Harvard
Business
Review

AI And Machine Learning

Is GenAI's Impact on Productivity Overblown?

by Ben Waber and Nathanael J. Fast

January 08, 2024

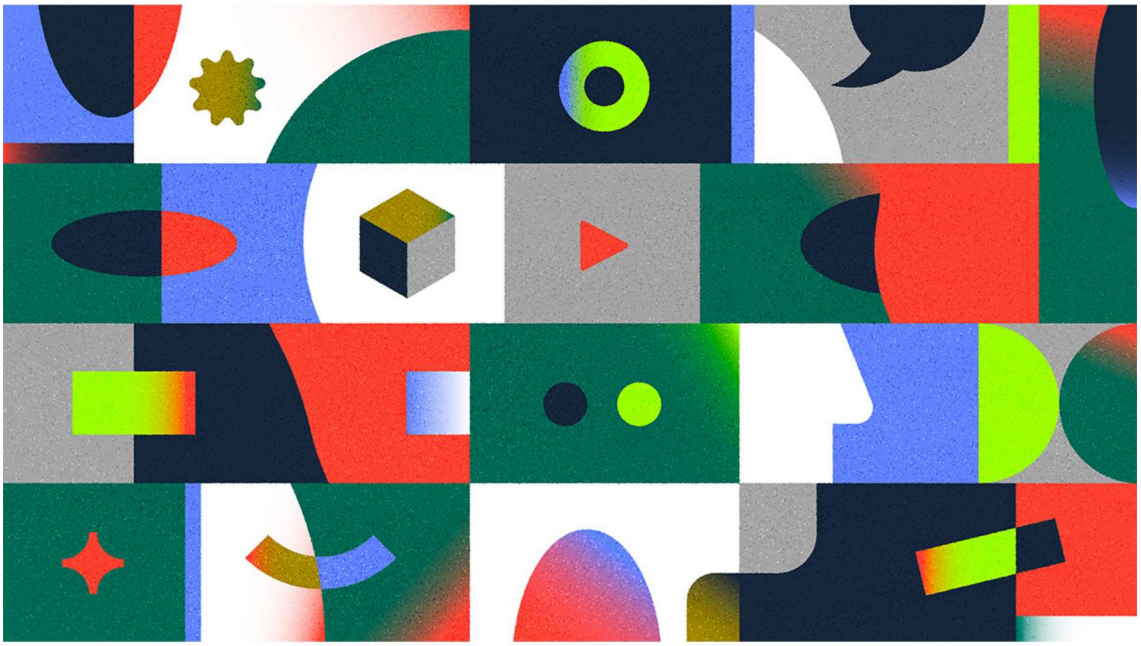

Illustration by Kemal Sanli

Summary. Generative AI like LLMs have been touted as a boon to collective productivity. But the authors argue that leaning into the hype too much could be a mistake. Assessments of productivity typically focus on the task level and how individuals might use and benefit... **more**

Ler em português

Large language models (LLMs) have been heralded as a boon to collective productivity. McKinsey boldly proclaimed that LLMs and other forms of generative AI could grow corporate profits globally by $4.4 trillion annually, and Nielsen trumpeted a 66%

increase in employee productivity through the use of these same tools. Projections like these have made finding ways to use these tools — and turbocharge productivity — a top priority for many companies over the past year. While we are intrigued and impressed by this new technology, we advise cautious experimentation instead of wholesale company-wide adoption.

Amid all the hype, there is reason to question whether these tools will have the transformative effects on company-wide productivity that some predict. One reason to take a slower approach is that assessments of productivity typically focus on the task level — summarizing a document, completing a slide deck, or answering a customer call, for example — and how *individuals* might use and benefit from LLMs. Using such findings to draw broad conclusions about *firm-level performance* could prove costly.

Consider recent research on the impact of generative AI in a call center environment, where a machine learning platform with an LLM interface was trained on chat and outcome data. The researchers looked at the average chat completion time to measure productivity and, on average, they saw a 14% improvement in chat completion time with the new tool.

A closer look, however, reveals a few worrying signs. Per the call center study we linked to, top employees' performance actually decreased with this system, which presents potential problems for innovation, motivation, and retaining a firm's best performers. In another study, researchers found more productivity gains from using generative AI for tasks that were well-covered by current models, but productivity decreased when this technology was used on tasks where the LLMs had poor data coverage or required reasoning that was unlikely to be represented in online text. Over time, external conditions (e.g., cultural values, known best practices) can materially change, causing benefits to either disappear or even lead to significant productivity *decreases*.

The consequences of introducing new products, including the possibility of turnover among experts whose output is used to train these systems, hasn't been examined. We argue that, in the absence of a more comprehensive, long-term analysis, looking at task-specific data reveals little about the true effect of a new technology like LLMs on overall firm performance.

As such, we suggest that organizations need to take a nuanced, data-driven approach to adopting LLMs. Leaders should consider where this technology actually helps and resist the urge to integrate it into every job and task throughout the organization. To do this, they need to understand two core problems of LLMs that are critical to their medium- and long-term business implications: 1) Its persistent ability to produce convincing falsities and 2) the likely long-term negative effects of using LLMs on employees and internal processes.

On the first, it is important to appreciate that LLMs' leaps in syntactic fluency don't translate into being better able to automatically look up facts — a problem that has confronted computer science for decades with incremental progress. On the second, the productivity effects of LLMs are often confined to performance on a self-contained task where a model has already been trained — a fact that can distort incentives for top performers and introduce systemic risks into complex workflows. When combined, these issues create organizational conditions that are ripe for systemic, hard-to-identify failures that can easily degrade organizational effectiveness if use cases for generative AI are not narrowly scoped and continuously monitored.

Plausible Fabrication

LLMs and machine learning in general by their nature predict future patterns based on what worked (or, more accurately, what persistently *occurred*) in the past. A full explanation of how LLMs work is beyond the scope of this piece (one good explainer is here), but at a basic level, these extremely large models take in huge amounts of text (at this point, nearly all text on the web) and

build a statistical model of next word prediction. After initial training, most companies pay annotators to give feedback on prompts to reduce the likelihood of toxic output.

Importantly, notice that this model doesn't have any concept of truth or fact (it was trained on the internet, after all). LLMs provide answers that are statistically likely to occur in public text. To the extent that truth is more likely to have occurred in the training data, LLMs are more likely to provide factual output. A quick perusal of recent news articles provides ample examples of when these models confidently provide blatant falsities. My (Ben) favorite example right now is to ask any of these models which African countries start with the letter "k." And my (Nate) favorite is the tendency for Google Bard's email tool to fabricate entire emails that were never sent; fully consistent with how LLMs work but not so helpful as a tool.

Unfortunately, this is not an easily fixable problem. Machine learning researchers have been working for decades to map questions into factual databases, and while LLMs provide a much more coherent front end, the core problem of retrieving facts based on natural language input remains unsolved. The fundamental innovation of LLMs — creating a model big enough with enough data to learn the statistical properties of syntax — is unrelated to factual retrieval.

People trick themselves into thinking that they can prompt LLMs into giving them only factual output, but that is simply not how the technology works. Entering something like "only give me output where you can find a source" doesn't change the model fundamentally, it just means that it will complete text that looks like what comes after when someone asks for a source. Sometimes that will work, to the extent that a particular segment of text frequently appears in the data, and sometimes it won't. A humorous description for LLMs we came across is "mansplaining as a service." Sometimes it's right, sometimes it's wrong, but it always sounds authoritative.

Stuck in the Past

The fact that this language also occurs in the past is an important consideration when considered in an organizational context. Take the call center example from before. If a company releases a new product, there are no chat logs dealing with that product to train on. Even assuming the output was correct in the past, it could be completely wrong moving forward.

One might say "fine, we'll need to retrain again," but retraining also raises a number of issues. First, it assumes that people know enough about performance changes to understand there is a problem. A new product release may be easy enough, but what about a change in marketing strategy? What about a change in an API that a programmer uses in a code completion LLM? Companies will need to implement extensive new processes to monitor these potential conflicts effectively, likely at great cost. Moreover, while changes in task completion speed are easy to measure, changes in accuracy are less detectable. If an employee completes a report in five minutes instead of 10, but it's less accurate than before, how would we know, and how long will it take to recognize this inaccuracy?

Second, the incentives for top performers to contribute to the retraining of these tools has shifted. Remember, reproducing the behavior of top performers doesn't help their performance; in the above study, it hindered it. If they're getting paid less and everyone else is getting paid more, they will become far less likely to engage in the exploratory behavior they previously exhibited to find innovative solutions. They may also be more likely to leave the company, degrading the performance of the overall system.

Model Collapse

As these systems start to be trained on their own output, organizations that rely on them will face the problematic issue of model collapse. While originally trained on human-generated text, LLMs that are trained on the output of LLMs degrade rapidly in quality. Given that these systems will need to be continuously retrained by humans in a real environment, and that the text they

are trained on will be generated at least partially from previous LLM output, this indicates that systems will deliver low or even negative value in a few training cycles.

This is only one of the roadblocks for people who claim that these models will continue to improve at a breakneck pace. There's simply not another internet's worth of text to train on, and one of the primary innovations of LLMs was the ability to ingest massive amounts of text. Even if there was, that text is now polluted by LLM output that will degrade model quality. There's already some evidence that model performance in the current paradigm has peaked.

A Long-Term Perspective on LLM Effects

To fully appreciate the problem of less-than-factual output, you have to take a long view. The "ChatGPT lawyer" is particularly instructive. In this case, a lawyer used ChatGPT to write a legal brief. When details in the brief turned out to be wrong, it created a scandal and a cascade of work for the court and these lawyers.

Viewed through a task-performance view, the use of ChatGPT in this case was a success. Rather than take days to write a legal brief, these lawyers saw their individual productivity skyrocket by using ChatGPT to write one in minutes. At a system view, this was a colossal failure. Because the outputs of ChatGPT *seem* authoritative — going so far as to use psychological tricks to encourage trust such as replying with "I" — even people who know they should double check the output are much less likely to do so. The use of ChatGPT here dramatically decreased the productivity of the overall court system.

This was a case where it was easy to point to ChatGPT as the culprit. But now imagine companies using LLMs to, say, write an employee handbook. While employees should check the entire handbook closely, after reading a few pages of authoritative-sounding, coherent text, they will likely skim the rest of it. If an error is introduced into the handbook, it might not show up for years. Imagine that an automatically generated employee

handbook omitted important details on penalties for sexual harassment. Later, if sexual harassment happens in the workplace and the company finds itself unable to fire the perpetrator, it will be extremely challenging to pin it on the use of an LLM for the handbook. These kind of risks cannot be properly quantified at the task level or in the short term. One needs a holistic, organizational, longitudinal evaluation.

With Prejudice

It is critical to address the role of LLMs in reinforcing and amplifying biases, which has been validated across many studies. While we would argue that it is enough to make an argument that this is ethically wrong and that organizations should be cautious in using these systems, it can be helpful to also focus on the economic effects.

Research has so often demonstrated the benefits of a diverse and inclusive workforce that asset managers are now using these metrics to drive investment and executive compensation decisions. Technology like LLMs that erases types of language that marginalized communities use or minimizes their contribution through poor summarization could make these communities feel unseen or unwelcome. For native speakers of languages that don't have enough text online to train LLMs, LLMs will have less data to draw from to provide accurate translations, further reinforcing their exclusion.

Since generative AI is disproportionately likely to show results that reinforce the social status quo, companies that make greater use of this technology run the risk of alienating their marginalized employees. Higher attrition from those groups will be costly in its own right, in addition to limiting idea generation.

Risky Business

Taken together, these points indicate large classes of work where using LLMs is risky. For projects and workflows where the truth matters, any claim of productivity improvements from this class of technology carries a high burden of proof that must address

many of the issues raised above (and probably more, such as the environmental cost of training and using these models, cybersecurity risks, etc.) in a longitudinal, holistic fashion. Task-level experiments are not enough.

When work involves summarizing and synthesizing evidence, LLMs could prove to be unreliable. For policy and process development or implementation, dispute resolution, report generation, and more, existing evidence indicates that LLMs may actually reduce overall performance rather than support it. Early research also indicates that when it becomes known that generative AI tools are being used for content generation in interpersonal communication, trust can be significantly reduced. This has profound implications for the ability of teams to have difficult discussions, engage in brainstorming, and pursue other mission-critical processes.

It's important to note that there are other significant ethical issues with this class of technology that we didn't address here. These issues include everything from the expansion and ossification of societal biases to problems of copyright infringement, as these models tend to memorize particularly unique data points. These issues are significant and their impact on the legal permissibility of LLMs does create additional risks, but they are best examined in a more thorough treatment.

Where Do We Go From Here?

In this article we have voiced skepticism about the hype surrounding LLMs, arguing for a more cautious approach. Making grandiose claims about LLMs may help people sell software or books in the short term, but in the long term, unthinking organization-wide application of these models could very well lead to significant losses in productivity. However, these productivity losses will be hard to measure, and the danger is that this hard-to-quantify drag on performance will continue unchallenged due to the deep integration of these tools into inappropriate workflows.

This is not to say that the technology isn't useful for certain classes of work but rather that users and developers must be clear about when we can use LLMs effectively and confidently. When people are writing in a foreign language, for example, using an LLM to clean up existing text to make it sound more natural and easily understood by others has the potential to level the playing field between native and non-native speakers. AI also holds promise for tasks where generating lots of non-factual ideas quickly is useful. It's easy to imagine tech products that focus on these use cases in a way that makes it easy for organizations to experiment with small scale, targeted applications. Leaders should be on the lookout for contexts where adopting LLMs proves helpful, neither blindly adopting nor blindly rejecting the technology.

In closing, not every new work technology leads to firm-level productivity improvements, though the hype surrounding AI could allow companies to rationalize replacing high-paid workers with low-paid workers, thereby increasing short-term profits even at the expense of productivity. With generative AI, we have the potential to avoid this trap, but only if we channel, test, and use it intelligently.

Ben Waber is the president and co-founder of the workplace analytics company Humanyze and the author of *People Analytics: How Social Sensing Technology Will Transform Business and What It Tells Us About the Future of Work*.

Nathanael J. Fast is the Jorge Paulo and Susanna Lemann Chair in Entrepreneurship and associate professor of management at the Marshall School of Business at the University of Southern California. He directs the USC Neely Center for Ethical Leadership and Decision

Making and co-directs the Psychology of Technology Institute. His research examines responsible leadership in the age of AI.

Recommended For You

Leading in a World Where AI Wields Power of Its Own

Demystifying Emissions Reporting

PODCAST
Should Businesses Take a Stand on Societal Issues?

What Will Working with AI Really Require?