What will GPT-2030 look like?

by **Jacob Steinhardt** 7th Jun 2023

Forecasts (Specific Predictions) | Scaling Laws | AI | Curated

GPT-4 surprised many people with its abilities at coding, creative brainstorming, letter-writing, and other skills. Surprises in machine learning are not restricted to GPT-4: I was previously surprised by Minerva's mathematical abilities, as were many competitive forecasters.

How can we be less surprised by developments in machine learning? Our brains often implicitly make a zeroth-order forecast: looking at the current state of the art, and adding on improvements that "feel reasonable". But what "seems reasonable" is prone to cognitive bias, and will underestimate progress in a fast-moving field like ML. A more effective approach is first-order forecasting: quantifying the historical rate of progress and extrapolating it forward, while also considering reasons for possible slowdowns or speedups.[1]

In this post, I'll use this approach to forecast the properties of large pretrained ML systems in 2030. I'll refer throughout to "GPT_{2030}", a hypothetical system that has the capabilities, computational resources, and inference speed that we'd project for large language models in 2030 (but which was likely trained on other modalities as well, such as images). To forecast GPT_{2030}'s properties, I consulted a variety of sources, including empirical scaling laws, projections of future compute and data availability, velocity of improvement on specific benchmarks, empirical inference speed of current systems, and possible future improvements in parallelism.

GPT_{2030}'s capabilities turn out to be surprising (to me at least). In particular, GPT_{2030} will enjoy a number of significant advantages over current systems[2], as well as (in at least some important respects) current human workers:

1. GPT_{2030} will likely be superhuman at various specific tasks, including coding, hacking, and math, and potentially protein engineering (Section 1).
2. GPT_{2030} can "work" and "think" quickly: I estimate it will be 5x as fast as humans as measured by words processed per minute [*range: 0.5x-20x*][3], and that this could be increased to 125x by paying 5x more per FLOP (Section 2).
3. GPT_{2030} can be copied arbitrarily and run in parallel. The organization that trains GPT_{2030} would have enough compute to run many parallel copies: I estimate enough to perform 1.8 million years of work when adjusted to human working speeds [*range: 0.4M-10M years*] (Section 3). Given the 5x speed-up in the previous point, this work could be done in 2.4 months.
4. GPT_{2030}'s copies can share knowledge due to having identical model weights, allowing for rapid parallel learning: I estimate 2,500 human-equivalent years of learning in 1 day (Section 4).
5. GPT_{2030} will be trained on additional modalities beyond text and images, possibly including counterintuitive modalities such as molecular structures, network traffic, low-

level machine code, astronomical images, and brain scans. It may therefore possess a strong intuitive grasp of domains where we have limited experience, including forming concepts that we do not have (Section 5).

These capabilities would, at minimum, accelerate many areas of research while also creating serious vectors for misuse (Section 6). Regarding misuse, GPT_{2030}'s programming abilities, parallelization, and speed would make it a potent cyberoffensive threat. Additionally, its rapid parallel learning could be turned towards human behavior and thus used to manipulate and misinform with the benefit of thousands of "years" of practice.

On acceleration, a main bottleneck will be autonomy. In a domain like mathematics research where work can be checked automatically, I'd predict that GPT_{2030} will outcompete most professional mathematicians. In machine learning, I'd predict that GPT_{2030} will independently execute experiments and generates plots and write-ups, but that graduate students and research scientists will provide direction and evaluate results. In both cases, GPT_{2030} will be an integral part of the research process.

My forecast of GPT_{2030}'s properties are not intuitive from looking at today's systems, and they may be wrong, since there is significant uncertainty about how ML will look in 2030. However, properties (1.-5.) above are my median bet, and whatever GPT_{2030} is like, I doubt it will be "GPT-4 but a bit better".

If I'm right, then whatever the impacts of AI are, they won't be small. We should be preparing for those impacts now, asking what will happen at the largest scales (on the order of $1T, 10M lives, or significant disruptions to social processes). It's better to be surprised now, rather than in 7 years when the system is already being rolled out.

1. Specific Capabilities

I expect GPT_{2030} to have superhuman coding, hacking, and mathematical abilities. I also expect it to be superhuman in its ability to read and process large corpora for patterns and insights and to recall facts. Finally, since AlphaFold and AlphaZero had superhuman abilities in protein engineering and game-playing, GPT_{2030} could as well, for instance if it was trained multimodally on similar data to the AlphaFold/AlphaZero models.

Programming. GPT-4 outperformed a strong human baseline on LeetCode problems posed after its training cutoff (Bubeck et al. 2023, Table 2), and passed the mock interview for several major tech companies (Figure 1.5). The velocity of improvement remains high, with a 19% jump from GPT-3 to 4. On the more challenging CodeForces competition, GPT-4 does less well, but AlphaCode is on par with the median CodeForces competitor. On the even more challenging APPS dataset, Parsel further outperforms AlphaCode (7.8%->25.5%). Looking forward, the forecasting platform Metaculus gives a median year of 2027° for 80% on APPS, which would exceed all but the very best humans.[4]

Hacking. I expect hacking to improve with general coding ability, plus ML models can scour large codebases for vulnerabilities much more scalably and conscientiously than humans.

ChatGPT has already been used to generate exploits, including polymorphic malware, which is typically considered to be an advanced offensive capability.

Math. Minerva achieved 50% accuracy on a competition math benchmark (MATH), which is better than most human competitors. The velocity of progress is high (>30% in 1 year), and there is significant low-hanging fruit via autoformalization, reducing arithmetic errors, improving chain-of-thought, and better data[5]. Metaculus predicts 92% on MATH by 2025°, and gives a median year of 2028° for AI winning a gold medal at the International Math Olympiad, on par with the best high school students in the world. I personally expect GPT_{2030} to be better than most professional mathematicians at proving well-posed theorems.[6]

Information processing. Factual recall and processing large corpora are natural consequences of language models' memorization capabilities and large context windows. Empirically, GPT-4 achieves 86% accuracy on MMLU, a broad suite of standardized exams including the bar exam, MCAT, and college math, physics, biochemistry, and philosophy; even accounting for likely train-test contamination, this probably exceeds the breadth of knowledge of any living human. Regarding large corpora, Zhong et al. (2023) used GPT-3 to construct a system that discovered and described several previously unknown patterns in large text datasets, and scaling trends on a related task in Bills et al. (2023) suggest that models will soon be superhuman. Both of these works exploit the large context windows of LLMs, which are now over 100,000 tokens and growing.

More generally, **ML models have a different skill profile than humans**, since humans and ML were adapted to very different data sources (evolution vs. massive internet data). At the point that models are human-level at tasks such as video recognition, they will likely be superhuman at many other tasks (such as math, programming, and hacking). Furthermore, additional strong capabilities will likely emerge over time due to larger models and better data, and there is no strong reason to expect model capabilities to "level out" at or below human-level. While it is possible that current deep learning approaches will fall short of human-level capabilities in some domains, it is also possible that they will surpass them, perhaps significantly, especially in domains such as math that humans are not evolutionarily specialized for.

2. Inference Speed

(*Thanks to Lev McKinney for running the performance benchmarks for this section.*)

To study the speed of ML models, we'll measure how quickly ML models generate text, benchmarking against the human thinking rate of 380 words per minute (Korba (2016), see also Appendix A). Using OpenAI's chat completions API, we estimate that gpt-3.5-turbo can generate 1200 words per minute (wpm), while gpt-4 generates 370 wpm, as of early April 2023. Smaller open source models like pythia-12b achieve at least 1350 wpm with out-of-the-box tools on an A100 GPU, and twice this appears possible with further optimization.

Thus, if we consider OpenAI models as of April, we are either at roughly 3x human speed, or equal to human speed. I predict that models will have faster inference speed in the future, as there are strong commercial and practical pressures towards speeding up inference. Indeed, in

the week leading up to this post, GPT-4's speed already increased to around 540wpm (12 tokens/second), according to Fabien Roger's tracking data; this illustrates that there is continuing room and appetite for improvement.

My median forecast is that models will have **5x the words/minute of humans** (range: [0.5x, 20x]), as that is roughly where there would be diminishing practical benefits to further increases, though there are considerations pointing to both higher or lower numbers. I provide a detailed list of these considerations in Appendix A, as well as comparisons of speeds across model scales and full details of the experiments above.

Importantly, **the speed of an ML model is not fixed**. Models' serial inference speed can be increased by k^2 at a cost of a k-fold reduction in throughput (in other words, k^3 parallel copies of a model can be replaced with a single model that is k^2 times faster). This can be done via a parallel tiling scheme that theoretically works even for large values of k^2, likely at least 100 and possibly more. Thus, a model that is 5x human speed could be sped up to 125x human speed by setting $k=5$.

An important caveat is that speed is not necessarily matched by quality: as discussed in Section 1, GPT_{2030} will have a different skill profile than humans, failing at some tasks we find easy and mastering some tasks we find difficult. We should therefore not think of GPT_{2030} as a "sped-up human", but as a "sped-up worker" with a potentially counterintuitive skill profile.

Nevertheless, considering speed-ups is still informative, especially when they are large. For language models with a 125x speed-up, cognitive actions that take us a day could be completed in minutes, assuming they were within GPT_{2030}'s skill profile. Using the earlier example of hacking, exploits or attacks that are slow for us to generate could be created quickly by ML systems.

3. Throughput and Parallel Copies

Models can be copied arbitrarily subject to available compute and memory. This allows them to quickly do any work that can be effectively parallelized. In addition, once one model is fine-tuned to be particularly effective, the change could be immediately propagated to other instances. Models could also be distilled for specialized tasks and thus run faster and more cheaply.

There will likely be enough resources to run many copies of a model once it has been trained. This is because training a model requires running many parallel copies of it, and whatever organization trained the model will still have those resources at deployment time. We can therefore lower bound the number of copies by estimating training costs.

As an example of this logic, the cost of training GPT-3 was enough to run it for 9×10^{11} forward passes. To put that into human-equivalent terms, humans think at 380 words per minute (see Appendix A) and one word is 1.33 tokens on average, so 9×10^{11} forward passes corresponds to ~3400 years of work at human speed. Therefore, the organization could run 3400 parallel copies of the model for a full year at human working-speeds, or the same number of copies for 2.4 months at 5x human speed.

Let's next project this same "training overhang" (ratio of training to inference cost) for future models. It should be larger: the main reason is that training overhang is roughly proportional to dataset size, and datasets are increasing over time. This trend will be slowed as we run out of naturally-occuring language data, but new modalities as well as synthetic or self-generated data will still push it forward.[7] In Appendix B, I consider these factors in detail to project forward to 2030. I forecast that models in 2030 will be trained with enough resources to perform **1,800,000 years of work** adjusted to human speed [*range: 400k-10M*].

Note that Cotra (2020) ° and Davidson (2023) estimate similar quantities and arrive at larger numbers than me; I'd guess the main difference is how I model the effect of running out of natural language data.

The projection above is somewhat conservative, since models may be run on more resources than they were trained on if the organization buys additional compute. A quick ballpark estimate suggests that GPT-4 was trained on about 0.01% of all computational resources in the world, although I expect future training runs to use up a larger share of total world compute and therefore have less room to scale up further after training. Still, an organization could possibly increase the number of copies they run by another order of magnitude if they had strong reasons to do so.

4. Knowledge Sharing

(*Thanks to Geoff Hinton who first made this argument to me.*)

Different copies of a model can share parameter updates. For instance, ChatGPT could be deployed to millions of users, learn something from each interaction, and then propagate gradient updates to a central server where they are averaged together and applied to all copies of the model. In this way, ChatGPT could observe more about human nature in an hour than humans do in a lifetime (1 million hours = 114 years). Parallel learning may be one of the most important advantages models have, as it means they can rapidly learn any missing skills.

The rate of parallel learning depends on how many copies of a model are running at once, how quickly they can acquire data, and whether the data can be efficiently utilized in parallel. On the last point, even extreme parallelization should not harm learning efficiency much, as batch sizes in the millions are routine in practice, and the gradient noise scale (McCandlish et al., 2018) predicts minimal degradation in learning performance below a certain "critical batch size". We'll therefore focus on parallel copies and data acquisition.

I will provide two estimates that both suggest it would be feasible to have at least ~1 million copies of a model learning in parallel at human speed. This corresponds to **2500 human-equivalent years of learning per day**, since 1 million days = 2500 years.

The first estimate uses the numbers from Section 3, which concluded that the cost of training a model is enough to simulate models for 1.8M years of work (adjusted to human speed). Assuming that the training run itself lasted for less than 1.2 years (Sevilla et al., 2022), this means the organization that trained the model has enough GPUs to run 1.5M copies at human speed.

The second estimate considers the market share of the organization deploying the model. For example, if there are 1 million users querying the model at a time, then the organization necessarily has the resources to serve 1 million copies of the model. As a ballpark, ChatGPT had 100 million users as of May 2023 (not all active at once), and 13 million active users/day as of January 2023. I'd assume the typical user is requesting a few minutes worth of model-generated text, so the January number probably only implies around 0.05 million person-days of text each day. However, it seems fairly plausible that future ChatGPT-style models would 20x this, reaching 250 million active users/day or more and hence 1 million person-days of data each day. As a point of comparison, Facebook has 2 billion daily active users.

5. Modalities, Tools, and Actuators

Historically, GPT-style models have primarily been trained on text and code, and had limited capacity to interact with the outside world except via chat dialog. However, this is rapidly changing, as models are being trained on additional modalities such as images, are being trained to use tools, and are starting to interface with physical actuators. Moreover, models will not be restricted to anthropocentric modalities such as text, natural images, video, and speech---they will likely also be trained on unfamiliar modalities such as network traffic, astronomical images, or other massive data sources.

Tools. Recently-released models use external tools, as seen with ChatGPT plugins as well as Schick et al. (2023), Yao et al. (2022), and Gao et al. (2022). Text combined with tool use is sufficient to write code that gets executed, convince humans to take actions on their behalf, make API calls, make transactions, and potentially execute cyberattacks. Tool use is economically useful, so there will be strong incentives to further develop this capability.

> ChatGPT is reactive: user says X, ChatGPT responds with Y. Risks exist but are bounded. Soon it will be tempting to have proactive systems - an assistant that will answer emails for you, take actions on your behalf, etc. Risks will then be much higher.
>
> — Percy Liang (@percyliang) February 27, 2023

New modalities. There are now large open-source vision-language models such as OpenFlamingo, and on the commercial side, GPT-4 and Flamingo were both trained on vision and text data. Researchers are also experimenting with more exotic pairs of modalities such as proteins and language (Guo et al., 2023).

We should expect the modalities of large pretrained models to continue to expand, for two reasons. First, economically, it is useful to pair language with less familiar modalities (such as proteins) so that users can benefit from explanations and efficiently make edits. This predicts multimodal training with proteins, biomedical data, CAD models, and any other modality associated with a major economic sector.

Second, we are starting to run out of language data, so model developers will search for new types of data to continue benefiting from scale. Aside from the traditional text and videos, some of the largest existing sources of data are astronomical data (will soon be at exabytes

per day) and genomic data (around 0.1 exabytes/day). It is plausible that these and other massive data sources will be leveraged for training GPT_{2030}.

The use of exotic modalities means that GPT_{2030} might have unintuitive capabilities. It might understand stars and genes much better than we do, even while it struggles with basic physical tasks. This could lead to surprises, such as designing novel proteins, that we would not have expected based on GPT_{2030}'s level of "general" intelligence. When thinking about the impacts of GPT_{2030}, it will be important to consider specific superhuman capabilities it might possess due to these exotic data sources.

Actuators. Models are also beginning to use physical actuators: ChatGPT has already been used for robot control and OpenAI is investing in a humanoid robotics company. However, it is much more expensive to collect data in physical domains than digital domains, and humans are also more evolutionarily adapted to physical domains (so the bar for ML models to compete with us is higher). Compared to digital tools, I'd therefore expect mastery of physical actuators to occur more slowly, and I'm unsure if we should expect it by 2030. Quantitatively, I'd assign 40% probability to there being a general-purpose model in 2030 that is able to autonomously assemble a scale-replica Ferrari as defined in this Metaculus question°.

6. Implications of GPT-2030

We'll next analyze what a system like GPT_{2030} would mean for society. A system with GPT_{2030}'s characteristics would, at minimum, significantly accelerate some areas of research, while also possessing powerful capacities for misuse.

I'll start by framing some general strengths and limitations of GPT_{2030}, then use this as a lens to analyze both acceleration and misuse.

Strengths. GPT_{2030} represents a large, highly adaptable, high-throughput workforce. Recall that GPT_{2030} could do 1.8 million years of work[8] across parallel copies, where each copy is run at 5x human speed. This means we could simulate 1.8 million agents working for a year each in 2.4 months. As discussed above, we could also instead run 1/5 as many copies at 125x human speed, so we could simulate 360,000 agents working for a year each in *3 days*.

Limitations. There are three obstacles to utilizing this digital workforce: skill profile, experiment cost, and autonomy. On the first, GPT_{2030} will have a different skill profile from humans that makes it worse at some tasks (but better at others). On the second, simulated workers still need to interface with the world to collect data, which has its own time and compute costs. Finally, on autonomy, models today can only generate a few thousand tokens in a chain-of-thought before getting "stuck", entering a state where they no longer produce high-quality output. We'd need significant increases in reliability before delegating complex tasks to models. I expect reliability to increase, but not without limit: my (very rough) guess is that GPT_{2030} will be able to run for several human-equivalent days before having to be reset or steered by external feedback. If models run at a 5x speed-up, that means they need human oversight every several hours.

Therefore, the tasks that GPT_{2030} would most impact are tasks that:

1. Leverage skills that GPT_{2030} is strong at relative to humans.
2. Only require external empirical data that can be readily and quickly collected (as opposed to costly physical experiments).
3. Can be a priori decomposed into subtasks that can be performed reliably, or that have clear and automatable feedback metrics to help steer the model.

Acceleration. One task that readily meets all three criteria is mathematics research. On the first, GPT_{2030} will likely have superhuman mathematical capabilities (Section 1). On the second and third, math can be done purely by thinking and writing, and we know when a theorem has been proved. There are furthermore not that many mathematicians in total in the world (e.g. only 3,000 in the US) so GPT_{2030} could simulate 10x or more the annual output of mathematicians every few days.

Significant parts of ML research also meet the criteria above. GPT_{2030} would be superhuman at programming, which includes implementing and running experiments. I'd guess it will also be good at presenting and explaining the results of experiments, given that GPT-4 is good at explaining complex topics in an accessible way (and there is significant market demand for this). Therefore, ML research might reduce to thinking up good experiments to run and interfacing with high-quality (but potentially unreliable) write-ups of the results. In 2030, grad students might therefore have the same resources as a professor with several strong students would have today.

Parts of social science could also be significantly accelerated. There are many papers where the majority of the work is chasing down, categorizing, and labeling scientifically interesting sources of data and extracting important patterns—see Acemoglu et al. (2001) or Webb (2020) for representative examples. This satisfies requirement (3.) because categorization and labeling can be decomposed into simple subtasks, and it satisfies requirement (2.) as long as the data is available on the internet, or could be collected through an online survey.

Misuse. Beyond acceleration, there would be serious risks of misuse. The most direct case is cyberoffensive hacking capabilities. Inspecting a specific target for a specific style of vulnerability could likely be done reliably, and it is easy to check if an exploit succeeds (subject to being able to interact with the code), so requirement (3.) is doubly satisfied. On (2.), GPT_{2030} would need to interact with target systems to know if the exploit works, which imposes some cost, but not enough to be a significant bottleneck. Moreover, the model could locally design and test exploits on open source code as a source of training data, so it could become very good at hacking before needing to interact with any external systems. Thus, GPT_{2030} could rapidly execute sophisticated cyberattacks against large numbers of targets in parallel.

A second source of misuse is manipulation. If GPT_{2030} interacts with millions of users at once, then it gains more experience about human interaction in an hour than a human does in their lifetime (1 million hours = 114 years). If it used these interactions to learn about manipulation, then it could obtain manipulation skills that are far greater than humans—as an analogy, con artists are good at tricking victims because they've practiced on hundreds of

people before, and GPT_{2030} could scale this up by several orders of magnitude. It could therefore be very good at manipulating users in one-on-one conversation, or at writing news articles to sway public opinion.

Thus in summary, GPT_{2030} could automate almost all mathematics research as well as important parts of other research areas, and it could be a powerful vector of misuse regarding both cyberattacks and persuasion/manipulation. Much of its impact would be limited by "oversight bottlenecks", so if it could run autonomously for long periods of time then its impact may be larger still.

Thanks to Louise Verkin for transcribing this post to Ghost format, and Lev McKinney for running empirical benchmark experiments. Thanks to Karena Cai, Michael Webb, Leo Aschenbrenner, Anca Dragan, Roger Grosse, Lev McKinney, Ruiqi Zhong, Sam Bowman, Tatsunori Hashimoto, Percy Liang, Tom Davidson, and others for providing feedback on drafts of this post.

Appendix: Runtime and Training Estimates for Future Models

A. Words per minute

First we'll estimate the word per minute of humans and of current models. Then we'll extrapolate from current models to future models.

For humans, there are five numbers we could measure: talking speed, reading speed, listening speed, and both "elliptic" and "extended" thinking speed. Regarding the first three, Rayner and Clifton (2009) say that reading speed is 300 words per minute[9] and speaking is 160 words per minute[10], and that listening can be done 2-3 times faster than speaking (so ~400 words per minute)[11]. For thinking speed, we need to distinguish between "elliptic" and "extended" thought—it turns out that we think in flashes of words rather than complete sentences, and if we extend these flashes to full sentences we get very different word counts (~10x different). Korba (2016) find that elliptic thought is 380 words per minute while extended thought is ~4200 words per minute. Since most of these numbers cluster in the 300-400 wpm range, I'll use **380 words per minute** as my estimate of human thinking speed. Using the 4:3 token to word ratio suggested by OpenAI, this comes out to **500 tokens per minute**.[12]

(*Thanks to Lev McKinney for running the evaluations in the following paragraphs.*)
Next, let's consider current models. We queried gpt-3.5-turbo and gpt-4, as well as several open source models from EleutherAI, to benchmark their inference speed. We did this by querying the models to count from 1 to n, where n ranged from 100 to 1900 inclusive in increments of 100. Since numbers contain more than one token, we cut the model off when it reached n tokens generated, and measured the time elapsed. We then ran a linear regression with a bias term to account for latency in order to estimate the asymptotic number of tokens per second.