Forest Identification*
    2. Policy Analyst Persona*
    3. Professor Persona*
    4. Education Researcher Persona, Part 1: As Assigned*
    5. Education Researcher Persona, Part 2: With WebPilot*
    6. Education Researcher Persona, Part 3: Browser Exploration*
    7. Education Researcher Persona, Part 4: With BrowerOp Plugin*
    8. Synthesis Draft*

*Generated using GPT-4 with the July 20th, 2023 version of ChatGPT. All others use the version from May 24th, 2023.