NBER WORKING PAPER SERIES

GENERATIVE AI FOR ECONOMIC RESEARCH:
LLMS LEARN TO COLLABORATE AND REASON

Anton Korinek

Working Paper 33198
http://www.nber.org/papers/w33198

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2024

I would like to thank Hemanth Asirvatham, Paul Bousquet, Kevin Bryan, Alan Chan, and Sam Manning for helpful comments on this paper, and David Romer for wisely asking me to commit to producing regular updates of this rapidly evolving material when publishing my first paper on the topic in the Dec. 2023 issue of the JEL. I have learned so much from working on these updates! All remaining errors are my own. I gratefully acknowledge financial support from the Complexity Science Hub Vienna, where part of this research was conducted. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Generative AI for Economic Research: LLMs Learn to Collaborate and Reason
Anton Korinek
NBER Working Paper No. 33198
November 2024
JEL No. A10, B4, C88, O33

ABSTRACT

Large language models (LLMs) have seen remarkable progress in speed, cost efficiency, accuracy, and the capacity to process larger amounts of text over the past year. This article is a practical guide to update economists on how to use these advancements in their research. The main innovations covered are (i) new reasoning capabilities, (ii) novel workspaces for interactive LLM collaboration such as Claude's Artifacts, ChatGPT's Canvas or Microsoft's Copilot, and (iii) recent improvements in LLM-powered internet search. Incorporating these capabilities in their work allows economists to achieve significant productivity gains. Additionally, I highlight new use cases in promoting research, such as automatically generated blog posts, presentation slides and interviews as well as podcasts via Google's NotebookLM.

Anton Korinek
Department of Economics
University of Virginia
Monroe Hall 246
248 McCormick Rd
Charlottesville, VA 22904
and NBER
anton@korinek.com

Online resources for this paper are available at https://www.GenAIforEcon.org

Reader's guide:

This paper provides an overview of the advances in generative AI over the past year (as of November 2024). It is all new but part of a series of papers exploring the use of generative AI for economic research. The first paper of the series appeared as NBER Working Paper 30957 (Korinek, Feb. 2023b) and was published as "Generative AI for Economic Research: Use Cases and Implications for Economists," *Journal of Economic Literature* 61(4), pp. 1281-1317, (Korinek, Dec. 2023a). New readers may want to also explore the original *2023 JEL* paper for a broader introduction to LLMs in economics.

The online resources associated with this paper are available on the website `https://www.GenAIforEcon.org`. They provide instructions for how to get started with Generative AI in economic research, allow users to sign up for bi-annual updates to this guide, and contain further resources.

Both expert users and newcomers to generative AI may find it most useful is to browse through the use cases documented below and try them out in their own research.

1 Introduction

Large language models (LLMs) have experienced remarkable progress over the past year, characterized by significant gains in speed, cost efficiency, accuracy, and the capacity to process larger amounts of text. These advancements have not only enhanced existing capabilities but have also enabled entirely new ways of interacting with LLMs, demonstrating how substantial quantitative improvements can lead to paradigm shifts in functionality.

The paper starts with an updated overview of the landscape of LLMs, highlighting the rapid evolution and current state of leading models and their applications. All frontier AI labs have released new models in recent months. OpenAI still ranks first on a range of LLM benchmarks, with an updated version of its GPT-4o model. Google DeepMind has released a significantly updated version of Gemini 1.5 Pro 002 with a 2m token context window, making it able to process about 3000 pages of text simultaneously. Elon Musk's xAI has shot to the #3 spot and is tightly integrated into the X (formerly Twitter) ecosystem. Claude 3.5 Sonnet excels at writing-related tasks. But there are now also excellent open-source LLMs from Meta and Alibaba that are close in capabilities to the models of the other four labs.

A key focus of this update are the new access modes enabled by the cumulative performance gains over the past year. Workspaces for interactive collaboration, such as Anthropic's Claude Artifacts and OpenAI's ChatGPT Canvas, Microsoft Copilot or Cursor, are changing how we interact with LLMs. They create an environment where users can iteratively develop and refine ideas, shifting away from static chat-style interactions towards more dynamic, document-oriented collaboration. They allow users to work in tandem with LLMs, offering real-time feedback and allowing for iterative editing. Another example is a new generation of real-time voice assistants that can also assist with research tasks.

A second recent breakthrough are LLM-based reasoning capabilities, exemplified by OpenAI's o1 series. I describe why traditional LLMs were not very good at reasoning and how work on a new generation of reasoning models is helping to overcome these barriers, enabling AI models to engage in multi-step problem-solving and logical deduction. This advancement opens new avenues for LLM use in economic research.

Finally, LLM-powered search, newly integrated into ChatGPT in November 2024 and also offered by Google Gemini and startups like Perplexity, is starting to become a useful tool to provide up-to-date answers to questions that are grounded in facts found on the internet, together with the requisite citations—a crucial capability for researchers.

The paper also describes technical advances in the space of structured outputs and prompt caching as well as practical considerations like declaring LLM use and watermarking.

In addition to describing these new developments, I also cover several other novel examples and use cases for LLMs in the subsections below, many of which center around new ways of promoting research outputs—the current generation of LLMs is highly capable of processing the main insights of research papers and accurately translating and conveying them into a wide range of formats:

- Advanced mathematical derivations using o1-preview
- Sophisticated coding tasks using o1-preview
- Creating a 'deep dive' podcast of your research paper
- AI-powered search
- Creating presentation slides
- Drafting blog posts
- Conducting interviews

2 Advances at the Frontier of LLMs

Table 1 provides an overview of the top proprietary and open-source LLM providers as of November 21st, 2024. The table is ranked by the score of each provider's leading models in the LMSYS leaderboard (column 4), which pits randomly-selected pairs of LLMs against each other and employs user ratings to compile an Elo-like score for each model.[1][2] Columns 5 and 6 of the table list how many tokens (or syllables of text)

[1]The Elo-system was designed by the physicist Arpad Elo to rank chess players by their relative skills. It is designed so that a score difference of D points between two players (or LLMs) corresponds to the higher-ranked one having a probability of $1/_{1+10^{D/400}}$ of winning in a direct match-up.

[2]Like all ranking systems that condense the capabilities of candidates who differ across many dimensions into a single dimension, the LMSYS score offers only a partial and imperfect snapshot of

AI Lab	Best Model	Released	LMSYS	Tokens	Data Cutoff	URL
OpenAI	gpt-4o-latest	Nov 2024	1360	128k	Oct 2023	chat.com*
GoogleDM	Gemini 1.5 Exp	Nov 2024	1343	2m	Nov 2023	gemini.google*
xAI	Grok-2	Aug 2024	1290	128k	Mar 2024	x.ai / x.com*
Anthropic	Claude 3.5 Opus	Oct 2024	1286	200k	Apr 2024	claude.ai
Meta	Llama 3.1-405b	Jul 2024	1267	128k	Dec 2023	OS / meta.ai
Alibaba	Qwen 2.5-72b	Sep 2024	1263	128k	Sep 2024	OS (GitHub)

Table 1: Overview of top proprietary and open-source LLM providers according to their best model score in the LMSYS leaderboard.
Source: `https://lmarena.ai/?leaderboard`. See Chiang et al. (2024). Last accessed on Nov. 21st, 2024.
* denotes chatbots that can also access real-time information on the internet

the models can process simultaneously, and the date on which their training data cuts off. Models generally do not have knowledge of facts that occurred past this date, except if they have the capacity to access the internet. The last column lists the URLs under which the models can be accessed. The designation "OS" reflects that the model is available on an open-source basis, i.e., that it can be freely downloaded, run, and modified by researchers.

Several observations stand out from the table:

1. The field is moving fast—all six of the listed models have been released or updated in the past four months. In fact, older models quickly fall in the rankings. For example, if OpenAI had not released any model updates since April 2024, it would currently rank at the bottom of Table 1.

2. OpenAI continues to be the clear leader in the space with the latest update to its GPT-4o model.

3. The gap between the LMSYS scores of the top models is, however, relatively small. For example, using the Elo formula from footnote 1, OpenAI's GPT-4o would win against the next-ranked Google DeepMind Gemini 1.5 Pro in 52.5% of match-ups – hardly a decisive victory. In the words of Microsoft CEO Satya Nadella, LLMs are becoming "more of a commodity."

4. The open-source models by Meta and Alibaba, listed in the bottom two rows of the table, have caught up and are now close to the frontier – a very different

LLM capabilities. I chose to use it for the overview table here because it has almost universal coverage of LLMs, it is updated in close to real-time, and it aggregates many different types of use cases when evaluating models. The LMSYS score is also highly correlated with other benchmarks of general LLM performance such as the MMLU.

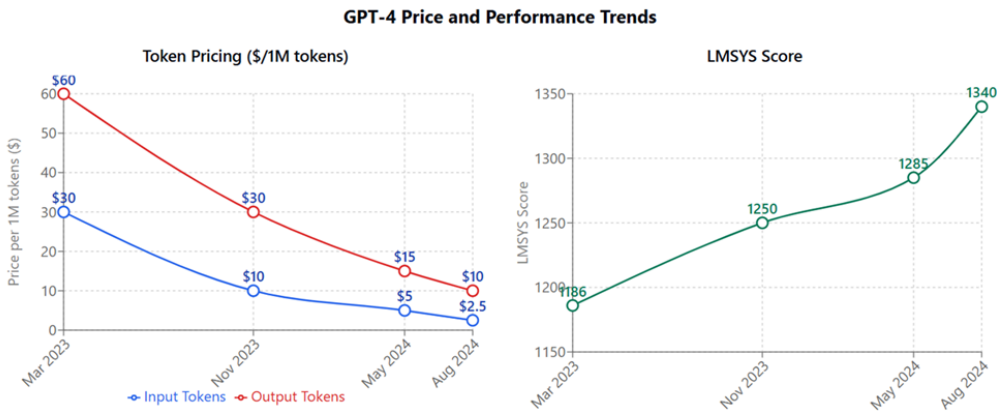


Figure 1: Decline in operating costs and quality improvement of GPT-4 models
Source: compiled by author, last updated Oct. 2024.

situation from a year ago, when open-source models were significantly behind proprietary models.

5. Chinese-made LLMs have ascended particularly rapidly, as reflected in the last row. Since LMSYS rankings are based on mostly Western user preferences, they may in fact understate the capabilities of Qwen 2.5. What is remarkable is that the model ranks so close to the best Llama 3.1 model, even though its parameter count (72bn) is just a fraction of the latter's (405bn).

Speed of Progress To provide data on the speed of progress, I list a few quantitative indicators from OpenAI's series of GPT-4 models as an example. Since the initial release of GPT-4 in March 2023—less than two years ago—the models' context window size has increased 16-fold, allowing it to process far more content at once, the quality of the model's responses has significantly improved (the current LMSYS score of the original GPT-4 is only 1186), and the speed of output generation has increased 3-fold. Figure 1 illustrates the steep decline in the cost of reading and generating text (input and output tokens) of GPT-4 level models since March 2023—by 92% and 83% respectively—even though their LMSYS score steadily improved. See Ho et al. (2024) for a detailed examination of algorithmic progress in LLMs.

In the following, I describe the leading LLM products of the frontier labs listed in Table 1. Readers who are most interested in the conceptual advances may want to skip to Section 2.1 describing advances in reasoning, followed by a description of novel collaborative access modes for LLMs.

Each of the labs listed in Table 1 offers families of models of different sizes that reflect different trade-offs between model performance, speed, and cost. Larger models are more "intelligent" and generally offer better performance and greater capabilities, but they also require more computational resources and take longer to process requests, making them more expensive. Smaller models, on the other hand, are faster and more cost-effective, but may not provide the same level of quality in their outputs. This allows users to consider their specific needs and budget when choosing the appropriate model size for their applications.

Leading Proprietary Models

The first four labs listed in the table offer proprietary models, which means that their models can only be accessed via the labs' computer servers. They do not share the source code, architecture, and model weights of their LLMs but allow users to access them via chatbots, web-based experimentation platforms, or APIs, subject to the certain conditions and controls.

OpenAI's GPT-4o model, last updated in November 2024, continues to lead the market for LLMs in terms of both general capabilities and popularity. (OpenAI's o1 model, released in September 2024, demonstrates new advances in LLM-based reasoning that are extremely valuable for research, as described in Section 2.1, but less valuable for general use, resulting in a lower LMSYS score than GPT-4o.) GPT-4o is an evolution of the original GPT-4 model of March 2023 that is considerably smaller, faster, cheaper, and more capable, as shown in Table 1. The suffix "o" stands for "omni" to reflect that the model can process text, images, and sound. GPT-4o also offers workspace extensions that make it easy to interact collaboratively with the model, including Canvas and Advanced Data Analysis (described in Sections 2.2 and 3.5 below), and the ability to search the web (described in Section 2.3 below). GPT-4o is subject to usage limits in the free version of ChatGPT. The model's smaller sibling. GPT-4o-mini, is faster and 94% cheaper but would still rank in the number 5 spot in Table 1, making it an attractive choice for bulk data processing.

Google DeepMind's Gemini series of LLMs carries the distinction of having a 2m token context window—the longest of all publicly available LLMs, which allows it to simultaneously process a few dozen books or several hundred papers.This offers new use cases—for example, it allows researchers to upload a significant body of their work all at once and process queries based on it, or to simultaneously process videos or large corpora of images. The most powerful version is currently *Gemini 1.5 Exp,* updated November 2024, and is only available to paying subscribers. It also comes with a smaller sibling, *Gemini 1.5 Flash*, which offers greater speeds at lower cost but slightly lower performance. Gemini is also accessible via an eponymous chatbot that can access the internet to include real-time information in its responses and allows users to cross-check results and follow links to its sources.

xAI's Grok-2 is a relative newcomer in the LLM space. xAI was founded by Elon Musk in March 2023, and its Grok-2 model has ascended into the top-3 a bit over a year after the lab's founding, offering state-of-the-art peformance in most tasks. xAI benefits from its close relationship with X, formerly Twitter, which Elon Musk took over in 2022 and uses for training data. This allows Grok-2 to be up-to-date on news.

Moreover, it distinguishes itself by not imposing any limits on user queries, following instructions and generating controversial content that many may consider unethical, reflecting Elon Musk's "free-speech absolutism."

Anthropic's Claude 3.5 Sonnet, by contrast, brands itself as being a helpful, honest, and harmless assistant, employing a process called constitutional AI to train the LLM to follow a set of high-level ethical principles (Bai et al., 2022). Claude is the model I use most for writing as I like its succinct, elegant and insightful writing style. The latest update, released in October 2024, ranks the model in the top spot of several technical benchmarks. Claude 3.5 has a context window of 200k tokens, which makes it able to process about 150,000 words in one go—for example, several academic papers. Anthropic pioneered many LLM applications and access modes, for example the chatbot format before ChatGPT or, more recently, interactive collaboration in workspaces called "Artifacts" (see Section 2.2) and autonomous computer use (see Section 2.2.3). Another recent update, PDF support (beta), allows Claude to visually process PDF documents uploaded in its chat interface or via its API so that it can read figures and graphs in PDFs, which is highly valuable in processing academic papers or other documents that contain visual information such as charts or figures.

Table 2 compares the cost of the models listed above—it has become industry practice for leading labs to offer two main models: a more expensive frontier model and a cheaper model well-suited for bulk data processing. xAI is only offering beta access to its models as of November 2024. OpenAI and Anthropic offer a 50% discount for batch processing that may be executed at a delay when their servers face a lower load; all three labs offer discounts for cached content. For Google DeepMind, the first 50 requests per day for its Pro model and 1500 requests per day for its Flash model are free, and using more than 128k tokens incurs double the cost displayed in the table.[3]

Model (cost per 1M tokens)	**Input Cost**	**Output Cost**
OpenAI GPT-4o	\$2.5	\$10
OpenAI GPT-4o-mini	\$0.15	\$0.60
Google DeepMind Gemini 1.5 Pro	\$1.25	\$5
Google DeepMind Gemini 1.5 Flash	\$0.075	\$0.30
Anthropic Claude 3.5 Sonnet	\$3	\$15
Anthropic Claude 3.5 Haiku	\$1	\$5

Table 2: Price comparison for input and output tokens across leading models.
Source: compiled by author, last updated Oct. 2024.

[3]Up-to-date pricing information for the three labs is available at `https://openai.com/api/pricing/`, `https://ai.google.dev/pricing` and `https://www.anthropic.com/pricing`.

Leading Open-Source Models

The top LLM providers that release their models open source are listed in the last two rows of Table 1. Their models are freely available to download, use, modify, and distribute.[4] This offers several benefits for economic research. Firstly, the transparency of open-source models allows researchers to examine the underlying architecture, enabling them to better understand the model's structure. Secondly, open-source projects allow anybody to innovate upon the model. This can help accelerate the development of LLMs tailored to specific needs. Thirdly, if researchers have access to low-cost computing resources, they can leverage open-source models for their work without incurring financial costs. Fourthly, open-source models that are operated locally offer significant privacy benefits as sensitive data does not need to be channeled over the internet to be processed on the servers of proprietary model providers. Finally, open-source models allow for greater reproducibility, which is helpful for ensuring scientific integrity in research as it enables other researchers to verify and build upon the reported results. These benefits make open-source language models an attractive choice for researchers seeking to harness the power of natural language processing in their work.

From an economic perspective, open-source models are highly beneficial as they freely distribute the economic social surplus created by LLMs and stimulate innovation (Korinek and Vipra, 2025). On the downside, as open-source LLMs become more capable, they also pose growing safety risks (Anderljung et al., 2023).[5]

Meta's LlaMA 3.1 is currently the most powerful series of open source models, which have been downloaded more than 350,000 times so far. The most powerful publicly available model is currently Llama 3.1-405B which features 405B parameters. However, as of November 2024, LlaMA is transitioning to version 3.2, offering multimodal versions with 11B and 90B parameters as well as text-only versions with 3B and 1B parameters, which can be operated on many devices. All available Llama models are also accessible on leading cloud computing platforms, including Microsoft Azure, AWS, and Hugging Face. NVIDIA released a fine-tuned version of Meta's 70B parameter model as Llama-3.1-Nemotron-70B-Instruct, which has obtained an LMSYS score of 1271 despite its smaller size.

Alibaba's Qwen 2.5 (short for Tongyi Qianwen, which translates to "Unified Thousand Questions") has made rapid progress and reached a spot in Table 1 in Sept 2024,

[4] More precisely, the models are "open weights," which means that the weights and software to run inference on the LLM can be freely downloaded but not the training source code and data. Frequently, there are additional restrictions. For example Meta imposes limits on the large-scale commercial use of their models. The AI lab TogetherAI has a project named RedPajama to reproduce and distribute an open source version of the LLaMA dataset.

[5] For example, LlaMA has already allowed researchers to construct adversarial attacks that circumvent the safety restrictions of all the LLMs listed above (Zou et al., 2023). Seger et al. (2023) discuss the pros and cons of open-sourcing LLMs as well as intermediate solutions between proprietary and fully open source models that may be desirable as LLMs become more capable and pose growing safety risks.

even though Alibaba, being located in China, is subject to export controls on cutting-edge GPU chips that are crucial for training LLMs. The Qwen 2.5 series consists of 100 open-source models with parameter sizes ranging from 0.5B to 72B, including multimodal models and excellent LLMs specialized in math and coding that reach state-of-the-art performance.

2.1 Advances in Reasoning

One of the most significant advances in recent months is that LLMs are becoming better at reasoning. Traditional LLMs generate output via token-by-token prediction, as described, e.g., in Section 2 of the originally published version of this paper (Korinek, 2023a). Although this basic architecture has proven surprisingly powerful, it makes it hard for basic LLMs to go back in the text that they have already generated to reason about it and iteratively improve it, as humans do when they write. A good analogy is that token generation by LLMs proceeds like a human's stream-of-consciousness. This makes it easy for such systems to emulate what Kahneman (2011) called system-1 thinking but difficult to perform cognitive tasks that correspond to system-2 thinking and require reasoning. For example, a famous test that poses no problem for fourth-graders but has regularly tripped up even the most advanced LLMs before o1 was the so-called strawberry test: asking an LLM "How many R's are there in strawberry?" typically delivers false responses.[6]

Aware of these limitations, researchers have worked hard on finding ways to enable LLMs to become better at reasoning (see, e.g., the surveys by Huang and Chang, 2023; Plaat et al., 2024). An influential mechanism to obtain better-reasoned results has been chain-of-thought prompting, which instructs LLMs to proceed step-by-step when generating responses to a prompt. This technique has delivered significant performance gains by guiding LLMs to break down complex questions into smaller logical steps that are easier to accomplish—akin to a student who performs better on an exam when asked to report his intermediate steps. Wei et al. (2022) show that chain-of-thought prompting considerably improves LLM performance on a range of arithmetic, commonsense, and symbolic reasoning task. For example, when given a question like "The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?" they show that OpenAI's GPT-3 from 2020 failed ("The answer is 27.") but succeeds when guided to reason through the calculation, producing the answer: "The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9."

[6]One of the reasons for this rather basic failure is that LLMs encode text not in letters, as we do in the English language, but in tokens that correspond to syllables or words and that imply that the spelling is not directly observable for LLMs when they process text. This implies that the model needs to reason about the English spelling corresponding to the underlying tokens. Dell'Acqua et al. (2023) use the term "jagged frontier" to observe that LLMs easily perform some tasks but fail at other tasks that are of seemingly similar difficulty for humans.

When academics reason about a novel research problem, an additional strategy that they employ is to perform a sort of tree search: they generate different hypotheses, evaluate them, pursue the most promising ones and refine them further. Recent advances in LLM-based reasoning attempt to emulate this process. For example, Yao et al. (2024) propose what they call a tree-of-thoughts technique, which extends chain-of-thought prompting by generating multiple intermediate steps or pro-verbial "thoughts" at each stage of the reasoning process. This allows LLMs to explore different paths of reasoning, evaluate their potential, and select the most promising ones to continue—much like, for example, a chess engine evaluating different moves. This approach has shown particular promise in solving complex reasoning tasks that require planning and strategic thinking. In spring 2024, Anthropic introduced a feature that allows Claude to engage in short episodes of reasoning that are hidden from the user behind a message like "Thinking deeply..." or "Ruminating..." before generating output. This led to clear performance gains, but with little fanfare.

OpenAI's o1 series of models, released on September 12, 2024, is the first that is explicilty designed for LLM-based reasoning (OpenAI, 2024).[7] Although an official description of the model's architecture is not publicly available, OpenAI seems to have employed reinforcement learning to hone the model's automated use of the two techniques described in the preceding two paragraphs: it employs a chain-of-thought technique to break down complex steps into simpler ones while also employing a form of tree search to attempt different approaches to solving a problem and to recognize and correct mistakes. Unlike earlier LLMs, o1 models react to prompts by first generating reasoning tokens that are hidden from the user—akin to a simulated inner monologue—as the model "thinks" through the problem at hand. Depending on the question, this may take from a few seconds to several minutes. Once the hidden thinking process is finished, the model generates a response for the user that summarizes the outcome of the reasoning process. This mechanism has enabled o1 to achieve significant gains in reasoning compared to GPT-4o, which itself was a leader in this category before o1's release. However, Mirzadeh et al. (2024) document that even the reasoning capabilities of o1 still have significant shortcomings, for example, that small changes in numbers or the addition of irrelevant information reduce model performance.

The o1 series consists of several models. In Sept. 2024, OpenAI released a preview version of the full model, `o1-preview`, as well as a smaller and more cost-efficient model, `o1-mini`, which is particularly well-suited for math and coding applications and better than `o1-preview` in some of these applications. The full version `o1` is the most powerful of the three and expected later in fall 2024. Figure 2 visualizes the advances compared to GPT-4o on three benchmarks: in the American Invitational Mathematics Examination (AIME), the qualifying exam for the US Mathematical Olympiad, o1 performed at a level corresponding to the top-500 contestants; in the Codeforces programming com-

[7]Originally nick-named "Q*" and later "strawberry," its creators argued that the system's architecture is sufficiently different from its earlier GPT series to merit a new name that is simply an abbreviation of "OpenAI 1."

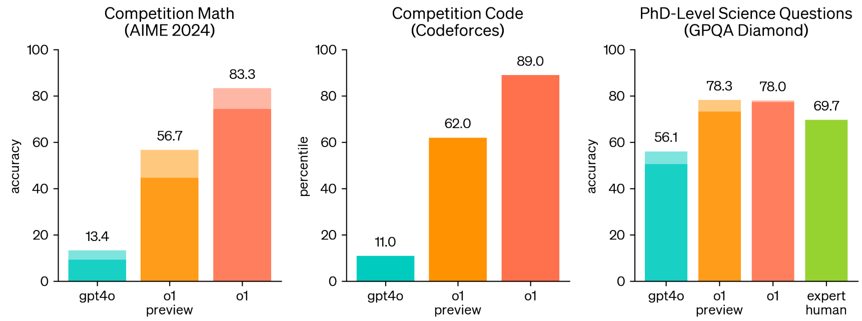


Figure 2: o1's performance on reasoning benchmarks. Source: OpenAI (2024).
Note: Solid bars show one-shot accuracy; shaded regions show performance of a majority vote of 64 responses.

petition, it performed at the 89th percentile of participants; in the hardest (diamond) category of the Graduate-Level Google-Proof Q&A benchmark, which tests for complex reasoning abilities in biology, chemistry, and physics, it outperformed PhD-level human experts.

Interestingly o1 also fundamentally changes the economics of operating frontier AI models: Under the traditional LLM paradigm, the computational cost of training models exceeded the cost of generating output (inference) by orders of magnitude. By contrast, o1 relies heavily on generating reasoning tokens during inference, and—in analogy to the scaling laws for training LLMs discussed in Section 2 of the original article—the model's responses improve predictably in the amount of reasoning tokens employed (OpenAI, 2024). This implies far greater variable costs compared to traditional LLMs. At present, o1 models are only available to OpenAI's paying customers, with limits on the number of queries that can be performed per week.

The following reproduces an interaction of mine with o1-preview for a task that was far beyond the capabilities of earlier LLMs. (Note that the latest version of Anthropic's Claude, released in October 2024, also successfully completed the assignments shown in the following two chats and required less thinking time.)

Chat 1 (o1-preview): The following is a no-arbitrage relationship between an interest rate on new long-term debt in period t, $R_{N,t}$, and expected future short-term rates, R_{t+s}:

$$\frac{1}{R_{N,t}+\delta} = E\sum_{i=0}^{\infty}\left[\frac{(1-\delta)^i}{\prod_{s=0}^{i}(1+R_{t+s})}\right] \tag{1}$$

Can you log-linearize this expression around the steady state value of the interest rates, which we denote by R^*, and which is the same for both interest rates? Denote deviations from steady state with lower-case variables.

Thought for 53 seconds [drop-down menu summarizing reasoning steps]

To log-linearize the given no-arbitrage relationship around the steady state R^*, we'll start by rewriting the original equation and then expand it using first-order approximations.

Original Equation:

$$\frac{1}{R_{N,t}+\delta} = E_t\left[\sum_{i=0}^{\infty}\frac{(1-\delta)^i}{\prod_{s=0}^{i}(1+R_{t+s})}\right]$$

Steady State Analysis:
At steady state, $R_{N,t} = R^*$ and $R_{t+s} = R^*$. The equation simplifies to:

$$\frac{1}{R^*+\delta} = \frac{1}{1+R^*}\cdot\frac{1}{1-k}$$

where $k = \frac{1-\delta}{1+R^*}$.

Log-Linearization Steps:
1. **Left-Hand Side (LHS):** Expand $\frac{1}{R_{N,t}+\delta}$ around R^*:

$$\frac{1}{R_{N,t}+\delta} \approx \frac{1}{R^*+\delta}\left(1-\frac{r_{N,t}}{R^*+\delta}\right)$$

2. **Right-Hand Side (RHS):**
Approximate the denominator and numerator for small deviations:

- **Denominator:**

$$\prod_{s=0}^{i}(1+R_{t+s}) \approx (1+R^*)^{i+1}\exp\left(\frac{1}{1+R^*}\sum_{s=0}^{i}r_{t+s}\right)$$

- **Numerator:**

$$(1-\delta)^i \approx e^{-i\delta}$$

- **Combining Terms:**

$$\frac{(1-\delta)^i}{\prod_{s=0}^{i}(1+R_{t+s})} \approx \frac{k^i}{(1+R^*)}\left(1-\frac{1}{1+R^*}\sum_{s=0}^{i} r_{t+s}\right)$$

3. **Simplify RHS Sum:**

$$E_t\left[\sum_{i=0}^{\infty} k^i\left(1-\frac{1}{1+R^*}\sum_{s=0}^{i} r_{t+s}\right)\right] = \frac{1}{1-k} - \frac{1}{(1+R^*)(1-k)}\sum_{s=0}^{\infty} k^s r_{t+s}$$

4. **Equate LHS and RHS:**

$$\frac{1}{R^*+\delta}\left(1-\frac{r_{N,t}}{R^*+\delta}\right) = \frac{1}{R^*+\delta}\left(1-\frac{1}{1+R^*}\sum_{s=0}^{\infty} k^s r_{t+s}\right)$$

5. **Simplify and Solve for $r_{N,t}$:**

$$r_{N,t} = \frac{R^*+\delta}{1+R^*}\sum_{s=0}^{\infty} k^s r_{t+s}$$

6. **Express $\frac{R^*+\delta}{1+R^*}$ in Terms of k:** Since $1-k = \frac{R^*+\delta}{1+R^*}$, we have:

$$r_{N,t} = (1-k)\sum_{s=0}^{\infty} k^s r_{t+s}$$

Final Log-Linearized Expression:

$$r_{N,t} = (1-k)\sum_{s=0}^{\infty} k^s r_{t+s}$$

This equation shows that the deviation of the long-term interest rate from its steady state is a weighted sum of expected future deviations of the short-term interest rate, where the weights decrease geometrically with factor k.

Answer:
Yes; the log-linearized expression is $r_{N,t} = (1-k)\sum_{s=0}^{\infty} k^s r_{t+s}$ where $k = \frac{1-\delta}{1+R^*}$.

Chat 1 (o1-preview) ctd.: Great. Can you express this as a forward difference equation in the style of $r_{N,t} = coeff1 * r_{N,t+1} + coeff2 * r_t$?

Thought for 22 seconds [drop-down menu summarizing reasoning steps]

Yes, we can express the log-linearized relationship as a forward difference equation of the form:

$$r_{N,t} = k\, r_{N,t+1} + (1-k)\, r_t$$

where $k = \frac{1-\delta}{1+R^*}$.

Derivation
[full derivation omitted for space reasons]

Final Answer
Yes; the log-linearized expression can be written as:

$$r_{N,t} = k\, r_{N,t+1} + (1-k)\, r_t$$

where $k = \frac{1-\delta}{1+R^*}$.

This example illustrates that even o1-preview, which is less capable than the ultimate o1, can perform complex mathematical derivations that can save researchers significant amounts of time. I should note that the prompt in Chat 1 contains an important hint for how to efficiently approach the problem, that the steady state of the two interest rate variables will be the same. Without this hint, the system struggled. Even with the hint, the model occasionally produced mistakes in repeat tests of derivations such as the one requested in Chat 1 but was frequently able to correct mistakes when I pointed them out. Moreover, the derivation steps reported in the example appear at times a bit clumsy—it looks like this LLM is still "in its infancy."

According to Jones (2024), o1 has impressed many scientists with its complex problem-solving performance, which is particularly beneficial in disciplines that rely on mathematical modeling. While o1 can reason through problems step-by-step, they also noted that these chains of thoughts can increase the risk of "hallucinations."

For complex coding tasks, o1 excels at following, generating, and debugging intricate code structures. Its reasoning capabilities enable it to tackle problems that require deeper contextual understanding and strategic thinking than earlier LLMs, making it particularly valuable for academics working on sophisticated coding projects. The following example documents o1-preview's response to a request to code a solution ot the Ramsey growth model, which earlier LLMs could not do (as documented in the December 2023 version of the paper):

Chat 2 (o1-preview): Can you write python code to solve the Ramsey growth model for standard parameter values?

Thought for 34 seconds [drop-down menu summarizing reasoning steps]

Certainly! Below is Python code that solves the Ramsey growth model using standard parameter values. The code sets up the differential equations representing the Ramsey model, computes the steady-state values, and then uses a shooting method to find the optimal initial consumption that leads to convergence to the steady state. It then simulates the model over time and plots the capital and consumption per capita.

```
import numpy as np
from scipy.integrate import solve_ivp
from scipy.optimize import root_scalar
import matplotlib.pyplot as plt

# Parameters
alpha = 0.3   # Output elasticity of capital
rho = 0.04    # Time preference rate
delta = 0.05  # Depreciation rate
theta = 2.0   # Coefficient of relative risk aversion (CRRA)
n = 0.01      # Population growth rate

# Steady-state capital per capita
k_star = (alpha / (rho + delta)) ** (1 / (1 - alpha))
# Steady-state consumption per capita
c_star = k_star ** alpha - (n + delta) * k_star

print(f"Steady-state capital per capita (k*): {k_star:.4f}")
print(f"Steady-state consumption per capita (c*): {c_star:.4f
   }")

# Initial capital per capita (start below steady state)
k0 = 0.5 * k_star

# Time span for the simulation
t_span = [0, 100]  # From t=0 to t=100
t_eval = np.linspace(t_span[0], t_span[1], 1000)  # Time
   points for evaluation

# Differential equations of the Ramsey model
def ramsey(t, y):
    k, c = y
    dkdt = k ** alpha - c - (n + delta) * k
```

```python
    dcdt = (1 / theta) * (alpha * k ** (alpha - 1) - rho - delta) * c
    return [dkdt, dcdt]

# Shooting method to find the optimal initial consumption c0
def shooting_function(c0):
    y0 = [k0, c0]
    sol = solve_ivp(ramsey, t_span, y0, t_eval=t_eval, method='RK45')
    k_T = sol.y[0, -1]
    return k_T - k_star  # We want k(T) to be close to k_star

# Initial guesses for c0
c0_lower = 0.1 * c_star
c0_upper = 2.0 * c_star

# Use root_scalar to find the root of the shooting function
res = root_scalar(shooting_function, bracket=[c0_lower, c0_upper], method='brentq')
if res.converged:
    c0 = res.root
    print(f"Optimal initial consumption per capita (c0): {c0:.4f}")
else:
    raise ValueError("Root-finding did not converge to a solution.")

# Simulate the model with the optimal initial consumption
y0 = [k0, c0]
sol = solve_ivp(ramsey, t_span, y0, t_eval=t_eval, method='RK45')
k = sol.y[0]
c = sol.y[1]
t = sol.t
# Plotting the results
plt.figure(figsize=(12, 5))
# Plot for capital per capita over time
plt.subplot(1, 2, 1)
plt.plot(t, k, label='Capital per capita (k)')
plt.axhline(y=k_star, color='r', linestyle='--', label='Steady-state k*')
plt.xlabel('Time')
plt.ylabel('Capital per capita (k)')
plt.title('Capital per Capita Over Time')
```

```
plt.legend()
# Plot for consumption per capita over time
plt.subplot(1, 2, 2)
plt.plot(t, c, label='Consumption per capita (c)')
plt.axhline(y=c_star, color='r', linestyle='--', label='
    Steady-state c*')
plt.xlabel('Time')
plt.ylabel('Consumption per capita (c)')
plt.title('Consumption per Capita Over Time')
plt.legend()
plt.tight_layout()
plt.show()
```

Explanation:

- Parameters: The code uses standard parameter values commonly found in the literature.
- Steady-State Calculation: It computes the steady-state values of capital and consumption per capita.
- Differential Equations: The ramsey function defines the system of differential equations for capital accumulation and consumption growth.
- Shooting Method: The shooting_function uses a shooting method to find the optimal initial consumption c0 that ensures convergence to the steady state.
- Simulation: It simulates the model over the specified time span using solve_ivp.
- Plotting: The results are plotted to show the paths of capital and consumption per capita over time. Output:

The code will display the steady-state values and the optimal initial consumption. It will then produce two plots showing how capital per capita and consumption per capita evolve over time, converging to their steady-state values.

Notes:

The shooting method is essential because the Ramsey model is saddle-path stable, meaning there is a unique consumption path that leads to the steady state from a given initial capital stock. If the root-finding does not converge, you may need to adjust the initial guesses for c0 or check the parameter values. This code provides a complete solution to the Ramsey growth model using standard parameter values and should serve as a solid foundation for further analysis or extensions of the model.

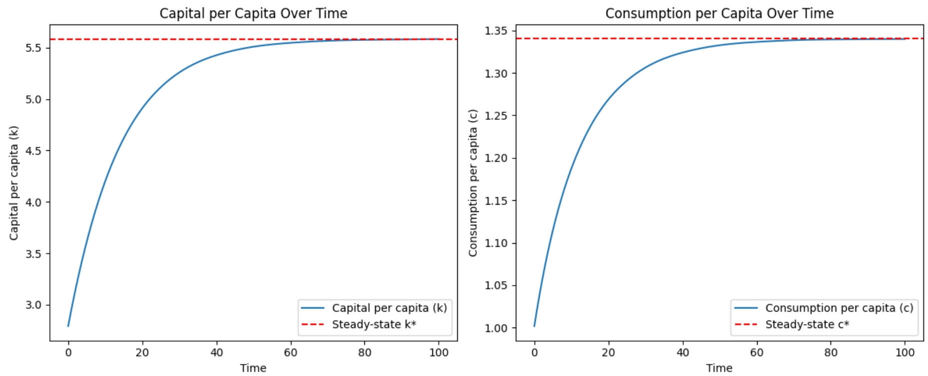


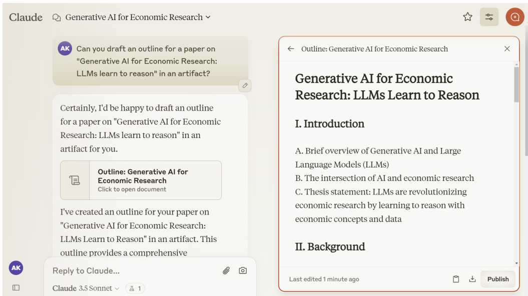
Figure 3: Solution to Ramsey model produced by o1

The code is written elegantly and explained well. The output produced by the generated code is shown in Figure 3.

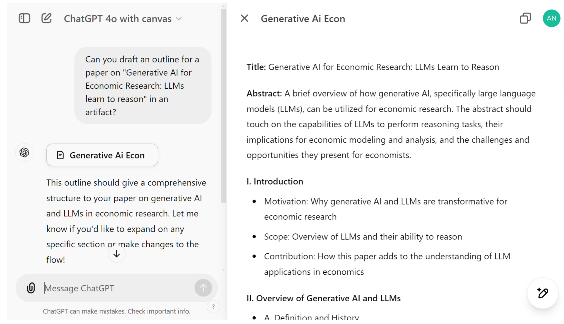
Meta recently published a paper on an approach called "Thought Preference Optimization" that demonstrates that the benefits of LLMs thinking can extend even further (Wu et al., 2024). Their method shows that optimizing thought generation before responding improves performance across a broad range of applications, from marketing and content creation to general knowledge tasks. It is likely that this approach will be implemented by future versions of Meta's open-source models.

The Road to LLM Agents A key strategic goals of frontier AI labs is to evolve LLMs into autonomous AI agents—systems that can maintain objectives across multiple exchanges, plan sequences of actions, and independently pursue specified goals (see e.g. Heikkilä, 2024). While current LLMs primarily react to immediate prompts, agents would proactively gather information, formulate plans, and take actions to accomplish tasks. The advances in reasoning capabilities described above represent a crucial step toward this goal, as autonomous agents require the ability to plan multi-step actions and revise strategies based on outcomes.

Two other key requirements for agency are already emerging: first, the ability to maintain coherent objectives over time through long-term memory and expanded context windows, as discussed earlier in this section; and second, the capability to interact with external tools and APIs to gather information and take actions in the world. The latter developments will be covered in the next subsection on access modes, particularly in the context of LLMs' autonomous "computer use" capabilities, and in the ensuing subsection on LLM-powered search, which gives LLMs real-time internet access. As these components come together, we may see LLMs evolve from passive tools into more active research collaborators.



(a) Claude Artifact



(b) ChatGPT Canvas

Figure 4: Anthropic's and OpenAI's workspaces for interactive LLM collaboration

2.2 Access Modes to Frontier LLMs

Several new access modes to frontier LLMs have emerged over the course of 2024—in addition to the traditional text-based interaction modes. The most notable are workspaces for interactive LLM collaboration, real-time voice assistants, and the emerging autonomous "computer use" capabilities of LLMs.

2.2.1 Workspaces for interactive LLM collaboration

Enabled by longer context windows and greater speeds, LLM providers have introduced innovative workspace environments in recent months that allow users to interactively collaborate with LLMs on content, as summarized in Table 3. Anthropic led the way with the introduction of Artifacts for Claude in June 2024 (left panel in Figure 4), followed by OpenAI's launch of Canvas for ChatGPT in October 2024 (right panel). Concurrently, developers of office package such as Microsoft and Google have increasingly integrated LLM capabilities into their existing workspaces. The resulting products allow users to move beyond the limitations of traditional chat interfaces and to interact with AI assistants in a more dynamic and collaborative environment for content creation, editing, and interaction. These features represent a significant leap forward in human-AI collaboration, providing dedicated spaces for users to credit, edit, and build upon AI-generated content in real-time.

Artifacts in Claude can be activated by clicking at the user button at the bottom left and choosing "Settings," which opens a menu that lets the user check the option "Enable artifacts." When Claude finds it useful, or when the user explicitly asks for an artifact, the chatbot opens a dedicated output window to the right of the chat interface where the user can instantly view and interact with the generated content. This feature supported a range of different outputs, including text (as in Figure 2a), code snippets, flowcharts, SVG graphics, websites, and interactive dashboards, which are all worth

Table 3: Interactive Workspaces for LLM Collaboration

Workspace	**Key Features**
Anthropic Claude Artifacts	• Dedicated output window • Supports text, code, flowcharts, SVG graphics, websites, dashboards • Real-time refinement and modification • Sharing and remixing capabilities
ChatGPT Canvas	• Separate collaboration window • Text editing and coding capabilities • Options for edits, length adjustment, reading level changes • Code review and porting features
OpenAI Advanced Data Analysis	• Data upload and analysis • Visualization capabilities • Python code execution in backend • Error correction and refinement
Claude Analysis Tool	• Fast exploratory data analysis • Interactive visualizations with real-time adjustments
Google NotebookLM	• Document upload for research grounding • Quick summarization and questioning • Citation and quote provision • "Deep dive conversation" podcast generation
Microsoft Copilot	• Integration with Microsoft 365 products • Assistance in Word, Excel, PowerPoint, etc. • Data analysis, formula construction
Google Gemini for Workspace	• Integration with Google's office suite • Assistance in Docs, Sheets, Slides, Gmail
Cursor AI Code Editor	• AI-assisted coding • Code suggestions and queries • Optimization recommendations • Debugging assistance • Real-time collaboration

trying out. The user can interact with the artifact by asking the chatbot to refine or modify the output in line with her instructions, allowing for rapid prototyping and iteration. The feature also allows users to publish and share artifacts with other users who can subsequently remix them, enabling easy collaboration.

Canvas in ChatGPT can be activated for paying users by selecting "GPT-4o with canvas" from the model selection menu at the top left of the screen. It is based on a similar concept as Claude Artifacts, opening in a separate window that allows users and ChatGPT to collaborate on writing and coding projects, but also offers some unique features that make it more powerful. The interface allows users to edit the content in the canvas like in a text editor or to select specific paragraphs and provide instructions on how to change the text or ask questions about it. Moreover, for text, the button at the bottom right of the canvas (Figure 2b) includes options to ask the LLM for suggested edits, to adjust the length of the content (shorter or longer) and the reading level (from kindergarten to graduate level), or to add a "final polish." For code, the button includes options to review and comment on the code, to port it to a different language, to fix bugs, and to add comments or logs—all while keeping the entire content in mind.

Advanced Data Analysis in ChatGPT has been available in a rudimentary form since mid-2023 (originally named "Code Interpreter," explored in Examples **??** and **??** below) but has been significantly improved after the release of GPT-4o in 2024, turning it into a collaborative workspace where the main focus is on interacting with the data. It allows users to upload data in a variety of formats, for example, spreadsheets, and analyze, visualize and process the data in a multitude of ways. On the backend, ChatGPT writes and executes python code to perform the necessary operations, implying that a wide range of analyses can be performed, limited only by the "intelligence" of the LLM that is writing the code. Simple routine tasks are usually no problem. For more complicated forms of analysis, Advanced Data Analysis may make errors but is frequently able to correct them automatically in response to python error messages or the user's requests.

Claude Analysis Tool is Anthropic's response to ChatGPT's Advanced Data Analysis, using JavaScript rather than python as its underlying engine. Released in October 2024, it allows Claude to visualize and explore data in an elegant, intuitive, and interactive manner. This makes the tool most valuable for quick data explorations and visualizations. For example, I created Figure 1 using the Claude Analysis tool. However, it is less suitable for specialized econometric analysis than ChatGPT's data analysis as it lacks the wide range of python libraries offering econometric tools.

NotebookLM by Google represents a more specialized approach to AI-assisted research, facilitating background research and literature reviews, leveraging the long context window of Google's Gemini models. Its key feature is to allow users to upload a wide range of relevant research papers or documents that ground the system AI in a specific corner of the literature and enables them to quickly summarize uploaded documents, ask targeted questions about the content, and brainstorm on the materials provided. Importantly for academic work, NotebookLM provides citations and relevant quotes from original sources during exchanges, supporting the rigorous documentation required in research.

In September 2024, NotebookLM introduced a groundbreaking feature called "Deep dive conversation," which allows users to generate podcasts based on the provided materials. These podcasts feature an insightful conversation between two virtual agents who explain the essence of the material in simple and clear terms for non-experts, narrated in a highly engaging manner. For research use, this is one of the most impressive capabilities that have emerged in recent months. It offers a novel way to disseminate complex economic research to a broader audience, potentially increasing the impact and accessibility of economic studies. The tool is available for free at `https://notebook.google`.

Below is the beginning of a sample podcast generated based on uploading the original version of this research paper (Korinek, 2023a):

Chat 3 (NotebookLM): Generate deep dive conversation.
[Attachment: w30957.pdf]

Male host: Okay, so you're interested in the economics of AI...
Female host: Yeah.
Male host: ...specifically, all the buzz about these large language models, LLMs.
Female host: *mmhmm*
Male host: You're in the right place, because we're doing a deep dive into how these things actually work, what they could do for economists like you right now, and what the future might hold.
Female host: And it's a future that's changing rapidly. It seems like every day there's some new breakthrough in what these models can do.
Male host: Well, let's unpack that a bit.
Female host: *mhm*
Male host: Imagine having a research assistant who can not only brew you a perfect cup of coffee, but also write code in multiple languages, ...
Female host: Yeah.
Male host: ...summarize dense academic papers in seconds, and even brainstorm potential downsides to your economic model. That's the promise of LLMs, right?
Female host: Yep.
Male host: But how do they actually work? Is it really some kind of artificial intelligence magic?
Female host: It's not magic, but it's pretty close. Think of LLMs as incredibly

complex neural networks, but instead of crunching numbers, they're trained on massive data sets of text. We're talking about everything from academic journals and books to code repositories and even social media posts—they process all of that information, learning patterns, relationships between concepts, and even different writing styles... [Remainder omitted for space reasons.]
[Transcription using otter.ai. Full audio version available at `https://t.ly/GrC0d`.]

The generated podcast lays out the material in the paper in an engaging conversational format while covering the content in a clear and insightful manner targeted at laypeople. One interesting observation is that the male host mistakenly attributes physical capabilities to AI research assistant ("...who can not only brew you a perfect cup of coffee...").

Since October 2024, NotebookLM allows users to customize the generated podcasts with specific instructions. For researchers, a useful prompt may look like this: "Your audience are PhD economists who are eager to learn how to effectively use LLMs in their research work."

Whereas the workspace tools described so far are based on the strategy of creating a new interface from scratch, like Artifacts or Canvas, in order to facilitate interactions with LLMs, the following tools follow the opposite strategy—they incorporate LLMs or similar systems to automatically perform functions in existing workspaces. Given the wide user base of the associated products, this integration will likely lead to widespread distribution of the benefits of these tools:

Copilot in Microsoft 365 integrates LLM capabilities into Microsoft Office products for an add-on subscription fee of $20/month. Copilot is based on OpenAI's latest GPT-4o model (and o1 in a pilot) as well as Microsoft-internal LLMs and can serve as an assistant and tutor for a wide range of office tasks. Use cases in Microsoft Word include creating drafts, including by brainstorming or referencing existing files; transforming text according to criteria like length, tone, formality, or intended reader; and summarizing and asking questions about content. In Microsoft Excel, use cases including extracting, converting, or reformatting data (use cases in data analysis); constructing, editing, and explaining formulas, even complicated ones; creating tables and charts to analyze and visualize data. Use cases in Microsoft PowerPoint include brainstorming, outlining and creating slides; enhancing content with images, design elements, and interactivity; summarizing and organizing presentations to highlight key points and action items; and even anticipating the most likely audience questions. Copilot is also available for all other programs that are part of the Microsoft 365 package, including Outlook, Teams, OneNote, as well as for Microsoft Edge.

Gemini for Google Workspace uses Google DeepMind's Gemini series to offer a set of similar capabilities for Google's office suite, including in Google Docs for writing documents, Google Spreadsheets, and Google Slides for presentations, as well as in the Gmail service.

Cursor – The AI Code Editor is a tool that is specialized in AI-assisted coding, making it particularly relevant for researchers engaged in computational work, data analysis, and econometric modeling. Cursor takes advantage of the long context windows and greater speeds of the latest LLMs to take the code suggestions pioneered by Github Copilot to new heights. It integrates AI assistance into all aspects of the coding process, including code suggestions, queries about code, edits to selected code according to instructions, recommendations for code optimization, and help for debugging—all while keeping the entire code base of a project in its context window to recognize interdependencies. It also offers support real-time collaboration, facilitating teamwork on large-scale projects. Cursor is built on (or, more specifically, forked from) the popular VS Code environment and can employ multiple different LLMs, including GPT-4o, o1, and Claude 3.5 Sonnet. Peng et al. (2023) report that GitHub Copilot delivered productivity gains of 126% for coding back in 2023. Based on user reports, the gains from Cursor may be even larger.

In a similar vein, the LaTeX editor Overleaf offers a tool called Writefull that is specialized in producing and editing LaTeX code, including tables and equations.

Additional Tools for Providing LLMs with Context One of the challenges in using the current generation of frontier LLMs is that they are excellent at processing content, but it is difficult or time-intensive for the user to supply the most relevant context for a work task. Anthropic and OpenAI have developed two slightly different solutions to this problem:

Anthropic introduced *Projects for Claude* in June 2024, which allow users to upload background documents that are relevant for multiple chat interactions and to organize and bring together related chats and artifacts in one place. For example, I have created a project on "Generative AI for Economic Research," to which I added the earlier versions of this paper (Korinek, 2023a) and the project-specific custom instructions "Help me draft content for my research project on 'Generative AI for Economic Research' in a similar style to the earlier versions and in a format that is helpful for economic researchers." When I use Claude to work on content related to this paper, I start a new chat that is part of this project, automatically providing Claude with all the relevant content. Projects help ground the LLM's outputs in relevant context and background knowledge to effectively mitigate the "cold start" problem when opening a new chat. They can also be shared across teams, enabling more collaborative workflows.

OpenAI allows users to create *Custom GPTs* that enhance GPT-4o's functionality by adding specialized background knowledge, interactive tools, and customized instructions tailored to specific instructions, ranging from writing to economic analysis tools. Users can also create their own custom GPTs by clicking "Explore GPTs" at the top left and the "Create" button, and going through the ensuing process step by step.

The resulting custom GPTs can be used privately or shared publicly in a GPT Store. Custom GPTs created by others can also be located in the "Explore GPTs" menu at the top left, which offers users the ability to browse, install, and use a wide range of extensions and applications that are developed by third-party creators. Two custom GPTs that economists may find useful are (1) Wolfram, which provides access to computation, math, curated knowledge and real-time data from Wolfram Alpha, the maker of Mathematica, and (2) Consensus, which offers an AI-based research assistant that searches 200m academic papers to provide science-based answers with citations to the underlying articles.

2.2.2 Real-time Voice Assistants

A significant innovation in access modes is a new generation of real-time voice assistants. Earlier voice assistants transcribed a user's spoken language into text that was fed into an LLM; after processing the request, the LLM's response was translated into audio again. By contrast, the new generation natively processes spoken text with all its nuances in tone and emotional expression and responds accordingly. Moreover, it also allows users to jump in and interrupt the flow mid-sentence in a way that allows for a more natural and fluid conversation. Some users report that they keep the ChatGPT app open on their phone in voice mode throughout certain work tasks, or even throughout the day, so that they can easily draw on the their digital assistant at any point without raising a finger.

The following are the leading interactive voice assistants of this new generation:

- OpenAI's Advanced Voice Mode is a feature of its ChatGPT mobile app that offers perhaps the most natural interaction, using the GPT-4o model.[8] In a version that is not yet publicly released, the model can also use the mobile phone's camera to include a video view of the user or their surroudings in its interactions. A desktop app to use Advanced Voice Mode that can see and respond to the information on a user's desktop is also in the works.

- Google's Gemini Live also allows for fluid voice conversations with users based on the Gemini series of models. Google is working on integrating Gemini Live with apps across the Google eco-system, including GMail, Calendar, Docs, YouTube, and Maps, to turn it into a powerful productivity assistant. Moreover, it is also working on a "Project Astra" (**A**dvanced **s**eeing and **t**alking **r**esponsive **a**gent) that will incorporate vision features in Gemini Live.

- Apple Intelligence voice assistant is part of the latest round of operating system updates across all Apple devices. It introduces an assistant that integrates seamlessly with Apple's ecosystem, handling general queries, managing tasks, and

[8] At the time of writing, Advanced Voice Mode is not yet available to ChatGPT Plus or Free users in the European Union.

interacting fluidly across apps like Mail, Calendar, and Notes. For more complex or nuanced inquiries, Apple Intelligence selectively leverages ChatGPT, adding depth to its responses when necessary. Although Apple's proprietary AI system provides a smooth, integrated experience, some users report that it lacks the advanced capabilities of standalone ChatGPT, particularly in handling complex, multi-layered questions.

- For researchers interested in open-source solutions, the AI lab Standard Intelligence has publicly released Hertz-dev, accessible at `https://si.inc/hertz-dev/`. Their models provide a versatile, real-time voice assistant solution that facilitates natural spoken interactions with LLMs. Its efficient compression and ultra-low latency make it ideal for real-time applications, while its audio generation capabilities enable nuanced, responsive conversations. As an open-source platform, Hertz-dev offers high customizability, allowing researchers to tailor it to their specific needs, such as automated interviews or integration with other research tools for a seamless, interactive experience.

To combine voice interactions and traditional text-based interactions, both OpenAI's and Google's models provide users with transcripts of their voice interactions which can be copied and pasted for further processing and for written research products.

2.2.3 Autonomous Computer Use

The perhaps most breathtaking recent advance has been an autonomous desktop assistant, simply labeled "Computer use," which was released by Anthropic in beta mode in October 2024. The system enables Anthropic's most cutting-edge model, Claude 3.5 Sonnet, to directly interact with your computer's interface, allowing it to see your computer screen and giving it access to virtually any software application that can be installed on a computer. This implies that the model can control your cursor, click buttons, type into text fields, and even navigate through software interfaces—as if another intelligent being were sitting at your computer. Although still preliminary, computer use gives LLMs the ability to automate a wide range of tasks that require multiple applications or complex workflows on a computer, ranging from organizing files and updating software to conducting online research. In effect, this development gives LLMs nearly unlimited access to external tools, enabling them to seamlessly interact across platforms and applications without manual intervention. Anthropic's Claude with "Computer use" can currently be accessed through Anthropic's API, which enables users to programmatically direct Claude to perform any desired operations on a computer. An instructive demo video is available at `https://www.youtube.com/watch?v=ODaHJzOyVCQ`.

Google's Project JARVIS (acronym for "Just A Rather Very Intelligent System," which is inspired by the AI assistant in the *Iron Man* franchise) is an experimental AI assistant that operates within the Google's Chrome browser environment, where it can perform web-based tasks such as filling out forms, navigating websites, and making online purchases. Currently, Jarvis is in the experimental phase and expected to be available more broadly in December 2024. Its functions are limited to browser-based tasks, unlike Anthropic's Claude, which can interact with any software installed on a computer. This makes Jarvis highly useful for automating tasks online but less versatile for complex workflows that require access to local applications.

For researchers, autonomous desktop assistants offer significant potential. They can automate standard research workflows, such as organizing datasets, managing references, and conducting data analyses in econometric packages. Additionally, for bulk operations like systematically collecting information from multiple sources, running batch simulations, or automating data entry, an LLM with desktop control can handle repetitive tasks with ease and efficiency. However, these capabilities also introduce risks, including grave security risks and privacy concerns, as such systems obtain full control over the user's device or browser.[9] Researchers must weigh these risks carefully, ensuring that sensitive data and systems are protected when taking advantage of the automation benefits these tools can offer.

2.2.4 LLM-based research tools

There are also a growing number of dedicated research tools that are based on LLMs and facilitate or automate research tasks. I will highlight two:

Expected Parrot Horton et al. (2024) develop an open-source python package to facilitate research on LLM-based simulations and surveys. In a dig at the term "stochastic parrot" that was used to critique LLMs, they have developed *Expected Parrot Domain-Specific Language* (EDSL), which takes advantage of LLMs' ability to generate a wide range of context-specific data that closely mirror human behavior and social dynamics. EDSL allows researchers to define a set of as *Questions* that are answered by *AI Agents* simulated by defined *Models* to produce a set of *Results*, which can be grouped into *Surveys* and contextualized with *Scenarios* (capitalization used to refer to specific objects in EDSL).

This approach enables economists to efficiently manage large-scale tasks with intricate dependencies, agent behaviors, and model parameters without getting bogged down in programming details. As a result, EDSL offers a powerful toolkit to conduct LLM-based simulations of detailed surveys and experiments, label large datasets, augment existing data, and generate synthetic data. Researchers can design AI agents with specific traits, utilize multiple language models simultaneously, and incorporate complex

[9]For example, Anthropic notes that Claude with computer use sometimes erroneously follows instructions that it happens to read on open webpages or in images, thereby overriding the instructions that it has been given by its user. For this reason, they recommend that computer use is run on a dedicated virtual machine or container with minimal access privileges to prevent system attacks or accidents.

logic and agent memory into their surveys. EDSL's built-in analysis and visualization tools, integrated into the python ecosystem, allow for both seamless execution and interpretation of research outcomes.

Sakana.ai AI Scientist Lu et al. (2024) at the Japanese startup sakana.ai introduce an automated framework for end-to-end scientific paper generation in computer science based on LLMs. The AI Scientist, as they call it, is designed to autonomously generate research ideas, implement experiments by running code, analyze results, and produce complete academic papers. While currently limited to a specific area within computer science in which progress can be made simply by writing code (machine learning algorithms and architectures), this approach demonstrates the potential for LLMs to assist across the research process.

The AI Scientist operates by generating novel research ideas, writing code to implement experiments, executing those experiments, and then drafting a full scientific paper based on the results. The system incorporates an automated reviewing process to evaluate the generated papers, mimicking the peer review system in academic publishing. Sample papers are available at `https://sakana.ai/ai-scientist/`.

Although the current quality of the generated papers is mediocre, lacking the full originality, depth and rigor of research authored by human experts, the framework points towards the potential future capabilities of LLMs in scientific research. It serves as a proof of concept for how LLMs could be leveraged to augment and accelerate the scientific process in the future. As LLM capabilities continue to advance, especially as they make breakthroughs in reasoning (see Section 2.1), systems like the AI Scientist may evolve into powerful tools for idea generation and the execution of research even in fields like economics.

2.2.5 Traditional text-based access modes

The following summarizes the more traditional access modes for LLMs that have been available for the past two years:

- **Web-based Chatbots:** The models in Table 1 are all accessible as chatbots under the URLs listed in the last column. The chatbot interface, pioneered by Anthropic but first publicly released by OpenAI in the form of ChatGPT in Nov. 2022, allows users to prompt LLMs as assistants or tutors. Most of the examples documented in the use cases below illustrate this mode of interaction, which has been the most popular way of accessing LLMs over the past two years. However, I anticipate that LLM use will gradually shift towards the interactive workspaces described in Section 2.2.1.
  The free versions of the listed chatbots typically comes with usage restrictions or provides access to less powerful model versions. In my experience, this makes it worthwhile to pay the $20 monthly subscription fee that is typically required for

full access to the frontier models listed in the table.

All of these chatbots are also available via apps on Apple and Android mobile phones. Moreover, OpenAI and Anthropic have also developed desktop apps for their chatbots that are available for download at `https://openai.com/chatgpt/download/` and `https://claude.ai/download` respectively and, once installed, can be conveniently accessed via the keyboard shortcuts Ctrl+Space and Ctrl+Alt+Space.

- **Web-based Experimentation Platforms:** All major LLM providers also offer web-based interfaces that offer greater functionality and flexibility than chatbots but do not require programming knowledge. These platforms, such as OpenAI Playground (`https://playground.openai.com`), Google AI Studio (`https://aistudio.google.com`), and Anthropic Console (`https://console.anthropic.com`)), allow users to experiment with different model settings, like temperature and top-p sampling, and provide more control over the input and output formats compared to chatbots. Such experimentation platforms are particularly useful for exploring the capabilities of LLMs, testing prompts, and fine-tuning models for specific tasks.

- **Application Programming Interfaces (APIs):** For the maximum level of customization and integration, the listed models are also accessible through APIs, which allow programmers to integrate LLMs directly into their own software applications. This enables a wide range of more advanced and customized use cases, such as automating repetitive tasks or analyzing large datasets using natural language processing techniques. APIs provide more flexibility and control compared to above two options and can be employed on a pay-per-use basis, but they also require a higher level of technical expertise to use effectively. Accessing LLMs through APIs typically involves signing up for an API key from the model provider (which can be thought of as a credit card for LLM tokens), installing a client library in the programming language of choice, and writing code to interact with the API endpoints. While this process may be more complex than using a chatbot, it unlocks the full potential of LLMs for those with the necessary programming skills. The replication package for this paper demonstrates how to use APIs to automatically query LLMs.

- **Locally Operating LLMs:** Open-source models allow researchers to run LLMs on their own computers, offering advantages such as data privacy, cost-effectiveness, customization, and offline accessibility. The computational resource requirements imply that only small models can be executed at a reasonable speed on desktop computers. However, advances in computational capacity as well as rapid efficiency gains of LLMs that allow greater capabilities of smaller models are rapidly making the local use of LLMs more attractive. Two solutions that make it particularly easy to deploy LLMs locally are:

- LM Studio allows users to download and run a range of open-source LLM, including VLMs, on their personal computer or server.
- llamafile makes it possible to download LLMs in a single file and run it on a wide range of computer systems.

Centralized Hubs for LLM Interaction and Experimentation A useful website with a user-friendly chat interface that offers access to all leading LLMs is `https://poe.com`. Similarly, a website that offers users a web-based experimentation platform with access to a wide range of different models is `https://nat.dev/`.

Prompting Guides for Text-based Access Although it is useful for everyone to gain experience interacting with LLMs through their own experimentation and learning-by-doing, most leading model providers now also offer public documents that describes strategies for how to prompt LLMs. The following is a list of such guides:

- `https://platform.openai.com/docs/guides/prompt-engineering`
- `https://ai.google.dev/gemini-api/docs/prompting-intro` and an example of a research assistant at `https://aistudio.google.com/app/prompts/research-assistant`
- `https://docs.anthropic.com/claude/docs/intro-to-prompting` and an associated prompt library at `https://docs.anthropic.com/en/prompt-library/library`
- `https://llama.meta.com/get-started/#prompting`

Anthropic also offers a "prompt generator" that helps users to optimize the prompts that they employ for querying LLMs through the Anthropic Console Dashboard (click "Generate a Prompt"). This tool converts basic instructions in written language describing a user's intent into a more complex and specific set of instructions that induce LLMs to generate higher-quality output. Naturally, the user can edit and adjust the generated instructions before use to ensure that they capture their intent as well as possible. I illustrate this using the following example. Throughout the remainder of this section, I will present all examples generated by LLMs in boxes, with the prompt in the header and the LLM's generated response in the body of the box.

Chat 4 (Anthropic Prompt Generator): I have written an economics paper and would like to create a draft of a blog for Brookings.

Your task is to help draft a blog post for the Brookings Institution based on an economics research paper. Here is the full text of the paper:

<paper>
{{PAPER}}
</paper>
The target audience for this blog post is:
<audience>
{{AUDIENCE}}
</audience>
First, carefully read through the entire paper, taking notes on the key points, main findings, and implications in a <scratchpad> section.
Next, write a short outline for the blog post in a <outline> section. Keep the target audience in mind as you decide what to focus on and how to present the information.
Then, using your outline as a guide, write out the full draft of the blog post in a <blog_post_draft> section. Remember to:
- Use clear, engaging language that will be easily understood by the target audience
- Incorporate relevant examples, anecdotes or analogies where appropriate to make the post more relatable and interesting
- Avoid copying word-for-word from the original paper. Put the key concepts into your own words.
- Focus on the implications and "so what" of the research rather than getting too in-the-weeds with technical details
After you finish the draft, review it to make sure it covers the most important and interesting points from the paper in a way that will resonate with the target Brookings audience. Make any final tweaks before submitting your draft.

The generated prompt provides a remarkable level of detail that employs several cutting-edge prompting techniques to obtain the best possible results: it asks the model to start by taking notes in a "scratchpad" section, which induces the model to synthesize the main lessons of the text. Then it asks the model to prepare an outline, which induces the model to generate a plan. Only after these steps does the prompt ask the model to write a draft of the text. The prompt also leaves two locations for the users to insert the paper in question and to define the audience. I simply substituted these by writing "The full text of the paper is attached" and inserting "policymakers" as the target audience. The prompt generated useful results on all three of the leading LLMs that would save considerable time in such a project, although the generated blog post was not quite up to the depth and quality of a think tank like Brookings.[10]

[10] I asked Anthropic Prompt Generator to create prompts for blog posts for several different outlets. For some reason, asking for a blog targeted at Brookings gave rise to a higher-quality prompt than asking for any other outlet I attempted. My best explanation is that the model internalizes that Brookings is known for its high-quality analysis. (Disclaimer: The author is a non-resident fellow at Brookings.)

Bsharat et al. (2024) systematically study 26 different guiding principles for prompting leading LLMs and offer prompting advice based on a systematic evaluation of their effectiveness. Curiously, they find that strategies such as (fictitiously) offering LLMs a tip or threatening a penalty improve performance.

2.3 Advances in LLM-Powered Search

In recent months, Perplexity AI and OpenAI's ChatGPT Search have become increasingly capable at combining the strengths of traditional search engines with those of generative AI. Like search engines, these systems have access to real-time information via the web, which overcomes the shortcoming of earlier generations of LLMs that their world knowledge was confined to their increasingly outdated training data. However, like LLMs, they can process the results using the sophisticated synthesis capabilities of generative AI. While traditional search engines provide lists of links that users must manually browse and evaluate, LLM-powered search tools can perform this for the user and synthesize information from multiple sources into coherent, direct responses that can be formatted in any desired way, for example, in table format. They maintain all the powerful features of LLMs, including the ability to understand conversation context and support follow-up questions, but now enhanced with real-time knowledge such as the latest economic research, current market data, and emerging policy debates.

The following two chats provide examples, for which I uploaded an earlier version of this article:

Chat 5 (ChatGPT-4o with Search): Can you draft a new subsection on LLM-powered search for my paper on "Generative AI for Economic Research," focusing on ChatGPT Search and Perplexity. Search the web to find out what people find most useful about these search tools. Then draft four concise paragraphs in an engaging academic style that fits well with the paper:
1. Introduce LLM-powered search as one of the recent innovations - what's the benefit of combining LLMs with search?
2. Overview and comparison of ChatGPT Search and Perplexity AI
3. Applications for academic research
4. Risks (but emphasize that search greatly reduces hallucinations)
[Attachment: LLMs_2024h2.pdf]

LLM-Powered Search: Enhancing Information Retrieval in Economic Research
The integration of Large Language Models (LLMs) with search functionalities represents a significant advancement in information retrieval, particularly beneficial for economic research. By combining the extensive language understanding capabilities of LLMs with real-time data access, researchers can obtain synthesized, contextually relevant information efficiently. This fusion facilitates the extrac-

tion of nuanced insights from vast datasets, streamlining the research process and enabling more informed analyses.

Two prominent LLM-powered search tools are ChatGPT Search and Perplexity AI. ChatGPT Search, developed by OpenAI, extends the conversational abilities of ChatGPT by incorporating real-time web search capabilities, providing users with up-to-date information and direct citations. Perplexity AI, on the other hand, functions as an AI-driven search engine that delivers concise answers with source attributions, emphasizing accuracy and transparency. While both tools aim to enhance information accessibility, ChatGPT Search offers a more interactive experience, whereas Perplexity AI focuses on precise, citation-backed responses.

In academic research, particularly within economics, these tools offer substantial advantages. They enable scholars to swiftly access current data, literature, and analyses, thereby expediting the research process. The ability to retrieve information with source citations ensures the credibility of the data, which is crucial for rigorous academic work. Moreover, the conversational interfaces of these tools allow for iterative querying, facilitating deeper exploration of complex economic topics.

Despite their benefits, it is important to acknowledge potential risks associated with LLM-powered search tools. One concern is the possibility of generating inaccurate or misleading information, known as "hallucinations." However, the integration of search functionalities significantly mitigates this issue by grounding responses in real-time data and providing source citations. This enhancement not only improves the reliability of the information retrieved but also fosters transparency, allowing researchers to verify and cross-reference data effectively.

Chat 6 (Perplexity AI Pro): [same user prompt as in Chat 5]

Here's a draft subsection on LLM-powered search for your paper:

LLM-Powered Search: Enhancing Information Retrieval

The integration of Large Language Models (LLMs) with traditional search engines represents a significant leap forward in information retrieval. By combining the vast knowledge base and natural language understanding capabilities of LLMs with real-time web search, these tools offer a more intuitive and comprehensive search experience. LLM-powered search can interpret complex queries, synthesize information from multiple sources, and present coherent, contextually relevant responses, effectively bridging the gap between raw data retrieval and insightful analysis.

Two prominent examples of LLM-powered search are ChatGPT Search and Perplexity AI. ChatGPT Search, integrated into OpenAI's ChatGPT Plus, leverages the GPT-4 model to provide up-to-date information and web browsing capabilities. It excels in offering detailed, conversational responses and can follow up on complex queries. Perplexity AI, on the other hand, is designed as a standalone search engine that combines web search with AI-generated summaries. It distinguishes itself by providing source citations for its responses, allowing users to verify information easily. While both tools offer similar core functionalities, Perplexity AI's emphasis on transparency and its dedicated search interface may appeal more to researchers seeking clear attribution of sources.

For academic research, these tools offer several advantages. They can rapidly summarize the state of knowledge on a given topic, potentially accelerating literature reviews. Researchers can use them to explore interdisciplinary connections, generate research questions, or find relevant datasets and publications. The ability to ask follow-up questions and receive clarifications in natural language can be particularly useful when delving into unfamiliar fields or complex topics. Additionally, these tools can assist in identifying research gaps by synthesizing information across multiple studies and highlighting areas where consensus is lacking or evidence is sparse.

While LLM-powered search tools significantly reduce the risk of hallucinations compared to standalone LLMs by grounding responses in current web data, some risks remain. The potential for bias in search results and AI interpretations persists, and the recency and reliability of sources can vary. Researchers must maintain critical thinking and verify important information through traditional academic channels. Despite these considerations, the integration of LLMs with search functionality represents a powerful tool for enhancing research efficiency and exploration, offering a valuable complement to traditional research methodologies.

Both of these summaries are actually quite good and free of hallucinations although perhaps a bit sterile for my taste. However, they illustrate that LLM-powered search represents a useful step in the direction of future AI models that will be capable of autonomously writing future updates of this article. However, for now, let me also point out that the prompt that I employed provided the LLM with a lot of structure that was based on my understanding of how LLM-powered search fits into the overall landscape of generative AI for economic research.

Perplexity AI is accessible at `https://www.perplexity.ai` and offers a chat interface much like ChatGPT but is specialized in searching the web to compile responses to user prompts based on the sources it identifies. What is most useful is that it provides links supporting the statements that it makes. After submitting a query, the site first shows the keywords it searches for and then a preliminary list of the materials that it evaluates as preliminary readings. Based on these, it generates a response that starts with a clickable list of final sources, continues with a synthesis of the materials found, and ends with a conclusion that summarizes everything. The user can follow the sources provided to judge their reliability and verify the model's output. Perplexity has several different "Focus" options, including "Academic," which focuses its search on published academic papers, or "Math." It also offers a "Pro Search" option that dives deeper into topics and allows users to fine-tune the results but requires a subscription for frequent use. Another example of Perplexity AI in background research is provided in Chat 10 on page 44.

OpenAI released ChatGPT Search only recently, in October 2024, but the model represents the new state of the art in LLM-powered web search. It can be activated by explicitly instructing ChatGPT to search the web or by clicking a globe button at the bottom of the prompt window. It largely replicates the functionality of Perplexity AI but integrates this with the other capabilities of ChatGPT, making it better at multi-turn conversations and integration with applications that require data analysis. Google's Gemini chatbot also draws on an internet search for questions for which this is relevant.

2.4 Technical Improvements

The following two technical advances have been particularly useful for programming when using LLMs via API:

Structured Outputs The ability of leading LLMs to reliably produce structured outputs has advanced enormously over the past year. Traditionally, LLMs used to produce unstructured, free-text responses that could not directly be fed into other computer applications. By contrast, structured outputs enable LLMs to take in raw, unstructured, qualitative information and convert it to machine-readable data that seamlessly integrates with other more traditional computer programs, such as econometric packages. Structured outputs therefore bridge the gap between LLMs and traditional computer applications.

JSON (JavaScript Object Notation) is one of the most common formats for structured data. It organizes information into key-value pairs, which makes it both flexible and easy for computers to interpret. JSON can store everything from simple numerical values and text to complex hierarchical data, such as nested lists or objects. For example, a JSON output might look like this: {"country": "United States", "population": 332915073}. XML (Extensible Markup Language) serves similar purposes, with data represented in tagged structures, such as <country><name>United States</name><population>332915073</population></country>, but it is a bit more verbose. Both formats are compatible with many programming environments, making them widely applicable across research and data-driven applications. Here is an example:

Chat 7 (Claude 3.5 Sonnet API): List the the world's three largest economies as [{"country", "population", "GDP_per_capita"}] in JSON.


```
[
{
"country": "United States",
"population": 332915073,
"GDP_per_capita": 69288
},
{
"country": "China",
"population": 1411750000,
"GDP_per_capita": 12556
},
{
"country": "Japan",
"population": 125507472,
"GDP_per_capita": 39285
}
]
```


To illustrate the benefits of this, assume this JSON was obtained from an LLM by a python program via API and is stored in the variable `country_data`. Then individual entries, say for the first country (the index starts at 0) can simply be accessed by referring to `country_data[0]['country']` or `country_data[0]['population'])` in the python code. This type of structured output was difficult to obtain reliably from LLMs just two years ago but is now readily available so LLMs can be integrated into programmatic workflows.

Shorten et al. (2024) introduced the benchmark StructuredRAG to evaluate LLMs on their ability to produce structured outputs consistently and accurately, assessing tasks such as generating structured responses in JSON, based on criteria like precision, adherence to format, and reliability. On their benchmark, Anthropic's Claude 3.5 ranked first with an almost perfect score, followed by Google Gemini 1.5 Pro and OpenAI GPT-4o, which delivered less consistent results in August 2024. However, in a September 2024 update, OpenAI included new functionality to allow users to specify any JSON format for the LLM to adhere to, greatly improving GPT-4o's capabilities to produce structured outputs.[11]

For economists, structured outputs are useful in a variety of applications, from organizing country-level economic indicators as in the example above, to managing survey data, financial data, sentiment data, or a wide range of other data sources.

[11] For further details, see `https://platform.openai.com/docs/guides/structured-outputs/`.

Prompt Caching is a technique to enhance LLM performance by storing and reusing previously processed text. This can reduce computational redundancy, resulting in higher speed and lower costs. For example, if a user repeatedly requests similar analyses based on a long introductory prompt, caching of the prompt instead of processing it again and again multiple times can save costs and speed up processing.

Both OpenAI and Anthropic have implemented prompt caching systems in their APIs. OpenAI's implementation automatically activates when prompts exceed 1,024 tokens, caching the longest previously computed segment for reuse, which incurs only half of the usual cost of API use. Anthropic has developed a system that charges users 25% extra to write prompts into a cache but then allowing the reuse of cached information at a cost of only 10% of the regular price of text processing. For researchers, prompt caching is particularly appealing when performing text analysis in bulk as prompts can be cached and redundant computations be avoided, saving both time and money.

2.5 Practical Considerations for LLM Usage

Data Confidentiality An important issue for researchers is how to ensure the confidentiality of the data that they enter into LLMs. OpenAI offers a "Temporary Chat" option in its ChatGPT app as well as a privacy option in the user settings (turn off "Improve the model for everyone") to let users opt out from their inputs being used for training future LLMs. OpenAI does not use user data that are entered via APIs. Anthropic does not use user data for future training except with an explicit opt-in or, in rare circumstances, if it is flagged for safety review. Google advises users against entering confidential information into its Gemini apps since input data may be used for future training purposes. For highly confidential data, the safest way of using LLMs is to run a cutting-edge open-source model on a local computer.

Declaring LLM Use In economics, most AEA journals will soon require authors to declare whether and how they have employed LLMs in their research. Although I usually welcome transparency, my own perspective is that such a requirement is unnecessary and may be potentially counterproductive. LLMs are rapidly becoming essential tools in the research process, akin to word processors, calculators, or econometric software. If used responsibly, they do not inherently compromise the integrity or originality of research any more than these other widely accepted tools. It is crucial to remember that authors remain solely accountable for the content they submit, regardless of the tools used in its creation. While it may be beneficial to remind authors of this responsibility when submitting, a formal declaration requirement could inadvertently create unwarranted skepticism among readers and discourage the use of these powerful productivity-enhancing tools. Moreover, such declarations are difficult to verify conclusively, rendering them at odds with the spirit of the revelation principle, which emphasizes designing mechanisms that naturally encourage truthful disclosure.

My own perspective is to be cautious about introducing additional bureaucratic steps that may impede the research process without clear benefits. I advocate for the judicious and responsible use of LLMs in research, always coupled with careful verification of results—a practice no different from how we treat output from human research assistants or other analytical tools. The focus should remain on the quality and integrity of the final research product, rather than the specific tools used in its creation.

Watermarking Relatedly, watermarking of LLM outputs has become an important new consideration when using LLMs. Watermarking embeds markers in AI-generated text by introducing a specific fingerprint key in the pseudorandom token selection during the text generation process (Dathathri et al., 2024). This makes it possible to trace back the origin of the text to the LLM for those who know the associated key while remaining undetectable to regular readers. Google has implemented this watermarking method, called SynthID, in the output of its Gemini models, representing the first known large-scale deployment of text watermarking.

While watermarking could help establish provenance and potentially address concerns about academic integrity and unauthorized AI use, its implications deserve careful consideration. First, the markers are not reliable since they can be defeated through simple paraphrasing using other LLMs. Secondly, the practice raises privacy concerns since watermarks could theoretically enable tracking of AI-generated content back to individual users. For economic researchers, watermarking has important implications both as a subject of study (for example, regarding information asymmetries and verification mechanisms) and as a practical consideration when using LLMs for research tasks. Users of Google DeepMind's Gemini models should be aware of the watermarks contained in the generated output. It is unknown whether other labs employ similar mechanism.

Additional Resources Let me also point to two additional resources for readers interested in the topic of this paper. First, Ash et al. (2024) provides a survey of how LLMs have transformed text analysis in economic research. Dell (2024) offers a JEL survey of deep learning for economists, covering classifiers, regression models, generative AI, and embedding models, together with a companion website, EconDL.

I now turn to tangible uses of generative AI in economic research.

3 Updates On Use Cases of Generative AI

Reader's tip:

Some readers have reported that they found it useful to print out Table 4 on the next page and treat it as a "challenge list" to work through, ticking off each of the listed use cases one-by-one after they have experimented with it.

Please email me to let me know what you found most useful and what new use cases I could include in the next version!

Table 4 represents a list of use cases of generative AI for economic research together with my subjective rating of their usefulness, updating Table 1 from Korinek (2023a). The latest addition to the table is a new category on "Research Promotion," in which there are several impressive new use cases. In the following, I list brief updates for each category, but I only report details on the use cases that are new since December 2023. Readers who are interested in seeing details for the full list of use cases can access the official JEL update on the website associated with this project, `https://www.GenAIforEcon.org`.

In the third column of Table 4, I report a subjective rating of how useful I found each of the capabilities in November 2024. I no longer found the need to use empty circles (○) in Nov. 2024. A half-full circle (◐) describes capabilities that are already useful and likely to save time but are somewhat inconsistent so that they still require significant oversight. A full circle (●) reflects capabilities that are already highly useful and largely work in the expected manner. Incorporating these capabilities into your workflow will definitely save you time and make you more productive.

The new inclusions in the table since the original publication in December 2023 are marked with superscripts for 2024/06 or 2024/11 in the second column. The superscripts in the third column mark capabilities for which I have increased my ratings since December 2023. These cover the following: In the category "Writing," the ability of LLMs like GPT-4o to transcribe hand-written equations has improved significantly due to the greater vision capabilities of frontier LLMs. In the category "Data Analysis," LLMs can now classify and score text and extract sentiment due to greater ability to understand context and reason. In "Coding," LLMs have become highly useful in writing clean code, explaining code, and even debugging code, with OpenAI's new reasoning model o1 having led to significant breakthroughs. The same model is also highly useful for deriving equations in the category "Math."

3.1 Ideation and Feedback

This category has significantly benefitted from the broad-based advances in the capabilities of LLMs and their ability to process ever larger context windows. LLMs now allow users to upload entire papers and obtain detailed feedback. Some users also report that it is highly useful to voice chat with leading LLM chatbots to talk through

Category	Task	Usefulness
Ideation & Feedback	Brainstorming	●
	Feedback	◐
	Providing counterarguments	◐
Writing	Synthesizing text	●
	Editing text	●
	Evaluating text	●
	Converting hand-written equations$^{24/6}$	●$^{+}$
	Generating titles & headlines	●
Background Research	Summarization	●
	Condensing YouTube videos$^{24/6}$	●
	Literature Research	◐*
	LLM-Powered Search$^{24/6}$	◐
	Formatting References	●
	Translating Text	●
	Explaining Concepts	◐
Coding	Writing code	●$^{+}$
	Explaining code	●$^{+}$
	Translating code	●
	Debugging code	●$^{+}$
Data Analysis	Locating data sources$^{24/6}$	◐
	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	◐
Research Promotion	Social media posts	●
	Presentation slides$^{24/11}$	●
	Blog posts$^{24/11}$	●
	Conducting interviews$^{24/11}$	●
	Podcasts$^{24/11}$	●

The third column reports my subjective rating of LLM capabilities as of November 2024:
○: 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
Superscripts 24/6 or 24/11 in 2nd row represents year and month of new inclusions
Superscripts in last row denote upgraded ratings in 2024/06(*) and 2024/11(+)

Table 4: Summary of LLM capabilities and rating of usefulness

research problems, obtain feedback, or become unstuck from a problem.

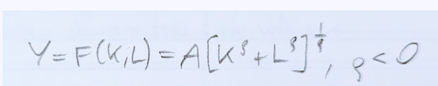
The ability of LLMs to synthesize the conventional wisdom captured in their training data, to be a tutor, or to offer a broader framing to questions makes them well-suited for this purpose. Based on my own experiments, I concur that interacting with LLMs via voice chat makes for a different experience than interacting via text – an experience that can be useful just like it is sometimes useful to talk through a problem with a colleague.

3.2 Writing

Writing text is one of the core competencies of LLMs. This category of use cases has benefitted enormously from the new workspaces for interactive LLM collaboration, described in Section 2.2.1. I regularly ask Claude or ChatGPT to draft a first version of a text based on bullet points or based on a recorded and transcribed lecture of mine, and then I interact with a Claude Artifact or ChatGPT Canvas to edit the text and make sure it accurately reflects the message I want to convey to the reader in all its nuances. The use cases listed in the table are significantly easier to access in the described collaborative workspaces.

I have also used the vision capabilities of LLMs to help me write up math. This is a capability that has seen dramatic improvements over the course of 2024. In my June 2024 update, I reported the following use case:

Chat 8 (GPT-4): Convert the attached image to LaTeX code.



The LaTeX code for the equation you provided would look like this:

$$Y = F(K, L) = AK^{\alpha}L^{\beta}, \quad \alpha, \beta \leq 0$$

By contrast, when I attempted the same query in November 2024, both ChatGPT and Claude transcribed the equation flawlessly:

$$Y = F(K, L) = A\left[K^{\rho} + L^{\rho}\right]^{\frac{1}{\rho}}, \quad \rho < 0$$

A dedicated tool for converting math in pictures to LaTeX format is `MathPix.com`.

3.3 Background Research

Recent advances in background research derive primarily from the capacity of frontier LLMs to process larger and larger amounts of text in their context window and from their ability to integrate real-time information from the web. Larger context windows enable cutting-edge systems to process multiple academic papers or books and even YouTube videos with growing accuracy. Systems like NotebookLM, described in detail on page 22, allow researchers to upload all the materials of a given project into one place and interact with it.

A significant time-saver for me has been the ability of Google DeepMind's Gemini to digest YouTube videos. This requires that the corresponding extension is activated in the settings of Gemini.

Chat 9 (Gemini): Provide a one-paragraph summary of this: `https://www.youtube.com/watch?v=sjL6Gl6ZIqs`

This video is about a new large language model, Claude 3 Opus, created by Anthropic. In the video, the creators test Claude 3 Opus's ability to analyze the world economy. The video shows that Claude 3 Opus can access and process information from the web using a tool called web view. It can then use this information to answer questions and complete tasks. For example, the video shows that Claude 3 Opus can generate a graph of US GDP trends and predict future GDP growth. The video also shows that Claude 3 Opus can break down complex problems into subproblems and distribute the work among other versions of itself, which is called dispatch sub agents. This allows Claude 3 Opus to analyze the economies of multiple countries simultaneously. Overall, the video demonstrates that Claude 3 Opus is a powerful tool that can be used to analyze complex economic data.

The summary provided is excellent. This new capability makes it possible to streamline the lengthy process of retrieving information from videos in a much simpler form, providing new sources of information for economic research.

3.3.1 Literature Research

When writing the original version of this article (Korinek, 2023a), LLMs were of limited usefulness for searching and referencing the academic literature. When asked for citations or references, they were often able to provide standard references that are frequently cited in the literature (say, Rothschild-Stiglitz), but basic LLMs frequently hallucinated authoritative-sounding papers that do not actually exist or made other simple factual mistakes.

The recent advances in LLM-powered search have significantly improved the ability of LLMs to help with literature research, as described in Section 2.3. Moreover, dedicated websites such as `https://elicit.org` allow literature searches and compile lists of papers on the requested topics, together with LLM-generated summaries.

However, when performing literature reviews, LLMs still struggle with synthesizing the state of the art on a given research topic—they lack the deep background knowledge that an experienced scholar has regarding the relative usefulness of different papers in their field of expertise and over-emphasize papers with titles that closely match the requested literature search. Further progress will require that these tools better learn to prioritize the most impactful academic references, perhaps as measured by citation counts and journal and researcher impact factors.

3.3.2 Background Internet Research

As already observed in Section 5, LLM-powered search tools have also become powerful instruments for factual research. Search-augmented LLMs still hallucinate occasionally, but they do so much less frequently than LLMs without web access. Moreover, they offer links supporting the statements that they make, which the user can follow to judge the reliability of the sources and verify the model's output.

The following is an example:

Chat 10 (Perplexity.AI Pro): Can you explain the EU AI Act?

The EU AI Act is a groundbreaking piece of legislation that aims to regulate artificial intelligence systems within the European Union. Here's a comprehensive overview of its key aspects:

Risk-Based Approach

The AI Act classifies AI systems into four risk categories [1] [2]:

1. **Unacceptable risk**: These AI systems are prohibited outright.
2. **High risk**: Strictly regulated with specific requirements.
3. **Limited risk**: Subject to transparency obligations.
4. **Minimal risk**: Allowed with minimal restrictions.

Key Provisions

[2 pages of descriptions]

Conclusion

The EU AI Act represents the world's first comprehensive legal framework for AI, aiming to foster innovation while protecting fundamental rights and ensuring ethical AI development. It sets a global standard for AI regulation and is likely to have far-reaching impacts on AI development and deployment worldwide.

The summary it provided is satisfactory, although I had to ask a follow-up question for it to report the date when the act was passed. Note that the leading LLM-based chatbots without search (e.g., Anthropic) were not able to correctly answer the question since their training data cut off before the passage of the act.

3.4 Coding

Coding is perhaps the category that has delivered the greatest productivity gains from the advances of the past year. Tools like Copilot or Cursor (described in Section 2.2.1

can increasingly leverage LLMs to collaborate with researchers in quickly producing high-quality code, starting from the ability to plan:

Outlining a plan for a coding project Cutting-edge LLMs, in particular Claude 3.5 Sonnet, also have the ability to outline a plan for how to approach coding projects. In the spirit of chain-of-thought prompting, they frequently deliver better results if first asked to outline a plan on how to code a programming task and then executing the plan.

A significant breakthrough in recent months was OpenAI's o1 series of models, described in Section 2.1, which can reason through a coding problem before writing up the code.

3.5 Data Analysis

This category has benefited from the better ability to produce structured output and the growing accuracy of LLMs, which now rival humans in tasks like classifying text or extracting sentiment. For individual researcher use, collaborative tools for data analysis in ChatGPT and Claude are becoming increasingly powerful and user-friendly. For bulk data analysis, cost declines and better models are making it possible to process larger and larger amounts of data.

The following is a useful ability that was not described in the 2023 article:

3.5.1 Locating Data sources

LLMs can be useful starting from the first step of data analysis, i.e., in identifying and locating appropriate data sources. The following example illustrates the role that Claude 3 can play in advising users on what data sources to consider:

Chat 11 (Claude 3): What data would be useful to analyze the effects of AI on the labor market and where can I find it?

To analyze the effects of AI on the labor market, you would need data from various sources. Here are some key data points and potential sources:

1. Employment data by industry and occupation
   - Bureau of Labor Statistics (BLS) in the US
   - Eurostat for European countries
   - National statistical agencies of other countries
2. Adoption rates of AI technologies by industry

- Market research reports from companies like Gartner, IDC, or McKinsey
- Academic studies and surveys
- Company reports and press releases

[Points 3. to 9. omitted for space reasons.]
Combining data from these various sources can provide a comprehensive picture of how AI is affecting the labor market across industries, occupations, and demographic groups. It's important to note that data availability and quality may vary across countries and regions.

Although the list provided by Claude is somewhat of a laundry list and generic, it serves as a good brainstorming basis that may give the researcher ideas about novel data sources. Moreover, Claude is also happy to offer follow-up advice on how to best download data sources and how to best clean them.

3.6 Mathematical Derivations

Over the past year, there have been significant advances in this domain, even though it is not natural for general-purpose LLMs to perform mathematical derivations – as we observed in Section on page 10, they are better suited for "system 1"-type of tasks. Moreover, OpenAI's o1 model marks a breakthrough in this sphere. All in all, a growing number of mathematicians and economists report that current cutting-edge models are already quite useful at aiding in proofs, although they still need to be guided. The experience is akin to experienced researchers working with a graduate students who are fast at working through mathematical derivations but require guidance in their efforts.

3.7 Promoting Research

After writing a paper, the next step is to promote it. There are few areas in which LLMs have been as impactful as in marketing. This is also very relevant for research papers, which are highly structured and information-dense self-contained pieces of knowledge—the current generation of LLMs has become capable of processing the content of entire research papers and are excellent at transforming that content into different formats, including formats that are conducive for promotion or for conveying the message to a lay audience. Accordingly, there are several excellent use cases of LLMs in research promotion.

3.7.1 Social media posts

For the following exercise, I uploaded the original working paper version of this paper (Korinek, 2023b) to GPT-4o and asked it to generate social media posts, in this particular example for X (formerly Twitter):

Chat 12 (GPT-4o): Write a series of X posts to promote the attached paper. Use simple sentences and make them catchy.
[Attachment: w30957.pdf]

1/Exciting new possibilities! My latest NBER paper explores 25 ways language models like ChatGPT can boost research in economics. From ideation to coding, LLMs are transforming our workflows! #AI #EconResearch #LLM
2/Did you know? Large language models are now capable of automating micro-tasks like data analysis, writing, and even mathematical derivations in economic research. Check out my paper for more insights! #AIinEconomics #ProductivityBoost
3/Researchers, take note! LLMs aren't just about writing—they can also help with background research, coding, and more. Discover 6 key areas where they can assist in my latest paper. #CognitiveAutomation #AI #Economics
4/What's the future of cognitive labor? My paper dives into how LLMs could reshape the role of human economists, making us more efficient while raising new questions about the future of work. #FutureOfWork #AIRevolution
5/Want to stay ahead in research? Learn how to incorporate LLMs into your daily tasks and increase productivity. Don't miss the insights in my NBER paper! #Economics #AItools #GenerativeAI

The LLM adapts to the optimal format for different social media platforms. Upon request, it can also incorporate suitable emoticons or links.

3.7.2 Presentation slides

Successful promotion leads to invitations to present your paper. The latest generation of LLMs released in 2024 also have the capability to draft slides based on a research paper. This is a difficult exercise because it requires that the model exhibits detailed recall across a long paper.

Chat 13 (Claude 3): Can you create a 20-slide deck for the attached economic research paper?
[Attachment: w30957.pdf]

Here is a 20-slide deck summarizing the key points from the research paper "Language Models and Cognitive Automation for Economic Research" by Anton Korinek:
Slide 1: Title Slide

- Paper title: "Language Models and Cognitive Automation for Economic

Research"

- Author: Anton Korinek

Slide 2: Introduction

- Large language models (LLMs) can revolutionize economic research
- Describes 25 use cases of LLMs across 6 domains
- Provides instructions and examples for each use case
- Classifies usefulness of capabilities from experimental to highly useful

Slide 3: What are LLMs?

- LLMs are foundation models
- large deep learning models pre-trained on lots of data
- Form abstract representations and world model from training data
- Capabilities improve according to predictable scaling laws as compute and data increase
- Many capabilities are emergent, arising suddenly at specific thresholds

[Slides 4 – 18 omitted for space reasons]
Slide 19: Importance of Economic Ideas

- Economists can translate human concepts to analytic forms for AIs
- Economic ideas will influence worldview of future AI systems
- "Soon or late, it is ideas which are dangerous for good or evil" (Keynes)

Slide 20: Conclusion

- LLMs are becoming useful for automating research tasks
- Humans still have comparative advantage in generating research for now
- Important to adopt LLMs to boost productivity
- Prepare for potentially transformative long-term impacts

The generated slides provide an excellent first draft on which to build. What is impressive about the latest version of Claude that was used in the example is that it recalled the information contained in the paper in sufficient detail so that it could create 20 slides full of appropriate material without starting to forget content. It also chunked the material appropriately into the requested number of slides.

3.7.3 Blog Posts

One of the most effective ways of promoting research papers to a broader audience is to write blog posts that convey their message in simple terms. This is an easy task for modern LLMs—users can upload the underlying paper to provide the content, and the LLM can create a blog post to the user's specifications. The following is an example that builds on the prompt that generated with Anthropic Prompt Generator in Chat 4 to draft a summary of the first version of this paper (Korinek, 2023b).

Chat 14 (GPT-4o): Your task is to help draft a blog post for the Brookings Institution based on an economics research paper. The full text of the paper is attached.
The target audience for this blog post are economic researchers.
[continuing with the prompt generated in Chat 4]
[Attachment: w30957.pdf]

<**scratchpad**> [Omitted for space reasons]
—
<**outline**> [Omitted for space reasons]
—
<**blog_post_draft**>
How AI Is Reshaping Economic Research: Exploring Language Models for Cognitive Automation
Recent advances in artificial intelligence have brought transformative tools like large language models (LLMs) into the toolkit of economic researchers. These models, such as OpenAI's ChatGPT, are not just conversational agents; they represent a major leap in how cognitive tasks can be automated, promising to significantly alter the landscape of research in economics. In his recent paper, "Language Models and Cognitive Automation for Economic Research," Anton Korinek lays out 25 ways that these models can be leveraged, offering insights that are highly relevant for economic researchers looking to understand the potential impact of AI on their field.
Six Domains of LLM Utility in Economic Research
Korinek's paper categorizes the applications of LLMs into six key domains. These applications span from idea generation to coding, offering a comprehensive toolkit for researchers.
- **Ideation and Writing**: LLMs can assist researchers from the very start of a project, helping brainstorm ideas, generate counterarguments, and refine initial drafts.

[Remainder omitted for space reasons]

The draft blog posts generated by the current generation of frontier LLMs are generally well done and, after some editing, save significant time compared to having to write such a post from scratch.

3.7.4 Conducting interviews

Large language models (LLMs) have emerged as powerful tools for conducting and assisting with interviews related to academic research. Their capabilities in this domain can be leveraged in two primary ways: answering interview questions and generating insightful questions for various audiences.

LLMs can be used to answer interview questions on behalf of researchers, using their published work as a knowledge base. This approach is particularly useful when researchers are preparing responses for written interviews or need assistance in formulating clear and concise answers. For example, using a research paper as a reference, an LLM could provide detailed answers to questions about the impact of AI on economic research.

Chat 15 (Claude 3.5 Sonnet): I received an interview request on the impact of AI on economic research from The Chronicle of Higher Education. Can you help me draft responses based on the attached paper of mine? I am pasting the questions here:
1. You describe 25 use cases for large language models (LLMs) in economic research across six domains. Which of these capabilities do you think has the most potential to revolutionize how economists conduct research in the near term?
2. In the paper, you speculate on the longer-term implications of cognitive automation via LLMs for economic research. Could you elaborate on how you envision the role of human economists evolving as LLMs become more advanced?
3. You mention that LLMs can be used to simulate human subjects for research. What are the ethical considerations and potential pitfalls of this approach, and how might it change the landscape of behavioral economics?
4. The paper touches on the limitations of current LLMs, particularly in areas like mathematical derivations. How do you see these limitations being addressed in future iterations of LLMs, and what implications might this have for theoretical economics?
5. You draw parallels between the rise of LLMs and Garry Kasparov's observations about chess computers. How do you think economists should prepare for a future where AI systems might produce superior economic models? What skills should economists focus on developing?
[Attachment: w30957.pdf]

Certainly! I'd be happy to help you draft responses based on your paper. I'll provide suggested responses for each question, drawing directly from the content and ideas presented in your paper.
1. You describe 25 use cases for large language models (LLMs) in economic research across six domains. Which of these capabilities do you think has the most potential to revolutionize how economists conduct research in the near term?
Suggested response: "Based on my analysis, I believe the capabilities related to writing and data analysis have the most potential to revolutionize economic research in the near term. LLMs are particularly adept at tasks like synthesizing text, editing, and generating catchy titles or summaries. These capabilities can significantly streamline the writing process, allowing researchers to focus more on ideas and analysis rather than the mechanics of writing.
In terms of data analysis, LLMs show great promise in extracting data from text, reformatting data, and classifying and scoring text. These capabilities can automate many time-consuming tasks in data preparation and initial analysis, potentially accelerating the research process.
It's worth noting that I've rated several capabilities in these domains as 'highly useful' in my paper, indicating that incorporating them into research workflows can already save significant time and boost productivity. As these tools continue

to improve, I expect their impact on how we conduct research to grow substantially."
2. In the paper, you speculate on the longer-term implications of cognitive automation via LLMs for economic research. Could you elaborate on how you envision the role of human economists evolving as LLMs become more advanced?
Suggested response: "As LLMs become more advanced, I anticipate a shift in the role of human economists. In the medium term, I expect LLM-based assistants and tutors to become increasingly useful for generating content that makes up research papers. Human researchers will likely focus more on their comparative advantages - posing questions, suggesting directions for obtaining answers, discriminating which parts of the produced content are useful, editing, and providing feedback.
In essence, the role may evolve to be more akin to that of an advisor or research manager. Economists will likely spend more time organizing research efforts, coordinating teams, and procuring data sources.
However, in the longer term, as LLMs continue to advance, the need for human inputs, edits, and feedback may diminish. We may find ourselves increasingly just rubber-stamping the output produced by ever-more advanced LLMs. Ultimately, it's possible that sufficiently advanced AI systems may be able to produce and articulate superior economic models, potentially making the cognitive work of human economists redundant.
This potential future underscores the importance of economists focusing on tasks that leverage uniquely human traits and adapting to work alongside increasingly capable AI systems."
[Remainder omitted for space reasons.]

The LLM provided well-crafted responses that accurately reflect the perspective provided in the attached background paper from February 2023. In its response to the first question, it even employed the classical human trick of punting the exact question (asking for *one* specific capability) and focusing on a more general question that summarizes several categories of use cases.

This approach works well because the LLM can draw on the paper it was provided to generate insightful and relevant responses that align with the researcher's work and the interviewer's audience. The model maintains consistency with the original research while adapting the language to suit an interview format, making complex ideas more accessible to a broader audience.

LLMs can also be employed to generate thought-provoking and relevant interview questions. This capability can be tailored to different target audiences, enhancing the impact and reach of research promotion efforts. For instance:

Chat 16 (Claude 3.5 Sonnet): I am a journalist working for the Chronicle of Higher Education. I enjoyed reading the attached paper and scheduled an interview with the author. What would be five good questions to ask him based on the content of the paper?
[Attachment: w30957.pdf]

Based on the content of the paper, here are five good questions you could ask the author in your interview:
1. You describe 25 use cases for large language models (LLMs) in economic research across six domains. Which of these capabilities do you think has the most potential to revolutionize how economists conduct research in the near term?
2. In the paper, you speculate on the longer-term implications of cognitive automation via LLMs for economic research. Could you elaborate on how you envision the role of human economists evolving as LLMs become more advanced?
3. You mention that LLMs can be used to simulate human subjects for research. What are the ethical considerations and potential pitfalls of this approach, and how might it change the landscape of behavioral economics?
4. The paper touches on the limitations of current LLMs, particularly in areas like mathematical derivations. How do you see these limitations being addressed in future iterations of LLMs, and what implications might this have for theoretical economics?
5. You draw parallels between the rise of LLMs and Garry Kasparov's observations about chess computers. How do you think economists should prepare for a future where AI systems might produce superior economic models? What skills should economists focus on developing?
These questions cover a range of topics from the paper, including specific capabilities of LLMs, their potential impact on the field of economics, ethical considerations, current limitations, and future implications for the profession. They should provide a good basis for an in-depth discussion with the author about the key themes of his paper.

The questions—which happen to be the ones answered in Chat 15—are well-chosen and cover the most important parts of the paper. The capability of generating insightful questions is useful not only for journalists but also for economists for two key reasons.

First, economists often find themselves in situations where they are asked to suggest questions that would lead to interesting conversations targeted at specific audiences. Journalists, podcast hosts, and moderators at panel events regularly turn to their interviewees or speakers for such input. LLMs can be an invaluable tool in these scenarios, helping economists quickly generate relevant and engaging questions tailored to different audiences. For instance, the questions generated in the previous example would be markedly different from those aimed at a business newspaper, a policy think tank, or a general interest magazine. By leveraging LLMs, economists can effectively shape the discourse around their research and ensure it resonates with diverse audiences.

Second, LLMs allow researchers to prepare thoroughly for high-stakes interviews. In such cases, the prompt could be modified to generate challenging questions that the researcher might face, for example, by adding a phrase like "What would be some challenging questions that I may face?" This approach can help to anticipate difficult questions (especially if the interviewers themselves turn to LLMs) and prepare thoughtful responses in advance.

3.7.5 Podcasts

Perhaps the most impressive tools when it comes to promoting research papers is the ability of Google's NotebookLM to generate 'deep dive' podcasts. The system works by first generating an outline of the source material and identifying linkeages, revising that outline, producing a detailed version of the script, performing a round of critique and the associated modifications, and then adding disfluencies to make the conversation sound more natural. An example is reproduced in Chat 3 on page 23.

4 Conclusions

This paper has described the rapidly evolving landscape of large language models (LLMs) and their applications in economic research in Fall 2024. We explored dozens of use cases across seven key categories: ideation and feedback, writing, background research, coding, data analysis, mathematical derivations, and research promotion. The capabilities of LLMs have advanced dramatically even in the short time since the first version of this paper was published in December 2023, with new developments like improved reasoning abilities, interactive collaboration workspaces, real-time voice assistants, and powerful LLM-powered search emerging in recent months. These tools are proving ever more useful for automating many tasks in the research process, with the potential to significantly boost productivity.

Experiencing this breakneck pace of progress firsthand serves as a stark reminder that we urgently need to examine the broader economic and societal implications of ever-improving artificial intelligence systems. As economists, we have a responsibility to help prepare for the myriad policy challenges that the Age of AI will bring Korinek (2024). A simple yet crucial exercise for all researchers is to consider: how would transformative advances in AI affect your specific area of study? By engaging with this question proactively, we can better anticipate and shape the impacts of AI on our disciplines and society at large.

The rapid efficiency gains in AI as an input to research also necessitate rethinking our optimal research methodologies. As LLM capabilities grow, it becomes increasingly beneficial to incorporate these tools into our workflows. However, the pace of advances also raises questions about the optimal timing of certain research projects. Some inquiries may be better postponed, as imminent AI advances could soon make them far easier to tackle. Researchers must weigh the benefits of immediate pursuit against the potential for dramatically improved capabilities for certain tasks in the near future.

Lastly, as the production of research papers becomes cheaper and easier with AI assistance, we face an urgent challenge in evaluation. The bottleneck in the research process may shift from generation to assessment of ideas and results. This highlights the pressing need to develop robust methods for evaluating AI-augmented content, including exploring the feasibility of LLM-assisted peer review processes. While this presents significant challenges, it also offers opportunities to reimagine and improve our systems of knowledge validation.

Despite these complexities, the integration of LLMs and other AI tools into economic research also creates some cause for optimism. These technologies have the potential to dramatically accelerate the pace of discovery, allowing us to tackle increasingly complex problems and generate novel insights. As we experience this transition, our collective task is to use the power of AI responsibly, ensuring that it enhances rather than diminishes the quality and impact of our work. By embracing these tools thoughtfully and critically, we may be able to benefit from unprecedented productivity, creativity, and a greater capacity to address the world's most pressing challenges.

References

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., and et al. (2023). Frontier AI regulation: Managing emerging risks to public safety. *arXiv:2307.03718*.

Ash, E., Hansen, S., and Muvdi, Y. (2024). Large language models in economics. *CEPR Discussion Paper DP19479*.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., and others (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv:2212.08073*.

Bsharat, S. M., Myrzakhan, A., and Shen, Z. (2024). Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., and Stoica, I. (2024). Chatbot arena: An open platform for evaluating LLMs by human preference.

Dathathri, S., See, A., Ghaisas, S., Huang, P.-S., McAdam, R., Welbl, J., Bachani, V., Kaskasoli, A., Stanforth, R., Matejovicova, T., Hayes, J., Vyas, N., Al Merey, M., Brown-Cohen, J., Bunel, R., Balle, B., Cemgil, T., Ahmed, Z., Stacpoole, K., Shumailov, I., Baetu, C., Gowal, S., Hassabis, D., and Kohli, P. (2024). Scalable watermarking for identifying large language model outputs. *Nature*, 634:818–823.

Dell, M. (2024). Deep learning for economists. *Journal of Economic Literature*, forthcoming.

Dell'Acqua, F., McFowland III, E., Mollick, E., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Working Paper*, (24-013).

Heikkilä, M. (2024). How ChatGPT search paves the way for AI agents: Here's what needs to be done to make AI assistants truly helpful. *MIT Technology Review*. `https://www.technologyreview.com/`, accessed on November 10, 2024.

Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., Atkinson, D., Thompson, N., and Sevilla, J. (2024). Algorithmic progress in language models.

Horton, J., Filippas, A., and Horton, R. (2024). EDSL: Expected parrot domain specific language for AI powered social science. Whitepaper, Expected Parrot.

Huang, J. and Chang, K. C.-C. (2023). Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics (ACL)*, page 15.

Jones, N. (2024). In awe: Scientists impressed by latest ChatGPT model o1. *Nature*, 634:275–276.

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux, New York.

Korinek, A. (2023a). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317.

Korinek, A. (2023b). Language models and cognitive automation for economic research. *NBER Working Paper*, 30957.

Korinek, A. (2024). Economic policy challenges for the Age of AI. *NBER Working Paper*, w32980.

Korinek, A. and Vipra, J. (2025). Concentrating intelligence: Scaling and market structure in artificial intelligence. *Economic Policy, forthcoming.*

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. (2024). The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv:2408.06292*.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. (2024). GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv:2410.05229*.

OpenAI (2024). Learning to reason with LLMs. *OpenAI blog, https://openai.com/index/learning-to-reason-with-llms*. Accessed: 2024-10-08.

Peng, S., Kalliamvakou, E., Cihon, P., and Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot.

Plaat, A., Wong, A., Verberne, S., Broekens, J., van Stein, N., and Bï¿œck, T. (2024). Reasoning with large language models, a survey. *arXiv:2407.11511*.

Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., Winter, C., Arnold, M., hEigeartaigh, S. O., Korinek, A., Anderljung, M., Bucknall, B., Chan, A., Stafford, E., Koessler, L., Ovadya, A., Garfinkel, B., Bluemke, E., Aird, M., Levermore, P., Hazell, J., and Gupta, A. (2023). Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives.

Shorten, C., Pierse, C., Smith, T. B., Cardenas, E., Sharma, A., Trengrove, J., and van Luijt, B. (2024). StructuredRAG: JSON response formatting with large language models.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E. H., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Wu, T., Lan, J., Yuan, W., Jiao, J., Weston, J., and Sukhbaatar, S. (2024). Thinking llms: General instruction following with thought generation.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. (2024). Tree of thoughts: Deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NeurIPS '23.

Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv:2307.15043*.