DISCUSSION PAPER SERIES

DP17923

LANGUAGE MODELS AND COGNITIVE AUTOMATION FOR ECONOMIC RESEARCH

Anton Korinek

LABOUR ECONOMICS AND MACROECONOMICS AND GROWTH

CEPR

ISSN 0265-8003

LANGUAGE MODELS AND COGNITIVE AUTOMATION FOR ECONOMIC RESEARCH

Anton Korinek

Discussion Paper DP17923
Published 20 February 2023
Submitted 11 February 2023

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

LANGUAGE MODELS AND COGNITIVE AUTOMATION FOR ECONOMIC RESEARCH

Abstract

Large language models (LLMs) such as ChatGPT have the potential to revolutionize research in economics and other disciplines. I describe 25 use cases along six domains in which LLMs are starting to become useful as both research assistants and tutors: ideation, writing, background research, data analysis, coding, and mathematical derivations. I provide general instructions and demonstrate specific examples for how to take advantage of each of these, classifying the LLM capabilities from experimental to highly useful. I hypothesize that ongoing advances will improve the performance of LLMs across all of these domains, and that economic researchers who take advantage of LLMs to automate micro tasks will become significantly more productive. Finally, I speculate on the longer-term implications of cognitive automation via LLMs for economic research.

JEL Classification: A10, B41, J23, O30

Keywords: N/A

Anton Korinek - anton@korinek.com
University of Virginia, Brookings, GovAI and CEPR

Acknowledgements
Financial support from Brookings and from Longview Philanthropy is gratefully acknowledged. I thank Julian Hazell, Sid Srinivasan, and participants at several seminars for helpful conversations on the topic, Max Schnidman, Don Suh and Natasha Swindle for excellent research assistance, and GPT-3 and Claude for inspiration and editorial assistance.

Language Models and Cognitive Automation for Economic Research

by Anton Korinek*

February 10, 2023

Abstract

Large language models (LLMs) such as ChatGPT have the potential to revolutionize research in economics and other disciplines. I describe 25 use cases along six domains in which LLMs are starting to become useful as both research assistants and tutors: ideation, writing, background research, data analysis, coding, and mathematical derivations. I provide general instructions and demonstrate specific examples for how to take advantage of each of these, classifying the LLM capabilities from experimental to highly useful. I hypothesize that ongoing advances will improve the performance of LLMs across all of these domains, and that economic researchers who take advantage of LLMs to automate micro tasks will become significantly more productive. Finally, I speculate on the longer-term implications of cognitive automation via LLMs for economic research.

Keywords: artificial intelligence, large language models, cognitive automation

*Anton Korinek is a Professor at the University of Virginia, David M. Rubenstein Fellow at the Brookings Institution, the Economics of AI Lead at the Centre for the Governance of AI (GovAI), a Research Affiliate of the NBER and CEPR, and a Senior Advisor of the Partnership on AI. Financial support from Brookings and from Longview Philanthropy is gratefully acknowledged. I thank Julian Hazell, Sid Srinivasan, and participants at several seminars for helpful conversations on the topic, Max Schnidman, Don Suh and Natasha Swindle for excellent research assistance, and GPT-3 and Claude for inspiration and editorial assistance.

1 Introduction

Recent advances in large language models (LLMs) have the potential to revolutionize research in economics and other scientific disciplines. LLMs have just crossed the threshold to become useful across a wide range of cognitive tasks – as was illustrated by the viral reception of ChatGPT, a version of OpenAI's GPT3.5 model that was released on November 28th, 2022, gained more than 100m users in its first two months, and is now estimated to produce a volume of text every 14 days that is equivalent to all the printed works of humanity (Thompson, 2023). Google and Microsoft are about to give users of their search engines and office suites access to powerful LLMs.

The objective of this article is twofold. First, it describes 25 use cases of modern LLMs to interested economic researchers, based on my own recent exploration of the space.[1] I have categorized the use cases that I experimented with into six domains: ideation, writing, background research, data analysis, coding, and mathematical derivations. I provide general instructions for how to take advantage of each of these capabilites and demonstrate them using specific examples. Moreover, I attempt to classify each LLM capability from experimental to highly useful. (A summary is provided in Table 1 on page 29.) My hope is that this description will enable other researchers to take advantage of the rapidly growing capabilities of LLMs. At present, I view LLMs to be most useful as assistants that can automate small "micro tasks" that researchers engage in numerous times during the day but that are too small to be assigned to human research assistants. LLMs are useful for such tasks because of their high speed and the low transaction cost. Moreover, LLMs are also useful as tutors in coding and data analysis tasks as well as in ideation and writing. I posit that researchers can significantly increase their productivity by incorporating LLMs into their workflow.

Second, studying the current capabilities of LLMs is useful because it foreshadows what future generations of LLMs will be able to do. In recent years, the amount of compute (computational power) employed in training cutting-edge LLMs has doubled on average every six months, delivering rapid increases in capabilities. There is widespread anticipation that these advances will continue in the near future, and that systems even more powerful than the ones discussed in this article will be released soon. It is useful for researchers to familiarize themselves even with experimental capabilities because LLMs are advancing so rapidly. In the longer term, I hypothesize that LLMs may usher in an era of cognitive automation that may have profound implications for scientific progress in economics and other disciplines. Additionally, such cognitive automation may also have stark effects on the value of cognitive labor.

The reactions to the release of the latest generation of LLMs, such as ChatGPT, have been sharply divided: One camp of commentators label LLMs as nothing but "stochastic parrots" (Bender et al., 2021) or "advanced autocomplete." Another camp equates ChatGPT with the advent of artificial general intelligence (AGI), i.e., artificial

[1] Please email me at anton@korinek.com to suggest additional use cases that I may incorporate.

intelligence that possesses human-level intelligence across all domains. One of the reasons for such divergent views is that the capabilities or "intelligence" of LLMs are so different from human intelligence, making it hard for humans to relate and to compare. I therefore want to express two warnings before I proceed:

1. It is easy – and dangerous – to overestimate the capabilities of LLMs. Current LLMs can product text in a style that sounds highly authoritative – even when they "hallucinate," i.e., when the content is completely wrong. In human-written texts, there is a strong correlation between authoritative style and insightful content, but LLMs have learned the former without being reliable on the latter. This can trick unknowing readers into believing false content. In some ways, the capabilities of LLMs feel alien to humans. Their primary objective is to generate text; their creators are still working on ensuring that the content is consistently truthful and appropriate.

2. It is easy – and dangerous – to underestimate the capabilities of LLMs. Since they regularly hallucinate and make blatant mistakes, it is easy to dismiss LLMs. However, a former Chairman for Mensa International reports that ChatGPT has a tested IQ of 147 (99.9th percentile) on a verbal-linguistic IQ test (Thompson, 2023). Moreover, whereas the level of human intelligence is relatively static, LLMs are advancing rapidly, becoming more accurate and powerful with every new iteration.

Ultimately, I believe that the most useful attitude towards the current generation of LLMs is to heed the lessons of comparative advantage that Ricardo taught us two centuries ago: LLMs increasingly have comparative advantage in generating content; humans currently have comparative advantage in evaluating and discriminating content. LLMs also have super-human capabilities in processing large amounts of text. All this creates ample space for productive collaboration, as we will explore throughout the remainder of the paper.

Section 2 observes that LLMs are a category of foundation models, which represent a new paradigm in artificial intelligence. Foundation models are vast deep learning models that are pre-trained on large amounts of data to create a foundation that can then be adapted for different applications. LLMs are capable of learning the structure of their training data and forming higher-level abstract representations of concepts. LLMs have been improving according to predictable scaling laws as a function of the amount of computation, parameter count, and size of training data employed. This has led to a rapid rise in the capabilities of LLMs, many of which are emergent, i.e., they are not present in smaller models but suddenly emerge once a certain threshold is crossed. Modern LLMs are, in some ways, starting to blur the difference between the cognitive capabilities of humans and AI systems.

In Section 3, I lay out six different areas in which LLMs can be useful. In the process of ideation, LLMs can help to brainstorm, evaluate ideas, and provide counterarguments.

In writing, they can synthesize text, provide examples, edit and evaluate text, and generate catchy tweets or titles for a paper. In background research, they can be useful for searching and summarizing the literature, translating text, explaining concepts, and formatting references. LLMs are also very capable in coding, writing code based on instructions in natural language, explaining code, translating code between programming languages, and even debugging code. For data analysis, LLMs can extract data from text, reformat data, classify text, extract sentiment, and even simulate humans to generate data. Finally, LLMs are starting to display emergent capabilities in mathematical derivations, starting from setting up models and working through derivations to explaining models. At the end of the section, Table 1 provides a systematic overview of all the described use cases and my rating of their usefulness as of Feb 1, 2023.

In the final section, I speculate on the medium- and long-run implications of advances in LLMs for cognitive labor. I hypothesize that in the medium term, LLM-based assistants will become increasingly useful for generating more and more of the content that makes up research papers, while human researchers will focus on their comparative advantage, i.e., organizing research projects, prompting, and evaluating generated content. In the long term, AI systems may be able to produce and articulate superior economic models by themselves.

2 What Are LLMs?

2.1 Foundation Models As a New Paradigm

LLMs are a category of foundation models, which can be regarded as the new paradigm in artificial intelligence of the 2020s (Bommasani et al., 2021). Foundation models are large deep learning models, with parameter counts in the order of 10^{11} and growing. They are pre-trained on abundant data to create a foundation that can then be adapted for different applications via a process called fine-tuning. For example, an LLM can be fine-tuned to act as a chatbot (such as ChatGPT) or as system that generates computer code (such as Codex). As of early 2023, some of the cutting-edge LLMs are OpenAI's GPT-3.5, DeepMind's Chinchilla, Google's PaLM and LaMDA and Anthropic's Claude.[2]

The pre-training of foundation models uses massive amounts of compute and data in a process called self-supervised learning, whereby the model learns the structure inherent in the training data by successively predicting parts of the data that are masked. For example, to train an LLM, a model is fed text fragments with some words masked, and the model learns to predict what the missing words are. This process is performed on terabytes of data from Wikipedia, scientific articles, books, and other sources on the

[2]In the space of image generation, the leading foundation models are OpenAI's DALL-E, Midjourney, and Stable Diffusion. Some recent models like DeepMind's Gato combine multiple modalities, with useful applications, e.g., in robotics.

internet.

To predict the structure of its training data in a loss-minimizing way, the model needs to learn syntactic structures, relationships between words and the concepts they represent, the context of sentences and how different words might interact in that context, and how different sentences are related to each other. For example, the system learns that "she loves cats and dogs" refers to animal-lovers whereas "it's raining cats and dogs" refers to precipitation. During the training process, the model forms increasingly higher-level abstract representations of concepts and their relationships – in short, it develops an internal world model. Based on that world model, the foundation model can be fine-tuned for different applications.[3]

2.2 Scaling and Emergent Capabilities

What distinguishes foundation models and by extension LLMs from earlier generations of deep learning models is that their scale gives rise to increasingly broad capabilities. The deep learning models of much of the 2010s displayed powerful capabilities in specific applications such as recognizing images, but there remained a category difference between the broad capabilities of humans and the narrow capabilities of specific AI systems. That difference is starting to blur with the latest generation of LLMs, which display an increasingly broad range of capabilities.[4]The overall performance of LLMs improves according to predictable scaling laws, i.e., empirical regularities that have held for several generations of machine learning models. The scaling laws observe that the goodness-of-fit of LLMs, as measured by their log-loss, improves according to a power law function of the amount of "training compute," i.e., the number of computations performed to train the model, as well as of the parameter count and size of training data (Kaplan et al., 2020). These laws also imply that it is optimal to use increases in compute for scaling the parameter counts and the size of the training data of LLMs in approximately equal proportions (Hoffmann et al., 2022).

Over the past decade, the training compute of top-end deep learning models has doubled on average every six months, implying a thousand-fold increase every five years (Sevilla et al., 2022). This trend is also behind the rapid rise in the capabilities of LLMs and other foundation models in recent years. By some measures, today's LLMs rival the human brain in their complexity, making it perhaps unsurprising that they are starting to exhibit eerily similar capabilities (Carlsmith, 2020).

As the log-loss of LLMs continuously improves, new capabilities arise at discrete thresholds. Many of the capabilities of LLMs are emergent – in the sense that they are not

[3]At the risk of oversimplifying, a useful analogy may be that the pre-training of a foundation model can be compared to a liberal-arts education that provides a useful general background, whereas the fine-tuning can be compared to a graduate education that instills specific skills.

[4]There is a philosophical debate ongoing on whether LLMs display true understanding or are merely "stochastic parrots," as Bender et al. (2021) have argued. Ultimately, this debate goes back to Turing (1950)'s famous question 'Can machines think?' For our purposes in this article, we limit our focus to how LLMs can be employed to automate research tasks.

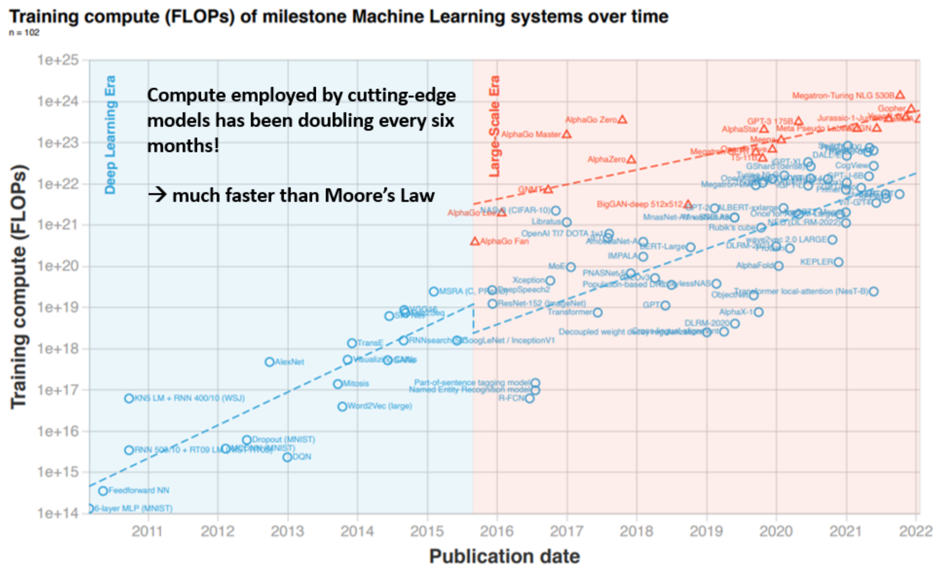


Figure 1: Training compute of cutting-edge ML systems over time
(Copyright (c) by Sevilla et al., 2022, under a CC BY 4.0 license.)

present in smaller models, suddenly emerge once a certain threshold is crossed, then improve quickly, and eventually mature. For example, Wei et al. (2022a) report that once a certain threshold of training compute is crossed, LLMs almost predictably develop the ability to perform arithmetic computations, to unscramble words, or to perform Q&A. Other significant capabilities that have emerged from language models include coding, translation, and rhyming. In fact, most of the useful capabilities for researchers that we document below have emerged only in recent years. An interesting phenomenon about many of these emergent capabilities is that they regularly surprise the creators of the systems – at the risk of anthropomorphizing, perhaps just like excellent students surprise their teachers. Many of the capabilities that emerge are discovered by chance after the systems have been released. This suggests that there may in fact be significant capabilities overhang, i.e., that many LLMs exhibit greater capabilities than what is known.

3 Applications

This section demonstrates use cases of LLMs in economic research classified along six domains: ideation, writing, background research, coding, data analysis, and mathematical derivations. For each domain, I will provide a general description and a few specific use cases for how to take advantage of LLM capabilities. I have attempted to refrain from cherry-picking and illustrate both the capabilities and failures of the LLMs I explored to provide a balanced version of their capabilities as of early 2023.

Unless otherwise stated, I am using the leading publicly available system at the time of writing, GPT-3 (`text-davinci-003`), which is slightly more powerful than ChatGPT but generates similar output. Interested users can register and access the system via a simple web interface at `https://platform.openai.com/playground`. To maximize reproducibility, I set the "Temperature" parameter of the model to 0, which makes the responses provided by the system deterministic. Let me describe a few important limitations of this particular LLM that potential users should be aware of. The system is trained on data that cuts off in 2021, so it has no knowledge of more recent events. It cannot access the Internet – the text it generates is based solely on the parameters acquired during the training process. Moreover, it has no memory so information does not carry over from one session to the other. The amount of text that it can process is limited to 4000 tokens per request, corresponding to about 3000 words, with the limit applying to the sum of the user prompt and the completion. Furthermore, note that the results generated by an LLM change depending on the prompt – even small changes in prompts, such as different spacing or punctuation, can lead to completely different outputs. This makes it important for users to experiment with different prompts and to learn how to optimally engineer their prompts. Finally, let me add a reminder that ultimate responsibility for any output produced by LLMs always rests with the human user.

One common theme in all the applications I will demonstrate is that LLMs exhibit such quick response times and low transaction cost that they are useful for outsourcing micro tasks in which they are still error-prone and in which similarly capable human research assistants would not be competitive. For example, I would not resort to human research assistance for micro-tasks such as spelling out the first-order conditions of an optimization problem while I am writing a paper – the associated delay would be too large. But the instantaneous response of LLMs makes it useful to outsource this micro task, even if there are occasional mistakes. Similarly, I would not hire a human research assistant who regularly commits basic logical fallacies while presenting their results with great confidence – I would consider them too unreliable. But after a short adjustment period, I have found it useful to incorporate LLMs that do precisely that into my workflow.

3.1 Ideation

One part of research for which LLMs are increasingly useful is the process of ideation, i.e., generating, selecting, and developing ideas. They can help both as assistants and as tutors. This is illustrates vividly how LLMs differ from earlier applications of deep learning in economics – they display a form of creativity that had long been reserved for humans. Although the capabilities of cutting-edge LLMs in ideation are impressive, they also have clear limitations. Some of these derive from the fact that I am experimenting with models that are not explicitly fine-tuned to economics, but some of them are rooted in the overall limitations of LLMs at the time of writing.

3.1.1 Brainstorming

Since LLMs are trained on a vast amount of data that represents a cross-section of all human knowledge, they are quite useful in brainstorming (or, perhaps more aptly, *net*-storming) ideas and examples related to a defined theme.

The following prompt illustrates a simple example. Throughout the remainder of this section, I will present all examples generated by LLMs in boxes, with the prompt that I entered in Sans Serif style and the system's generated response `in typewriter style`. For example, in the box below, the first sentence was my prompt, and the numbered bullet points below are what the system generated.

Please brainstorm economic channels through which advances in AI may increase inequality

```
1.  Automation of low-skill jobs, leading to job displacement and increased
unemployment
2.  Increased demand for highly-skilled workers, leading to wage inequality
3.  Increased access to capital for those with the resources to invest in AI
technology
4.  Increased surveillance of workers, leading to decreased privacy and
potential exploitation
5.  Increased use of AI-driven algorithms to make decisions about hiring,
lending, and other economic activities, leading to potential bias and
discrimination
6.  Increased use of AI-driven algorithms to optimize pricing, leading to
increased costs for consumers
7.  Increased use of AI-driven algorithms to optimize marketing, leading to
increased costs for businesses
8.  Increased use of AI-driven algorithms to optimize investments, leading
to increased returns for those with the resources to invest
9.  Increased use of AI-driven algorithms to optimize tax avoidance, leading
to decreased tax revenues for governments
10.  Increased use of AI-driven algorithms to optimize labor costs, leading
to decreased wages for workers
```

The channels listed by the model are not path-breaking, but they are on-topic, largely reasonable, and span a wide range of what I could think of plus some more. One can also observe a tendency that is all-too-familiar from human brainstorming – after about point 5., creativity declines and the responses become somewhat repetitive.

3.1.2 Evaluating ideas

They can also evaluate different ideas, in particular by providing pros and cons of different research plan. As the following example illustrates, the model I tested was programmed to refrain from providing strong opinions, but the example – with a prompt that is admittedly a bit silly – showcases how LLMs can be a useful input in deliberations about research directions.

I am writing a paper on the effects of AI on inequality. What do you think is more useful? A paper on how AI increases inequality or a paper on how AI reduces inequality?

```
That depends on the purpose of your paper. If you are looking to explore
the potential implications of AI on inequality, then a paper on how AI
increases inequality would be more useful. This would allow you to explore
the potential negative impacts of AI on inequality, and how it could
potentially exacerbate existing disparities. On the other hand, if you
are looking to explore potential solutions to inequality, then a paper on
how AI reduces inequality would be more useful. This would allow you to
explore the potential positive impacts of AI on inequality, and how it could
potentially reduce existing disparities. Ultimately, the choice of which
paper to write depends on the purpose of your paper.
```

As the model's response suggests, looking at how AI may increase inequality is more useful for positive work, whereas how it may decrease inequality is more useful for normative work.

3.1.3 Providing counterarguments

No matter what point we are arguing, there are always counterarguments. LLMs do not care about which side of an argument they are on – they are just as good at providing arguments in favor of a given point as they are for the counterarguments. This may be helpful to counteract the confirmation bias common to our human brains. The following is an example:

My hypothesis is that AI will increase inequality. What are the main counterarguments?

```
1. AI can be used to increase access to education and healthcare, which can
help reduce inequality.
2. AI can be used to automate mundane tasks, freeing up people to pursue
more meaningful work.
3. AI can be used to increase efficiency and productivity, leading to
higher wages and better job opportunities.
4. AI can be used to create new jobs and industries, providing more
opportunities for people to earn a living.
5. AI can be used to improve decision-making processes, leading to fairer
outcomes.
6. AI can be used to identify and address systemic biases, leading to more
equitable outcomes.
7. AI can be used to reduce the cost of goods and services, making them
more accessible to people with lower incomes.
8. AI can be used to improve the accuracy of predictions, leading to more
equitable outcomes.
```

Some of the counter-arguments are better and some are worse, but the output covers the main points that I am aware of – I consider points 3., 4., 6., and 7. among the most important ones.

3.2 Writing

The core competence of LLMs is to generate text. This implies that they are quite capable and useful for many tasks related to writing, from synthesizing text based on bullet points, changing the style of text, editing text, evaluating styles, and generating titles, headlines, and tweets. Let us cover each in turn.

3.2.1 Synthesizing text

One of the most useful capabilities of LLMs may be to translate rough bullet points into well-structured sentences that are clear and easy to read. The following example describes a typical use case for this. Note that it also asks the system to provide examples and offer solutions:

Please write a paragraph that integrates the following arguments and goes beyond. Include a topic sentence.
* large language models automate cognitive tasks
* the work of economists is largely cognitive [add examples]
* we economists should therefore worry about recent progress in language models
* in the short run, these models can help us be more productive
* in the long run, there is a good likelihood that they will displace human economists
* we should think how to prepare [offer solutions]

```
The recent progress in large language models has the potential to automate
many cognitive tasks, including those of economists. Economists are
responsible for analyzing data, making predictions, and providing policy
advice, all of which are cognitive tasks. Therefore, economists should
be concerned about the implications of these language models, both in the
short and long run. In the short run, these models can help us be more
productive, but in the long run, there is a good likelihood that they will
displace human economists. To prepare for this eventuality, economists
should think about how to use these models to their advantage and how to
develop skills that are not easily automated. This could include developing
expertise in areas such as communication, problem-solving, and critical
thinking.
```

This allows researchers to concentrate their energy on the ideas in their text as opposed to the writing process itself. The text is generally high-quality (although some observe that it is, naturally, a bit sterile and lacks the idiosyncracies and elements of surprise that characterize human writing – a feature that detectors of LLM-written text zero in on[5]).

[5]The most widely used detector for GPT-3.5-written text is GPTZero, which is available at

LLMs are also responsive to requests to write in a particular style. For example, the example above will read differently if we add "Write in an academic style", "colloquial style", "style that is comprehensible to non-economists" or "in the style of [enter your favorite politician]." When I added the sentence "Write in the style of a gangster." the system started with "Yo, economists, it's time to wake up and smell the coffee." Alternatively, when I wrote "Can you add a joke?" it included the following as the penultimate sentence: "After all, we don't want to be replaced by a machine that can't even tell a good joke!" Of perhaps greater relevance for our daily writing, LLMs can also write text in LaTeX format, which automatically surrounds variables by the $ signs required for LaTeX, and it can draft beamer slides in LaTeX code based on bullet points.

Writing is of course a skill that is not only useful for research - cognitive workers in many areas report using, e.g., ChatGPT to draft emails, memos, performance reviews, or summaries based on a few bullet points. Likewise, professors report generating the first draft of letters of recommendation or syllabi for standard courses using LLMs. The writing tool `https://lex.page` incorporates several related functions in an easy-to-use web interface.

3.2.2 Editing text (for mistakes, style, clarity, simplicity, ...)

Another useful capability is editing. LLMs can edit text for grammatical or spelling mistakes, style, clarity, or simplicity. This set of capabilities are perhaps most useful for non-native speakers who want to improve their writing. In the following example, I have bolded each word with a mistakes as well as the corrections made by the system for emphasis:

Can you correct the following sentence?
These paper **describe** six **domain** in which large language models (LLMs) **displays emergant capbilities** that are useful **to the** economic research.

```
This paper describes six domains in which large language models (LLMs)
display emergent capabilities that are useful for economic research.
```

Can you explain the main corrections you made?

```
I changed "These" to "This" to make the sentence singular, and I changed
"useful to" to "useful for" to make the phrase more grammatically correct.
```

As illustrated by the follow-up question, the system can even even explain its edits afterwards so that students can learn from any corrections it has made. Another useful application is that the system can convert text to be comprehensible to readers of different levels, for example, using a prompt like "Rewrite the following text so that it is comprehensible to an eight-year-old" or "comprehensible to an eighth-grader."

`https://gptzero.me/`.

3.2.3 Evaluating text

LLMs can also evaluate text for its style, its clarity, or similar. The following question about a draft abstract of this paper turned out to be a bit of a deflating experience:

What are the main stylistic shortcomings of the following text?
This paper describes six domains in which large language models (LLMs) display emergent capabilities that are useful for economic research: ideation, writing, background research, data analysis, coding, and mathematical derivations. I provide general instructions as well as specific use cases for how to take advantage of each of these and classify the LLM capabilities that are displayed on a range from "experimental" to "highly useful." I hypothesize that ongoing advances will improve the performance of LLMs across all of these domains, and that economic researchers who take advantage of LLMs to automate micro tasks will become significantly more productive. Finally, I speculate on the longer-term implications of cognitive automation via LLMs for economic research.

```
The main stylistic shortcomings of the text are that it is overly wordy and
lacks a clear structure. The sentences are too long and complex, making
it difficult to follow the author's argument. Additionally, the text lacks
transitions between ideas, making it difficult to follow the flow of the
argument. Finally, the text does not provide any concrete examples or
evidence to support the author's claims.
```

Can you rewrite the paragraph to correct for these shortcomings?

```
This paper examines six domains in which large language models (LLMs) can
be used for economic research: learning, ideation, writing, data analysis,
coding, and mathematical derivations. I provide instructions and use cases
for each domain, and classify the LLM capabilities from ‘‘experimental’’ to
‘‘highly useful.’’ I hypothesize that LLMs will continue to improve in all of
these areas, and that economic researchers who use them to automate micro
tasks will benefit from increased productivity. Finally, I discuss the
implications of cognitive automation via LLMs for economic research.
```

After some deliberation, I agree with all the shortcomings identified by the LLM. The rewritten version alleviates some of them but does not do so perfectly. The system can also respond to question such as "Which argument is the hardest to follow in the following text?"

3.2.4 Generating catchy titles & headlines

Another useful application is to come up with catchy headlines or paper titles, as illustrated by the following example based on the abstract of a recent paper of mine (Korinek and Juelfs, 2022):

Please suggest two or three catchy titles for a paper with the following abstract:
We analyze how to set up institutions that future-proof our society for a scenario of ever-more-intelligent autonomous machines that substitute for human labor and drive down wages. We lay out three concerns arising from such a scenario, culminating in the economic redundancy of labor, and evaluate recent predictions and objections to these concerns. Then we analyze how to allocate work and income if these concerns start to materialize. As the income produced by autonomous machines rises and the value of labor declines, we find that it is optimal to phase out work, beginning with workers who have low labor productivity and job satisfaction, since they have comparative advantage in enjoying leisure. This is in stark contrast to welfare systems that force individuals with low labor productivity to work. If there are significant wage declines, avoiding mass misery will require other ways of distributing income than labor markets, whether via sufficiently well-distributed capital ownership or via benefits. Recipients could still engage in work for its own sake if they enjoy work amenities such as structure, purpose, and meaning. If work gives rise to positive externalities such as social connections or political stability, or if individuals undervalue the benefits of work because of internalities, then there is a role for public policy to encourage work. However, we conjecture that in the long run, it would be more desirable for society to develop alternative ways of providing these benefits.

```
1.  "Future-Proofing Society:  Preparing for Autonomous Machines and the
Decline of Labor"
2.  "The End of Work?  Navigating the Impact of Autonomous Machines on
Labor"
3.  "The Impact of Autonomous Machines on Labor:  How to Allocate Work and
Income"
```

All three suggested titles would fit the described work quite well.

3.2.5 Generating tweets to promote a paper

The following exercise is similar but converts the abstract I pasted into five tweetable chunks that would work well for Twitter:

Write five tweets that summarize the following text. Write simple sentences and make them as catchy as possible:
We analyze how to set up institutions [...same abstract as above...]

```
1. Worried about the future of work? We analyze how to set up
institutions to future-proof our society. #FutureOfWork #Automation
#IncomeRedistribution
2. As autonomous machines rise and labor value declines, it's optimal
to phase out work starting with those with low labor productivity.
#FutureOfWork #Automation #IncomeRedistribution
3. If wage declines occur, avoiding mass misery will require other ways
of distributing income than labor markets. #FutureOfWork #Automation
#IncomeRedistribution
4. If work gives rise to positive externalities, there's a role for public
policy to encourage work. #FutureOfWork #Automation #IncomeRedistribution
5. In the long run, it's more desirable to develop alternative ways of
providing benefits. #FutureOfWork #Automation #IncomeRedistribution
```

We can only speculate on what this capability will imply for the future of content on Twitter. Goldstein et al. (2023) provide an overview of the resulting threats and potential mitigations.

3.3 Background Research

There are a variety of tasks that researchers perform in order to conduct background research and learn about both existing and new work in their reserach area. These include summarizing text, searching relevant literature, seeking to understand new concepts, and translating text from other languages.

3.3.1 Summarizing Text

First, LLMs are very capable at summarizing. They are usually adept at extracting the most relevant content from a passage of text, as illustrated here:

Please summarize the following text in one sentence:
Large language models (LLMs) have the potential to revolutionize economic research. This paper describes six domains in which LLMs are starting to become useful: ideation, background research, writing, data analysis, coding, and mathematical derivations. I provide general instructions and demonstrate specific use cases for how to take advantage of each of these and classify the LLM capabilities that are displayed on a range from "experimental" to "highly useful." I hypothesize that ongoing advances will improve the performance of LLMs across all of these domains, and that economic researchers who take advantage of LLMs to automate micro tasks will become significantly more productive. Finally, I speculate on the longer-term implications of cognitive automation via LLMs for economic research.

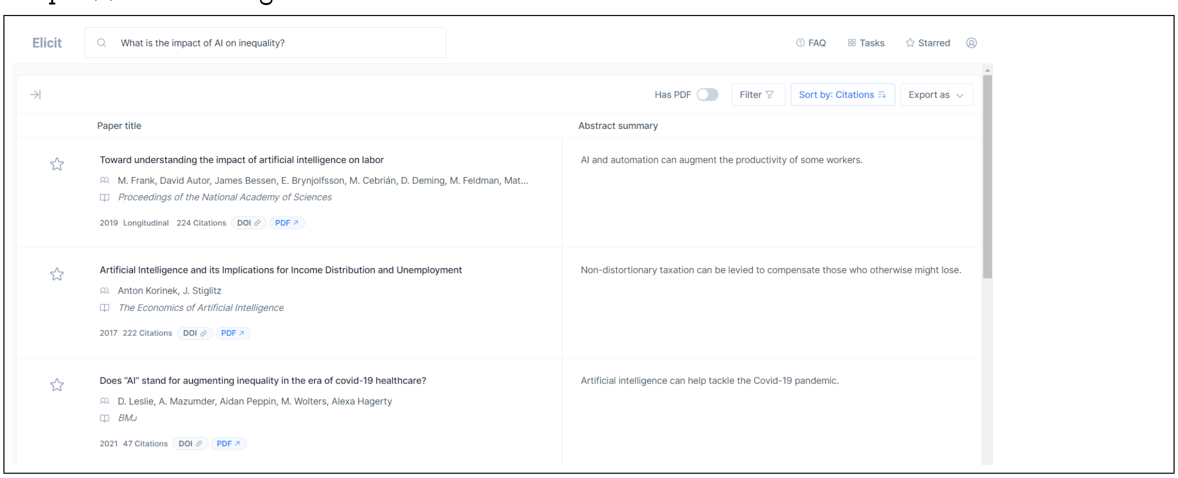
```
This paper discusses the potential of large language models to revolutionize
economic research by providing general instructions and demonstrating
specific use cases for how to take advantage of six domains in which LLMs
are becoming useful.
```

The summary sentence it provided covers all the important bases. Summaries of longer texts are limited by the transformer technology on which the systems are based. As already mentioned, for GPT3.5, the amount of text that can be processed is limited to 4000 tokens or about 3000 words.

3.3.2 Literature Research

LLMs are of limited usefulness for searching the literature. When asked for citations, there is a risk that they may hallucinate and make up authoritative-sounding papers that do not actually exist. They are frequently able to provide standard references that are frequently cited in the literature, but even so, all references they provide should always be double-checked.

On the other hand, a tool that I found useful when exploring new areas of literature that I am not familiar with – and that only reports paper that exist – is available at `https://elicit.org`:



3.3.3 Formatting References

Once references are found and it is verified that they are not hallucinated, LLMs are very capable of formatting them in the desired manner:

Please convert the following reference to bibtex format:
Anton Korinek (2023), Large Language Models and Cognitive Automation for Economic Research, NBER Working Paper 30957, February 2023.

```
@techreport{Korinek2023,
title={Large Language Models and Cognitive Automation for Economic
Research},
author={Anton Korinek},
year={2023},
institution={National Bureau of Economic Research},
number={30957},
month={February},
}
```

Now format it in APA style:

```
Korinek, A. (2023).  Large language models and cognitive automation for
economic research.  National Bureau of Economic Research Working Paper,
30957.
```

I have omitted it for space reasons, but it goes without saying that the system can convert a whole bibliography in batch format, for example from APA to Chicago format or into bibtex format (subject to the limit of about 3000 words per request). In the example above, notice that the LLM even knew what the abbreviation "NBER" stands for.

For highly-cited works that the LLM has frequently encountered in its training dataset, a prompt like "bibtex reference for stiglitz weiss" also works well and saves the user from having to copy or type the detailed citation information of the work in question. However, to be sure, I would double-check the information for all but the most highly cited papers – the system blatantly made up articles and citation information when I asked it to generate bibtex references for works with lower citation counts.

3.3.4 Translating Text

Jiao et al. (2023) demonstrate that LLMs such as ChatGPT perform competitively with commercial translation products on high-resource European languages. The following example using GPT3.5 originates from a demonstration I gave at the Bank of Finland in January 2023. According to my human hosts, the translation was impressive.

Please translate the following abstract into Finnish:
Large language models (LLMs) have the potential to revolutionize economic research. This paper describes six domains in which LLMs are starting to become useful: ideation, background research, writing, data analysis, coding, and mathematical derivations. I provide general instructions and demonstrate specific use cases for how to take advantage of each of these and classify the LLM capabilities that are displayed on a range from "experimental" to "highly useful." I hypothesize that ongoing advances will improve the performance of LLMs across all of these domains, and that economic researchers who take advantage of LLMs to automate micro tasks will become significantly more productive. Finally, I speculate on the longer-term implications of cognitive automation via LLMs for economic research.

```
Suuret kielimallit (LLM:t) voivat mullistaa taloustieteellisen tutkimuksen.
Tässä artikkelissa kuvataan kuusi aluetta, joissa LLM:t alkavat olla
hyödyllisiä:  ideointi, taustatutkimus, kirjoittaminen, datan analysointi,
koodaus ja matemaattisten johtopäätösten tekeminen.  Tarjoan yleisiä ohjeita
ja osoitan erityisiä tapauksia, joilla hyödyntää näitä ja luokittelen LLM:n
kyvyt asteikolla "kokeellinen" - "erittäin hyödyllinen".  Hypoteesinani
on, että jatkuva kehitys parantaa LLM:n suorituskykyä kaikilla näillä
alueilla, ja että taloustieteelliset tutkijat, jotka hyödyntävät LLM:ää
mikrotöiden automatisoimiseen, tulevat merkittävästi tuottavammiksi.
Lopuksi spekuloin LLM:n kognitiivisen automatisoinnin pitkän aikavälin
vaikutuksista taloustieteelliseen tutkimukseen.
```

Performance on lower-resource languages for which less digitized text and fewer digitized translations are available is worse.

3.3.5 Explaining Concepts

LLMs can act as a tutor and explain many common economic concepts very clearly – at a level that is quite useful for students who are trying to learn new, and even for more advanced researchers who are branching out beyond their main area of expertise.

Why are instrumental variables useful?

```
Instrumental variables are useful because they can help to identify causal
relationships between variables.  They are used when there is a correlation
between the independent and dependent variables, but it is not clear which
variable is causing the other.  Instrumental variables can help to identify
the causal relationship by introducing a third variable that is correlated
with the independent variable, but not with the dependent variable.  This
allows researchers to isolate the effect of the independent variable on the
dependent variable.
```

Other concept questions that the LLM answered superbly included the following, for which I am leaving out the responses for space reasons:

1. What is the invisible hand theorem and why does it hold?

2. What is Monte Carlo simulation?

3. How does Cournot equilibrium differ from Bertrand equilibrium?

4. What are the leading theories of exchange rate determination?

5. [several more]

But sometimes the system can go terribly wrong:

What is the second theorem of welfare economics?

```
The Second Theorem of Welfare Economics states that any competitive
equilibrium is Pareto efficient.  This means that any competitive market
outcome is the most efficient outcome possible, as it cannot be improved
upon without making at least one individual worse off.
```

I thought that was the first theorem?!

```
No, the First Theorem of Welfare Economics states that any Pareto efficient
outcome can be achieved through some competitive equilibrium.
```

Ok, can you explain again what is the first theorem of welfare economics and what is the second theorem?

```
The First Theorem of Welfare Economics states that any Pareto efficient
outcome can be achieved through some competitive equilibrium.  This means
that any outcome that is beneficial to all parties involved can be achieved
through a competitive market.  The Second Theorem of Welfare Economics
states that any competitive equilibrium is Pareto efficient.  This means
that any competitive market outcome is the most efficient outcome possible,
as it cannot be improved upon without making at least one individual worse
off.
```

The two fundamental theorems of welfare economics are cornerstones of modern economics, but for some reason the LLM confuses the two. This illustrates that – although they give clear and impressive responses in many instances – the responses of LLMs cannot always be trusted, even on what seems like rather basic questions about fundamental concepts.

3.4 Coding

LLMs can serve both as assistants for micro tasks and as tutors for coding. They can write, edit, modify, translate, or debug snippets of code based on instructions in plain English (or other common languages). Moreover, they can act as tutors when using new libraries, functions, or even programming languages that the user is not very familiar with by quickly generating output that shows what libraries and functions are needed for specific types of operations or what syntactic structures to use in a given programming language. This allows the human programmer to consult the LLM and weave together code from many different snippets generated by it.

GPT3.5 has been trained on large amounts of computer code and is accordingly quite powerful at coding. OpenAI's Codex is accessible through the `code-davinci-002`-model or, alternatively, is integrated as Copilot into GitHub. The language model `text-davinci-003` is a descendant of `code-davinci-002` and is therefore not only capable of generating natural language but also computer code. Because of its training data, the two programming languages in which it is most well-versed are python and R, but it is quite capable in any common programming language – from basic functions in Excel to complex C/C++ code.[6]

3.4.1 Writing code

LLMs can be very useful for writing code, in particular for standard programming tasks, data manipulation and repetitive tasks. The following is an example:

Python code to calculate the Fibonacci numbers

```
def Fibonacci(n):
  if n<0:
    print("Incorrect input")
  # First Fibonacci number is 0
  elif n==1:
    return 0
  # Second Fibonacci number is 1
  elif n==2:
    return 1
  else:
    return Fibonacci(n-1)+Fibonacci(n-2)

# Driver Program

print(Fibonacci(9))
```

Another example for which the system is very useful is for plotting graphs. I modified the prompt above to "# Python code to calculate the Fibonacci numbers and plot the first 10 numbers and compare against an exponential curve". The result worked well. The system can also be told to repeat graphs with different variables or changes in formatting. Turning to data manipulation tasks, an example is the prompt "# Python code to reformat a name in standard format to [Last name], [First name]".

The current capabilities of publicly available LLMs are insufficient to write the full code to simulate most economic problems without human help – for example, I tried various

[6] One of the reasons why recent progress by the leading AI labs has been so rapid is that they have developed very powerful LLMs that automate many programming tasks and greatly increase the productivity of their own programmers and developers.

prompts to solve basic economic problems like optimal consumption smoothing, optimal monopoly pricing, etc., and the results were rather inconsistent. However, LLMs can still be useful assistants for automating micro tasks involved in such problems. Moreover, they can act as tutors that help users to find the packages and functions that they need for specific use cases.

3.4.2 Explaining code

Whereas the previous example wrote code based on English language instructions, LLMs can also do the reverse and explain what a given piece of code does in plain English, akin to a tutor:

[continued from previous example]
Explain the code above:

```
The code above is a function that calculates the Fibonacci numbers.  The
function takes an integer n as an argument and returns the nth Fibonacci
number.  The function uses recursion to calculate the Fibonacci numbers.
The base cases are when n is 1 or 2, in which case the function returns 0 or
1 respectively.  For any other value of n, the function returns the sum of
the two previous Fibonacci numbers.
```

This functionality is particularly useful when working with programming languages one is not very familiar with.

3.4.3 Translating code

Frequently, we have code in one programming language that needs to be converted into another language, for example, because we are porting a project to another platform, or because we have found code snippets in an online coding forum such as StackExchange that is useful but in the wrong language. The following example speaks for itself:

[continued from the example in 3.4.1]
Translate the function above from Python to Matlab

```
function F = Fibonacci(n)
 if n < 0
  disp('Incorrect input')
 elseif n == 1
  F = 0;
 elseif n == 2
  F = 1;
 else
  F = Fibonacci(n-1) + Fibonacci(n-2);
 end
end
```

Current LLMs are quite reliable in translating short pieces of code in this fashion across most common languages. For longer sequences, human assistance is still required.

3.4.4 Debugging code

One of the most frustrating parts of programming is to debug code. LLMs display some useful capabilities in this domain, as shown in the following example:

What is the mistake in the following code? [emphasis in bold added after text generation]
def Fibonacci(n):
  elif n==1:
    return 0
  elif **n=2**:
    return 1
  else:
    return Fibonacci(n-1)+**Fibonaci**(n-2)

```
1.  The keyword "elif" should be "if" in the first two lines.
2.  The operator "=" should be "==" in the second line.
3.  The function "Fibonaci" should be "Fibonacci" in the last line.
```

As illustrated, here, the system is highly useful for catching typos and violations of basic syntax. It has some capabilities that go beyond this, e.g., when indices are mixed up and similar situations. For high-level mistakes such as mistakes in the algorithms underlying the code, human debugging is still required. For example, when I changed one of the numbers returned in the code above to a nonsensical number (say “2” instead of “0”), the system did not find anything wrong about that.

3.5 Data Analysis

LLMs can format data, extract data from plain text, classify and score text, extract sentiment, and even simulate human test subjects. Moreover, what is perhaps most useful is that these capabilities can be accessed not only through a web interface, like what is shown in the demonstrations below, but also via an API (Application Programming Interface) that allows large amounts of data to be formatted, extracted, classified etc. These operations can be performed in batches to remain within the token limit for each request. (It goes without saying that LLMs can write computer code in order to access their own APIs – for example, try out “Python code to ask GPT3 to extract birth dates from text input”.)

3.5.1 Extracting data from text

The following example illustrates how LLMs can extract data from written text.

Mark got an A in econ and a B+ in math. Sally got an A- in both econ and math. Frank got a B in econ and a C in math.
Reformat as follows:
Name & econ grade & math grade

```
Mark & A & B+
Sally & A- & A-
Frank & B & C
```

The example provided is kept short for illustrative purposes. The data are names and grades, but other applications include numerical data, e.g., stock prices from news articles or dosage information from drug databases. When I prompted the system with "Can you provide examples of what kinds of numbers you can extract from text?" it answered the following 10 types: phone numbers, zip codes, social security numbers, credit card numbers, bank account numbers, dates, times, prices, percentages, measurements (length, weight etc.)

3.5.2 Reformatting data

LLMs are equally versatile translating data into the desired format. Building on the previous example, I asked the system to format the data first into comma-separated values (CSV) format and then to format it as a LaTeX table:

[continuing the previous example]
Reformat in CSV format:

```
Name,econ grade,math grade
Mark,A,B+
Sally,A-,A-
Frank,B,C
```

Write as a LaTeX table:

```
\begin{tabular}{lcc}
\hline
Name & econ grade & math grade \\
\hline
Mark & A & B+ \\
Sally & A- & A- \\
Frank & B & C \\
\hline
\end{tabular}
```

It excelled at both these and other similar tasks.

3.5.3 Classifying and scoring text

Social science research frequently employs statistical techniques to represent text as data (Gentzkow et al., 2019). Modern LLMs can go beyond traditional techniques for this because they increasingly understand the meaning of the sentences that they are fed. The following example asks GPT3.5 to classify whether a given task listed in the US Department of Labor's Occupational Information Network (O*NET) database is easy or hard to automate and to justify its classification:

Here are 5 tasks from O*Net:
1. Monitor operational records and reports to ensure accuracy and efficiency.
2. Develop and implement policies and procedures for recordkeeping.
3. Prepare and maintain employee records.
4. Maintain and update filing, inventory, mailing, and database systems.
5. Compile and prepare reports, charts, and graphs.
Please classify each of these tasks into easy or hard to automate. Provide your output in the form:
number; task; classification; five-word justification

```
1; Monitor operational records and reports to ensure accuracy and
efficiency; Hard to Automate; Requires Human Judgment/Decision-Making
2; Develop and implement policies and procedures for recordkeeping; Hard to
Automate; Requires Human Judgment/Decision-Making
3; Prepare and maintain employee records; Hard to Automate; Requires Human
Judgment/Decision-Making
4; Maintain and update filing, inventory, mailing, and database systems;
Easy to Automate; Repetitive and Structured Tasks
5; Compile and prepare reports, charts, and graphs; Easy to Automate;
Repetitive and Structured Tasks
```

The results are reasonable, although I should emphasize that they are not entirely robust. Just like a human evaluator who may make different calls depending on circumstances, the system's answers to what is essentially the same question change when the wording of the prompt is changed – and in some cases even when the order of the listed tasks is modified.

3.5.4 Extracting sentiment

LLMs can also extract sentiment from text. They can, for example, classify tweets into "positive" or "negative." Similarly, they can categorize statements by the Federal Open Market Committee (FOMC) that sets US interest rates, as illustrated in the following example:

The following is the December 2022 FOMC statement: [accessed from `https://www.federalreserve.gov/newsevents/pressreleases/monetary20221214a.htm`]
Recent indicators point to modest growth in spending and production. [remainder omitted for space reasons]
Is the statement hawkish or dovish?

```
The statement is hawkish. The Committee is raising the target range for
the federal funds rate and reducing its holdings of Treasury securities and
agency debt and agency mortgage-backed securities. The Committee is also
strongly committed to returning inflation to its 2 percent objective.
```

The system's assessment is correct and well-argued. I also explored whether the system could identify whether the December 2022 or February 2023 FOMC statement was more hawkish, but its ability to assess Fed-speak was not quite nuanced enough – it focused mainly on the interest rates in February 2023 being higher as opposed to the small and nuanced changes in the text of the statement that could indicate a change in direction.

3.5.5 Simulating human subjects

Argyle et al. (2022) propose the use of LLMs to simulate human subjects, based on the observation that the training data of LLMs contains a large amount of information about humanity. They condition GPT3 on the socio-demographic backstories of real humans and demonstrate that subsequent answers to survey questions are highly correlated with the actual responses of humans with the described backgrounds, in a nuanced and multifaceted manner. Horton (2022) showcases applications to economics, using simulated test subjects to replicate and extend upon several behavioral experiments. The following example illustrates the concept:

For each of the following registered voters, would they approve or disapprove of a policy proposal loosen US immigration restrictions on well-educated workers?
1. A 32-year old female lawyer in Houston, TX who is Hispanic
2. A 73-year old male retired doctor in Rome, GA who is Caucasian
3. A 25-year old software engineer in San Francisco, CA who is Asian and identifies as non-binary
4. A 45-year old female grocery check-out clerk in Brooklyn, NY who is second-generation Italian
5. A 55-year old insurance agent in Lake City, FL who is Caucasian

```
1. Approve
2. Disapprove
3. Approve
4. Approve
5. Disapprove
```

There is a significant risk that the simulated results simply propagate false stereotypes, and they must hence be used with great care. However, they also contain valuable information. If used correctly, they can provide useful insights about our society, from which all the data used to train the LLMs ultimately originate.

3.6 Mathematical Derivations

The ability of LLMs to perform mathematical derivations is emerging but still limited at this point. Noorbakhsh et al. (2021) show that LLMs can be fine-tuned for mathematical tasks. Frieder et al. (2023) develop a dataset of graduate-level mathematical questions and conclude that ChatGPT's mathematical abilities are significantly below those of an average mathematics graduate student – perhaps providing some solace for the short term. However, current LLMs are trained mostly on text and only the occasional math paper. I would not be surprised if the new dataset created by Frieder et al. (2023) will be useful for making future LLMs significantly better at math. This is an area in which further progress would be very valuable.

3.6.1 Setting up models

The following prompt illustrates how LLMs can be useful for setting up models. It describes a consumer with consumption utility that is iso-elastic in one good and linear in the other good, written in LaTeX, and assigns variables for the prices, akin to what one may write on an undergraduate micro problem set. Based on this prompt, the LLM knows to continue with the consumer's optimization problem. The left column displays the generated text, and the right column shows the typset version compiled by LaTeX.

`Utility is $u(x)=\frac{x^{1-\rho}}{1-\rho}+y$`
`and the prices of the goods are $p$ and $1$. A`
`consumer` `has income $m$.`
`The consumer's optimal bundle is given by`
`$\max_{x,y} \frac{x^{1-\rho}}{1-\rho}+y$`
`subject to $px+y \leq m$.`
`The Lagrangian is`
`$L(x,y,\lambda)=\frac{x^{1-\rho}}{1-\rho}`
`+y+\lambda(px+y-m)$.`
`The first order conditions are`
`$\frac{\partial L}{\partial x}=`
`    \frac{x^{-\rho}}{1-\rho}-\lambda p=0$`
`$\frac{\partial L}{\partial y}=`
`    1-\lambda=0$`
`$\frac{\partial L}{\partial \lambda}=`
`    px+y-m=0$`
`Solving the first two equations for $x$`
`and $y$ yields`
`$x=\left(\frac{1-\rho}{\lambda p}\right)`
`    ^{\frac{1}{\rho}}$`
`$y=\lambda-1$`
`Substituting these into the third equation`
`yields`
`$\lambda=\frac{m}{p\left(\frac{1-\rho}`
`  {\lambda p}\right)^{\frac{1}{\rho}}+1}$`
`Solving for $\lambda$ yields`
`$\lambda=\left(\frac{m(1-\rho)}{p}\right)`
`    ^{\frac{1}{1+\rho}}$`
`Substituting this into the expressions for`
`$x$ and $y$ yields`
`$x=\left(\frac{p(1-\rho)}{m}\right)^`
`    {\frac{\rho}{1+\rho}}$`
`$y=\left(\frac{m(1-\rho)}{p}\right)^`
`    {\frac{1}{1+\rho}}-1$`
	Utility is $u(x)=\frac{x^{1-\rho}}{1-\rho}+y$ and the prices of the goods are p and 1. A consumer has income m.
	The consumer's optimal bundle is given by
	$\max_{x,y} \frac{x^{1-\rho}}{1-\rho}+y$ subject to $px+y \leq m$.
	The Lagrangian is
	$L(x,y,\lambda)=\frac{x^{1-\rho}}{1-\rho}+y+\lambda(px+y-m)$.
	The first order conditions are
	$\frac{\partial L}{\partial x}=\frac{x^{-\rho}}{1-\rho}-\lambda p=0$
	$\frac{\partial L}{\partial y}=1-\lambda=0$
	$\frac{\partial L}{\partial \lambda}=px+y-m=0$
	Solving the first two equations for x and y yields
	$x=\left(\frac{1-\rho}{\lambda p}\right)^{\frac{1}{\rho}}$
	$y=\lambda-1$
	Substituting these into the third equation yields
	$\lambda=\frac{m}{p\left(\frac{1-\rho}{\lambda p}\right)^{\frac{1}{\rho}}+1}$
	Solving for λ yields
	$\lambda=\left(\frac{m(1-\rho)}{p}\right)^{\frac{1}{1+\rho}}$
	Substituting this into the expressions for x and y yields
	$x=\left(\frac{p(1-\rho)}{m}\right)^{\frac{\rho}{1+\rho}}$
	$y=\left(\frac{m(1-\rho)}{p}\right)^{\frac{1}{1+\rho}}-1$

It can be seen that the LLM correctly fills in a suitable budget constraint and sets up the associated maximization problem. In the Lagrangian, the system includes the budget constraint with an unusual sign. It derives two of the three first-order conditions correctly – as if the sign on the budget constraint in the Lagrangian had just been a typo – but stumbles on the derivative of the iso-elastic utility function. The remainder of the derivations is largely gibberish.

Although it takes up time to read the generated text and spot the mistake, the LLM automatically writing out the maximization problem and Lagrangian and solving parts of the maximization problem within seconds still helps to save valuable researcher time.

3.6.2 Deriving equations

As the previous example demonstrates, current LLMs have already some limited capability in deriving equations. In fact, continuing on the example above, I corrected the mistake in the first-order condition and asked the system to generate the remainder.

Utility is `$u(x)=\frac{x^{1-\rho}}{1-\rho}+y$` and the prices of the goods are `$p$` and `$1$`. A consumer has income `$m$`.
The consumer's optimal bundle is given by `$\max_{x,y} \frac{x^{1-\rho}}{1-\rho}+y$` subject to `$px+y \leq m$`.
The Lagrangian is `$L(x,y,\lambda)=\frac{x^{1-\rho}}{1-\rho} +y+\lambda(px+y-m)$`.
The first order conditions are `$\frac{\partial L}{\partial x}=x^{-\rho}-\lambda p=0$` `$\frac{\partial L}{\partial y}=1-\lambda=0$` `$\frac{\partial L}{\partial \lambda}=px+y-m=0$`
To solve for `$x$` and `$y$` `we first solve for $\lambda$ from the second equation:`
`$\lambda=1$`
`Substituting this into the third equation we get`
`$px+y=m$`
`Substituting this into the first equation we get`
`$x^{-\rho}=p$`
`Solving for $x$ we get`
`$x=\left(\frac{1}{p}\right) ^{\frac{1}{\rho}}$`
`Substituting this into the equation for $px+y=m$ we get`
`$y=m-p\left(\frac{1}{p}\right) ^{\frac{1}{\rho}}$`
`Therefore, the consumer's optimal bundle is`
`$x=\left(\frac{1}{p}\right) ^{\frac{1}{\rho}}$ and $y=m-p\left(\frac{1}{p}\right) ^{\frac{1}{\rho}}$.`
	Utility is $u(x)=\frac{x^{1-\rho}}{1-\rho}+y$ and the prices of the goods are p and 1. A consumer has income m.
	The consumer's optimal bundle is given by $\max_{x,y} \frac{x^{1-\rho}}{1-\rho}+y$ **subject to** $px+y \leq m$.
	The Lagrangian is $L(x,y,\lambda)=\frac{x^{1-\rho}}{1-\rho}+y+\lambda(px+y-m)$.
	The first order conditions are
	$\frac{\partial L}{\partial x}=x^{-\rho}-\lambda p=0$
	$\frac{\partial L}{\partial y}=1-\lambda=0$
	$\frac{\partial L}{\partial \lambda}=px+y-m=0$
	To solve for x and y we first solve for λ from the second equation:
	$\lambda=1$
	Substituting this into the third equation we get
	$px+y=m$
	Substituting this into the first equation we get
	$x^{-\rho}=p$
	Solving for x we get
	$x=\left(\frac{1}{p}\right)^{\frac{1}{\rho}}$
	Substituting this into the equation for $px+y=m$ we get
	$y=m-p\left(\frac{1}{p}\right)^{\frac{1}{\rho}}$
	Therefore, the consumer's optimal bundle is
	$x=\left(\frac{1}{p}\right)^{\frac{1}{\rho}}$ and $y=m-p\left(\frac{1}{p}\right)^{\frac{1}{\rho}}$.

The system proceeded to correctly deriving the solution of the maximization problem, i.e., the optimal levels of x and y as a function of the exogenous parameters.

However, the mathematical abilities of the system are still quite limited: after obtaining the correct solution, I corrected the sign mistake in the Lagrangian and attempted to regenerate the remainder of the derivations, but the system produced gibberish. I attempted several other derivations and found the ratio of mistakes too high for the system to be useful in this application.

3.6.3 Explaining models

Current LLMs also have some limited ability to explain simple models. The following is an example of where the system explains the math underlying the famous bat-and-the-ball problem:

Solve the bat-and-the-ball problem and state all your intermediate steps:

```
The bat-and-the-ball problem is a classic problem in mathematics.  It states
that a bat and a ball cost $1.10 in total.  The bat costs $1.00 more than
the ball.  How much does the ball cost?
Intermediate Steps:
1.  Let x represent the cost of the ball.
2.  Since the bat costs $1.00 more than the ball, the cost of the bat is x +
1.
3.  Since the total cost of the bat and the ball is $1.10, the equation is x
+ (x + 1) = 1.10.
4.  Simplifying the equation, 2x + 1 = 1.10.
5.  Subtracting 1 from both sides of the equation, 2x = 0.10.
6.  Dividing both sides of the equation by 2, x = 0.05.
7.  Therefore, the ball costs $0.05.
```

Interestingly, when I tried variations of the prompt, for example by just stating "Solve the bat-and-the-ball problem:" the system frequently committed the classic human fallacy in the bat-and-the-ball problem, i.e., it suggested incorrectly that the ball cost \$.10. This is related to an interesting phenomenon: LLMs frequently generate more reliable content when asked to show their intermediate steps, called "chain-of-thought prompting" – akin to students who are less error-prone when they are asked to explain the intermediate steps behind the solutions that they report (Wei et al., 2022b).

3.7 Summary

Table 1 summarizes all the sample tasks illustrated in this paper, categorized by the six described domains of application of LLMs. In the third column of the table, I report my subjective rating of how useful I found the described LLM capabilities as of Feb 1, 2023. My rating ranges from 1 to 3, where 1 describes capabilities that I currently consider more experimental and that deliver inconsistent results, requiring significant human oversight ; 2 signifies capabilities that are useful and likely to save time but

Category	Task	Usefulness
Ideation	Brainstorming	3
	Evaluating ideas	2
	Providing counterarguments	3
Writing	Synthesizing text	3
	Editing text	3
	Evaluating text	3
	Generating catchy titles & headlines	3
	Generating tweets to promote a paper	3
Background Research	Summarizing Text	3
	Literature Research	1
	Formatting References	3
	Translating Text	3
	Explaining Concepts	2
Coding	Writing code	2
	Explaining code	2
	Translating code	3
	Debugging code	2
Data Analysis	Extracting data from text	3
	Reformatting data	3
	Classifying and scoring text	2
	Extracting sentiment	2
	Simulating human subjects	2
Math	Setting up models	2
	Deriving equations	1
	Explaining models	1

The third column reports my subjective rating of LLM capabilities as of Feb 1, 2023:
1 = experimental; results are inconsistent and require significant human oversight
2 = useful; requires oversight but will likely save you time
3 = highly useful; incorporating these into your workflow will save you time

Table 1: Summary of LLM capabilities and rating of usefulness as of Feb 1, 2023

are somewhat inconsistent so that they still require careful oversight; and 3 reflects capabilities that are already highly useful and work in the expected manner most of the time. Incorporating these latter capabilites into your workflow will definitely save you time and make you more productive.

4 Outlook and Concluding Thoughts

LLMs have become useful research tools for tasks ranging from ideation, writing and background research to data analysis, coding, and mathematical derivations. In the short term, cognitive automation via LLMs will allow researchers to become significantly more productive. I expect that a growing number of researchers will incorporate LLMs into their workflow. This could help to increase the overall speed of progress in economics, although it risks leaving behind those who do not take advantage of LLMs.[7]

In the medium term, I anticipate that LLM-based assistants and tutors will become increasingly useful for generating the content that makes up research papers. Human researchers will focus on their comparative advantage – by posing the questions, suggesting directions for obtaining answers, discriminating which parts of the produced content are useful, editing, and providing feedback, akin to an advisor. Moreover, they will also continue to play an important role in organizing research efforts – for example, by coordinating teams and procuring data sources, akin to a research manager.

Over time, further advances will imply that LLMs are performing their tasks better and better so that the need for humans to provide inputs, edits, and feedback will diminish. We may increasingly just rubber-stamp the output produced by ever-more advanced LLMs. The experience may be deflating. Eventually, our AI research assistants will graduate and become researchers of their own.

It is difficult to predict whether and how different areas of research will be differentially affected by cognitive automation – for example, will theorists be the last ones standing because their abilities prove difficult to replicate by LLMs, or will a more advanced LLM fine-tuned for mathematical applications outperform humans and automate theory work more quickly than other branches of economics? Will empiricists have a leg up because the process of collecting novel data involves many steps that are difficult to automate?

In the longer term, I believe that economists would be well advised to heed the "Bitter Lesson" of progress in AI, which Sutton (2019) described as follows: for most of the history of AI, researchers worked on making their AI systems smarter and more poweful by programming domain-specific knowledge into them – for example, teaching a chess computer the wisdom accumulated by generations of chess players. He observed that this strategy always helped in the short term, but the benefits of it eventually plateaued.

[7] To the extent that longer and more complex papers are the result of a positional arms race among researchers, greater productivity in generating text may also lead to further bloating of research papers without improving depth or quality (see, e.g. Frank, 1991).

In the long term, Sutton suggests that the brute scaling of compute has always proven the more successful strategy – for example, when DeepMind developed AlphaZero, a chess computer that used massive compute to learn chess by itself without any human input, it learned to beat all other chess computers in the world (and of course all humans) within 24 hours (Silver et al., 2017). This strategy corresponds to what DeepMind founder Demis Hasabis has called "solving intelligence, and then using that to solve everything else."

In our work as economists, we spend a lot of our time and energy on similar strategies to what Sutton described, expending tremendous resources on fine-tuning our models of human behavior and of the economy to obtain better results. Yet a similar bitter lesson may apply to economics: with enough compute, sufficiently advanced AI systems may be able to produce and articulate superior economic models, and the cognitive work of human economists – like that of all other researchers – may eventually become redundant.

Garry Kasparov (2017) distills the lessons he learned from observing decades of progress in chess computers, with important milestones including his 1997 defeat to Deep Blue and the 2017 release of AlphaZero, as follows (p. 254-255):

> "Thousands of years of status quo human dominance, a few decades of weak competition, a few years of struggle for supremacy. Then, game over. For the rest of human history, [...] machines will be better than humans at chess. The competition period is a tiny dot on the historical timeline. This is the unavoidable one-way street of technological progress in everything from the cotton gin to manufacturing robots to intelligent agents.
>
> The competition dot gets all the attention because we feel it intensely when it occurs during our lifetimes. The struggle phase often has a direct impact on our lives in real time, so we overinflate its relevance in the big picture. [...] it is almost always better to start looking for alternatives and how to advance the change into something better instead of trying to fight it and hold on to the dying status quo."

In Kasparov's terminology, LLMs have entered the period of "weak competition" with cognitive workers, including economic researchers. We are currently at the competition dot, and LLMs are garnering a lot of attention. Yet just like the chess champions of the 1990s, we should not let our anthropocentric bias blind us to the rise of AI, and we should remind ourselves that the competition period, which we may feel intensely in coming years, is just a tiny dot on the historical timeline.

Whereas my long-term predictions are clearly speculative, I am quite confident about my predictions on the short- and medium-term implications of LLMs. I also believe that the cognitive automation ushered in by the rapid rise of LLMs poses important and urgent new research questions to economists, of which I will brainstorm a few:

1. What will cognitive automation imply for labor markets? Will it also accelerate the automation of physical tasks? How can our society best prepare for the impending changes?

2. What are the implications of cognitive automation for education? Will human capital be devalued?

3. How will cognitive automation affect technological progress and economic growth? If human labor can be automated, what will be the bottlenecks to growth in the future?

4. ...

5. Finally, but perhaps most importantly, how can we best address the AI alignment problem, i.e., ensure that ever-more advanced and potentially super-intelligent AI systems pursue objectives that are aligned with human objectives?

Continuing on the last question, economists have the tools to translate concepts from the social sciences and humanities, such as "human objectives," into analytic concepts like preferences that are more easily accessible to machines. And we have experience analyzing agency and control problems and their solutions. Their contribution is urgently needed. In fact, there are two channels through which economists can make important contributions to this line of work: First, we can directly work on AI alignment; see, e.g., Korinek and Balwit (2023), for some tentative research directions. Second, our work will affect the concepts and representations through which future AI systems will view economic questions and, ultimately, through which they will view the world – just like our work influences that of our human students, whether they work as economists or policymakers, it will also influence future LLMs that perform economic research and that impact economic policy. As Keynes (1936) described so powerfully at the conclusion of his general theory,

> "...the ideas of economists and political philosophers, both when they are right and when they are wrong, are more powerful than is commonly understood. Indeed the world is ruled by little else. [...] I am sure that the power of vested interests is vastly exaggerated compared with the gradual encroachment of ideas. [...] soon or late, it is ideas, not vested interests, which are dangerous for good or evil."

At this point, human researchers, especially when AI-assisted, are still the best technology around for generating economic research!

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C., and Wingate, D. (2022). Out of one, many: Using language models to simulate human samples. *arXiv:2209.06899*.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., and et al. (2021). On the opportunities and risks of foundation models. *arXiv:2108.07258*.

Carlsmith, J. (2020). *How Much Computational Power Does It Take to Match the Human Brain?* Open Philanthropy.

Frank, R. H. (1991). Positional externalities. In Zeckhauser, R., editor, *Strategy and Choice*, pages 25–47. MIT Press, Cambridge, MA.

Frieder, S., Pinchetti, L., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., and Berner, J. (2023). Mathematical capabilities of ChatGPT.

Gentzkow, M., Kelly, B. T., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–74.

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., and Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. Technical report, Georgetown, Center for Security and Emerging Technology.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. *arXiv:2203.15556*.

Horton, J. J. (2022). Large language models as simulated economic agents: What can we learn from homo silicus? working paper.

Jiao, W., Wang, W., Huang, J.-t., Wang, X., and Tu, Z. (2023). Is ChatGPT a good translator? A preliminary study. *arXiv:2301.08745*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv:2001.08361*.

Kasparov, G. (2017). *Deep thinking: where machine intelligence ends and human creativity begins*. PublicAffairs.

Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money*. Palgrave Macmillan.

Korinek, A. and Balwit, A. (2023). Aligned with whom? Direct and social goals for AI systems. In Bullock, J. and et al., editors, *Oxford Handbook of AI Governance*. Oxford University Press.

Korinek, A. and Juelfs, M. (2022). Preparing for the (non-existent?) future of work. *NBER Working Paper*, 30172.

Noorbakhsh, K., Sulaiman, M., Sharifi, M., Roy, K., and Jamshidi, P. (2021). Pretrained language models are symbolic mathematics solvers too! arXiv:2110.03501.

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P. (2022). Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv:1712.01815.*

Sutton, R. (2019). The bitter lesson. Incomplete Ideas (blog). http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

Thompson, A. D. (2023). GPT-3.5 + ChatGPT: An illustrated overview. Technical report, LifeArchitect.ai.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, LIX(236):433–460.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022a). Emergent abilities of large language models. *arXiv:2206.07682.*

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2022b). Chain-of-thought prompting elicits reasoning in large language models. *arXiv:2201.11903.*