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 doctor, 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 evidence comparing GPT-4 research quality in the field of economics to researchers in the area. 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 reliably produces literature review summaries of superior or comparable quality to doctors in the field. 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.

These findings are similar to results found in other fields including medicine. Taloni et al found that humans performed better than GPT-3.5 in the self-assessment program of American Academy of Ophthalmology, but GPT-4 performed better than humans. Brin et al reported similar performance in an analysis of soft skill questions from the United States Medical Licensing Examination, and Mustafa Eray Kılıç found similar results in Turkish Medical Specialization Exam performance.

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.

Beyond Large Language Model Tasks

Accessing GPT-4 through ChatGPT includes native machine vision support, enabling tasks like optical character recognition, diagram reading, and image generation. GPT-4 includes a code compiler which allows native execution of code, massively improving mathematical correctness over ordinary large language

models.

ChatGPT enables a user to upload data files for in-browser analysis. Combined with the code compiler, a data analyst or researcher can often complete an exploratory data analysis entirely in the browser.

GPT-4 includes ChatGPT plugin access, which enables ChatGPT to overcome the training knowledge cutoff. Various plugins enable general internet access, with some limitations, and there are also plugins connecting ChatGPT to a substantial and growing body of academic literature. Some data sets of interest are also communicated to ChatGPT through a plugin. The World Bank, for example, has a ChatGPT plugin.

In addition to many new technical capabilities, GPT-4 performs notably better on many tasks that large language models are already capable of executing, such as code generation.

Methodology

This study involves comparative authoring of one-page summary literature reviews across four topics by human authors with varied educational attainment and also GPT-4 using a Plugin Forest prompt strategy. Some documents were also obtained through online services that offer undergraduate paper writing as a service.

Once these eighteen summary literature reviews were obtained, a sample of adults from the United States were questioned about their quality. Participants were also asked about the assessed education level of the document author and the likelihood of GPT-4 authorship. Participants were also asked about their own education and whether they have a degree in the field of economics. Appendix A provides further questionnaire details. The documents were normalized on font style and citation format prior to presentation.

Thirty participants were recruited across two channels[1]. Participants were provided an incentive of fifteen dollars each on average. Twenty-seven were recruited using the Prolific recruitment platform and three others were recruited by word-of-mouth. The participant pool is not balanced to the census and substantially overweights graduate degree holders, as the objective of the study is mainly to calibrate GPT-4 against graduate degree holders and researchers and not to draw a comparison with the typical American adult.

Survey responses were partially blind in that three participants also contributed content for assessment. Such respondents had knowledge over their own authorship, but not the authorship of other documents.

GPT-4 authored a paper in each of four given topics, while other authors contributed any number of summary literature reviews up to four. The list of

[1]Anonymized study data and analytical code are available at https://osf.io/vxgy4?view_only=ca664b55eae243339f13c59cc5dea3ec.

topics include:

1. Is modern Austrian economics distinct from neoclassical economics?
2. What key factors indicate the overall health of the macroeconomy? Include both general concepts and also specific public measures.
3. What is the impact of remote work on the gender wage gap and career progression in the post-pandemic labor market?
4. Suggest best practices for literature search with and without a large language model (LLM). Given the benefits and problems of such a process, do you expect the gains to researcher productivity from LLM augmentation to be large, small, or negative?

These topics were respectively selected to test a lesser-published topic in economics, a highly-published topic, a recent topic, and a topic in the technical domain of ChatGPT.

The Plugin Forest

GPT-4 authored summary literature reviews under the supervision of the principal investigator following a novel prompt engineering strategy called the Plugin Forest. The Plugin Forest is designed specifically as a best practice for use with ChatGPT plugins and it incorporates more than half a dozen evidence-based techniques for improving generative artificial intelligence task performance.

The name most directly refers to a tree of thoughts that leverages plugins. Over multiple rounds of prompting and across multiple distinct threads, the tree of thoughts is created and resolved by GPT-4 acting under a variety of roles that it deems appropriate to the task. This implements a mixture of experts in which each expert is initially ignorant of their peers. Later, particular expert results are synthesized into a consensus result.

To initialize the tree of thoughts, GPT-4 is told to act as a researcher with formal training in economics. As a researcher, GPT-4 is then asked to identify which plugins should be used to assist in authoring a journal-quality academic paper in the field of economics.

At document creation time, many potentially appropriate plugins existed in the ChatGPT plugin directory, but ChatGPT itself only allowed access to three plugins concurrently. Fortunately, the directory included exactly three so-called finder plugins. These finder plugins are able to search across the directory and suggest other plugins for use depending on a user's use case.

In the first round, GPT-4 uses the researcher role and the finder plugin to list nine plugins. That is, each active plugin may suggest three others for use. The plugins suggested often yield duplicates or hallucinations, but eventually a unique list of plugins between three and nine is identified. Given this list of plugins, GPT-4 is asked to identify kinds of people that it would like to discuss research with, and it is asked to assign those people three plugins from the list of plugins to be used. A response might be that GPT-4, as a researcher

in economics, would like to discuss research with a seasoned economic policy analyst or a distinguished professor of economics from a top university.

For each role and plugin set suggested by GPT-4 in the first round, a new ChatGPT thread is created where GPT-4 takes on the specified role and plugin set. They are then asked to write on whatever the given topic for the tree is. Once all personas have created a draft literature review, a final round of prompting is completed in a synthesis thread. These threads were transfer into PDF files and uploaded to the Open Science Foundation for later anonymized presentation to raters.

Performance Variation by Model Version

GPT-4 is a major model version with a number of subversions that have been deployed over time. The document creation period the GPT-authored documents used in this study ran across July 2023, and a new model version was deployed in the midst of authoring. GPT was used to construct four documents. As an accidental observation, two of the four GPT-authored summary literature reviews were constructed using a May version of the ChatGPT interface, which uses GPT-4 subversion GPT-4-0314, and two other documents were created using the July 20th version of the ChatGPT interface, which uses the GPT-4 subversion GPT-4-0613. Appendix B clarifies the specific mapping with links to the original and public ChatGPT threads.

Chen et al notes significant performance decreases for some tasks in GPT-4-0613 compared to its predecessor. To partial out this effect, the current paper calculates and reports on a regression coefficient for a dummy variable `is_march_gpt_model`, which takes a positive value when a document under assessment was generated by subversion GPT-4-0314.

Results

We begin with a case study discussion by examining results for a particular document, then proceed into descriptive statistics in order to check generalizability of case study findings. Regression analysis using ordinary least squares (OLS) and curvilinear regression is reported in the final section on results.

Case Study Results

Document 8 was written by GPT-4 on the topic of the impact of remote work on the gender wage gap and career progression in the post-pandemic labor market, but less than twenty-seven percent of respondents rated this document as a likely creation of GPT-4, where a likely event is scored a six or higher on a ten-point scale.

Document 8 had an average quality rating of 6.8, a median of seven, and an interquartile range from six to eight, indicating a strong majority assessment of high quality writing in the document. While the document was rated high in

quality, it was not calibrated at the doctorate level. On average, respondents assigned an education level between an undergraduate and a master's degree holder to the author of Document 8.

When responses are subselected for respondents with a doctorate-level education, the median rating remains at seven, although the standard deviation on the estimate shrinks. Notably, no respondent that holds a doctoral degree assigned the writing at the undergraduate level. From five responses by a doctorate, three assigned a master's education level to the author of Document 8, and two assigned it a Doctor of Philosophy or higher.

Generalized Case Study

In the case of Document 8, respondents assigned a high quality rating and an education level above the undergraduate level to GPT-4, while expecting the author was not GPT-4. These findings generalize across the other documents authored by GPT-4. In general, respondents assign a low likelihood of GPT-4 authorship to all papers, whether or not GPT-4 is in fact the author.

The average likelihood rating that respondents assigned for GPT-4 authorship across all papers, regardless of the actual author, was 4.39, while the average likelihood rating assigned for papers that were authored by GPT-4 was 4.63 ($p > 0.36$). Still, the output from GPT-4 feels somehow distinct from doctoral writing to the participants. Respondents assigned a Ph.D. education to GPT-4 about thirteen percent of the time, while they assigned such an education to proper Ph.D. authors twenty-two percent of the time.

This difference of approximately ten percent holds when we subselect the responses given by respondents with a doctoral education, although again we see overall ratings more favorable from these respondents. When a doctorate rated a paper authored by a doctorate, they assigned a Ph.D. or higher level of education about thirty-one percent of the time. When a doctorate rated a paper authored by GPT-4, they assigned a Ph.D. or higher level of education about twenty percent of the time.

Interestingly, doctors rated GPT-authored economic papers slightly higher on average (6.65) than those authored by their peers (6.49). While the means were not significantly different, GPT-authored papers had an edge on average and the quality variation among GPT-authored papers was smaller than the variation among doctors as assessed by doctors. This constitutes strong evidence that summary literature reviews produced by with a Plugin Forest and GPT-4 are reliably competitive with journal-quality content.

The edge for GPT-4 predictably grows when doctoral assessments are adjusted to exclude self-ratings. The average quality score given by Ph.D. holders to documents authored by other Ph.D. holders with self-authored documents excluded is 6.15. This lower mean quality assessment was not significantly different from documents constructed through the Plugin Forest in our study ($t = -1.07$, $p =$

0.292).

Table 1 gives an overview of document quality by topic and the kind of author. Ph.D. authors had the highest performance of all author kinds when writing a comparison between Austrian and Neoclassical economics. This result confirms the hypothesis of weaker performance from GPT-4 on relatively niche topics, which would be less available in GPT-4 training data.

Table 1: Performance spread:

Author Group	Best Performing Topic	Best Topic Avg. Quality	Worst Performing Topic	Worst Topic Avg. Quality	Performance Spread
GPT Authors	Macro Health	7.03	Austrian Neoclassical	6.23	0.80
Non-GPT Authors	Austrian Neoclassical	7.13	Macro Health	6.03	1.10
PhD Authors	Austrian Neoclassical	7.13	LLM Best Practices	6.27	0.87

Regression Results

Two ordinary least squares models are investigated to directly explain the relation between GPT-4 authorship, quality, and assessed author education level. The cartesian product of thirty respondents and eighteen documents yields 540 observations for these models, with participant fixed effects included as an independent variable.

Table 2 compares two simple regressions of assessed education level on GPT-4 authorship. Participant effects are excluded for brevity and none of them were significant in the first specification, so the second specification with improved Akaike information criterion (AIC) is preferred. Assessed education for the author of a document has three possible responses and the master's level is the intercept. The intercepts is robustly positive and significant across specifications. Author assessment at the undergraduate or lower level as well as the doctorate level are negatively related to GPT-4 authorship. This model helps to calibrate GPT-4 performance to the level of a master's degree holder.

	Model 1 (Participant FE)	Model 2 (Without Participant FE)
Intercept (Master's)	0.2587**	0.2623***
	(0.1048)	(0.0307)

	Model 1 (Participant FE)	Model 2 (Without Participant FE)
Author Assessed Undergraduate or Less	-0.0863	-0.0845
	(0.0562)	(0.0535)
Author Assessed Ph.D. or Higher	-0.0547	-0.0526
	(0.0433)	(0.0399)
R-squared	0.0057	0.0055
R-squared Adj.	-0.0550	0.0018
AIC	645.48	587.56

Standard errors in parentheses.

- p<.1, ** p<.05, ***p<.01

TODO:

- 3 model table

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.

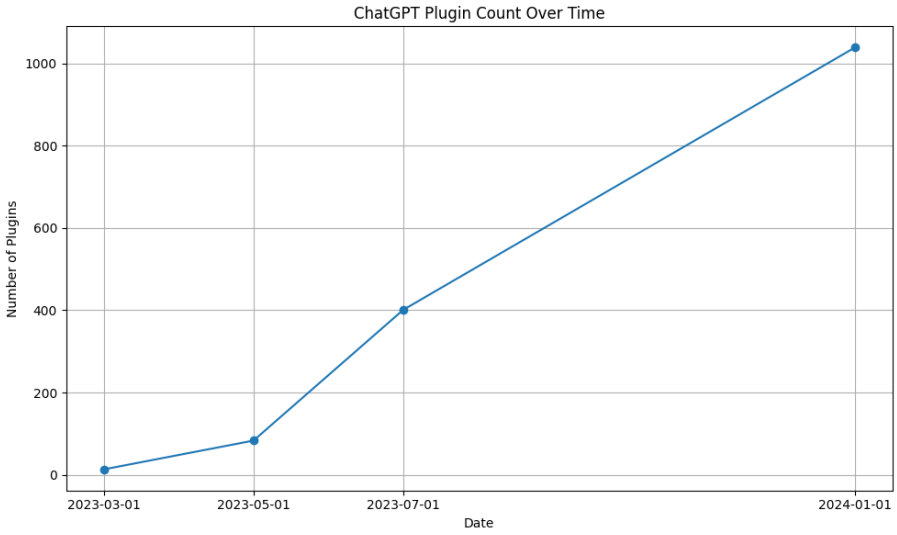

Figure 1: ChatGPT Plugins Over Time

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.

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.