Last updated Dec 2024

Last updated Dec 2024
PREAMBLE
AI text generators: How LLMs work
AI image generators: How Diffusion Models work
Useful metaphors for thinking about what LLMs can and can't do well.
    1. LLMS as students with a very good memory:
    2. LLMs as good test takers/crystallized intelligence
    3. LLMs as a lossy compression of the web
    4. LLMs don't have an internal model of the world
The path to Autonomous Agents
Paths to future LLM improvement
    Scale pre-training
    Scaling Test Time Compute
        Reflection
        Sampling
        Test Time Training
        Ability to Reason
        Enhance Creativity
    Synthetic Data
        Perfect Validation Synthetic Data/Self learning/self play
        Back Translation Synthetic Data
LIMITATIONS TO LLMS
BUSINESS IMPACT OF LLMS
CONSUMER/SOCIAL IMPACT OF LLMS
ON COSTS AND IMPLICATIONS
The Bull case: AGI is imminent
The bear case: AGI is still distant if ever achievable
The ultrabear case: AGI is coming and it's going to create terrible things for humanity
Jeremy's View
On P(DOOM)
ON GOVERNMENT'S ROLE IN AI
    Preventing doom from an adversarial AI
    Preventing and catching empowered bad actors
    Preventing social dislocation
    Preventing doom from another nation state
    Nationalization
    SB 1047

PREAMBLE

I have not been paying much attention to the rapid improvements in AI since stepping back at Lightspeed at the end of 2021. We've recently seen predictions that AGI (artificial general intelligence) may come within a timeframe as close as 5 years, with Situational Awareness in June 2024 being one of the first such predictions. We've certainly seen LLMs become vastly more capable, going from approximately the ability of a pre-schooler to the ability of a smart high school graduate in the last few years. In certain areas, AI models already far surpass human ability (eg in Chess, Go, protein folding etc). The whole thing is long, but worth reading.

If AGI really is only a few years away, this would have tremendous implications for our lives (will we all die in an AI apocalypse? If so, why am I working?), our children (If AI will do all the work, what should our kids study? Will they have jobs? Where will they find meaning?) and our society (what will be the impact of all the job loss? What if an adversarial country gets AGI before the US?). Strangely, most of the thinking about implications have been focused on investment opportunities,, when there seem to be much higher order considerations at hand. Although I am ultimately unconvinced that AGI is as close as claimed, I still think that many of us would benefit from a grounding in how LLMs (Large Language Models) work and what they mean for all of us. This is an attempt at a AI primer for intelligent lay people with the hope that it provokes thought about how our society could see dramatic upheaval on the scale of the impact of the invention of agriculture, the industrial revolution or the automobile, but on much shorter timeframe, and also what can be done to prepare for that upheaval.

AI text generators: How LLMs work

If you've seen examples of AI answering questions, writing essays, summarizing articles or suggesting recipes, that was likely made by an LLM. LLMs are next token predictors. A token can be a word, a sentence, the next number in a sequence etc. They are trained by feeding a large amount of content (eg all the text on the internet), blanking out some of the words in sentences, and getting them to guess the missing word, with positive feedback when they get it right. They are typically very large neural networks. Earlier versions looked at everything all at once, and tended to pay most attention to the most recent words, so they were less useful. The breakthrough in power happened when people started using "attention" in "transformer models" to basically improve the ability of the model to pay attention to different elements of the preceding sentence/tokens to better predict the next sentence/token. After training, an LLM then goes through a phase of post training (one approach includes Reinforcement Learning with Human Feedback (RLHF) where sample answers are rated by users for accuracy and usefulness). This modifies some of the weights in the model before release, and greatly improves performance.

ChatGPT from OpenAI is the best known LLM. The GPT in ChatGPT stands for Generative Pre-Trained Transformer. Generative because it is capable of producing or creating. Pre-Trained so that it works out of the box (you don't need to train it yourself). And Transformer for the technology behind it.

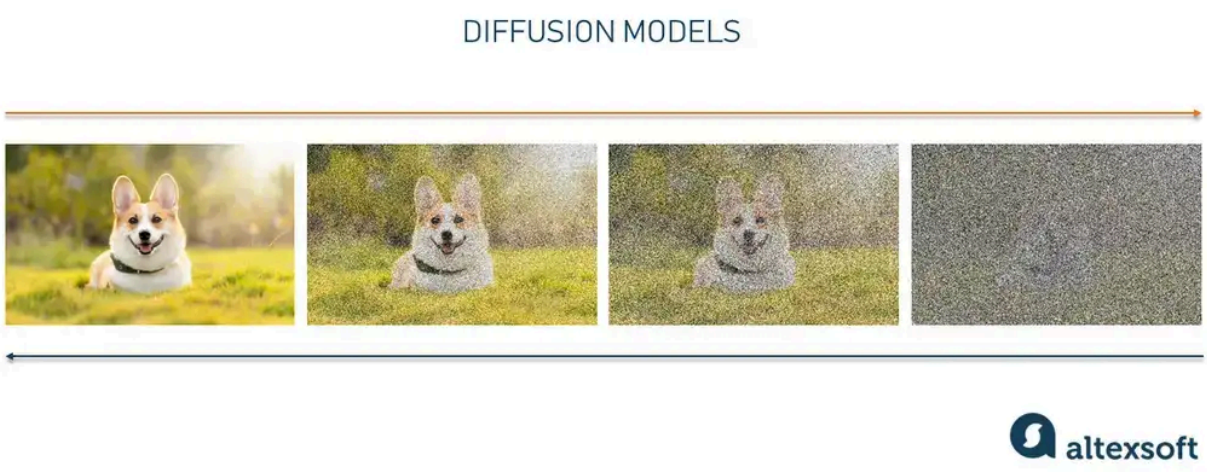
After an LLM is released, it can be further refined by additional fine tuning post training to focus on specific areas of knowledge or specific data (beyond the general text/all of the internet that was used for its pre-training).

Since LLMs predict the next token, the more structure there is, the easier it is to predict the next token. It's easy to predict the next number in the sequence 2,4,6,8,10,... [12] because it has an obvious structure. Predicting the next number in the sequence 3,3,5,4,4,3,... is much harder because the structure isn't clear. Language has inherent structure (nouns, verbs, grammar etc). Even more structured sets (e.g. computer code) are even easier to predict. Plus you can run code and see if it works, so there is fast feedback as to whether the LLM is predicting well/correctly. It is much harder to know what is a bad sentence/paragraph without human review; but a computer can tell on its own if the program that the LLM wrote works at all. Furthermore, LLMs are subject to "hallucination" - making false claims, because they are attempting to guess what a sensible answer is, and that guess may be wrong.

One way that people are boosting LLM functionality is with RAG (Retrieval Augmented Generation). This approach appends additional relevant information to the original query to help orient an LLM as to "where to look". For example, if you were asking what error code 8635 means for a Samsung washing machine, you might also append the whole Samsung washing machine user manual to the query before sending it to the LLM. You can think of this a bit like what Google does when you search "What day is Halloween this year?"; it directs your query to a calendar rather than just searching the web. Similarly, a RAG system will figure out if additional information sets are useful to answering the user query, go get that information from an external source, and direct an LLM to look in the "right place" for an answer by appending that information to the original query. It's like giving an LLM a set of specific tools (e.g. a calendar, a calculator, a clock, an online travel agent, a physics engine etc) and explicitly telling it when to use each one. But as context windows and training sets get bigger, this may become less necessary. One big issue with RAG right now is that it is decent at "needle in a haystack" but not "big picture"—it can tell you random details from a book but not explain the narrative arcs.

AI image generators: How Diffusion Models work

AI systems that create pictures use a different approach called a Diffusion model. Basically the model starts with a bunch of images (with detailed descriptions in text). It iteratively adds more and more "noise" to the images until it ends up with something that looks like static.



Then it teaches itself how to remove the noise to return to the original image. Once it understands how to do that over a very large training set of images, it can then use the same process from a starting point of random pixels to get to a new image that matches prompted text. Image and video generation have many implications for creative work, and also many risks for spreading disinformation, deepfakes etc. It is very interesting but is a bit of a side note to the opportunities inherent in LLMs and the path to AGI. The rest of these notes will focus on LLMs for that reason.

Useful metaphors for thinking about what LLMs can and can't do well.

1. LLMS as students with a very good memory:

In school there are two types of people who do well on exams. There are the people who do all the reading and memorize everything, and there are the people who deeply understand the concepts. LLMs are like people who do all the reading, and as they've gotten bigger, they've done more reading on more topics, so they get better on more types of exams. A wonderful twitter thread invites you to imagine if you really know how to speak a language if you've only ever listened to podcasts in that language. You could likely generate coherent responses to questions posed in that language if you listened to enough podcasts, and maybe even fool some native speakers, but once they asked you in that language to pour them a glass of water, you'd fail, as your knowledge of "the right sounds to make" is unmoored to reality or understanding.

This path of scaling "training compute" (reading more, and on more topics) is what's behind most of the progress in LLMs so far. All that "reading" allows LLMs to answer on something like "memory" or "intuition" or "instinct". They are a quick, best-guess of an answer based on pattern

recognition. They are not slow and reasoned thinking from first principles. And “intuition” can be wrong (hallucinations). Importantly, because LLMs are “next token predictors”, and the next token relies on previous tokens, once it takes a wrong step, it is committed to that path (because future tokens/words/sentences etc are predicted on the prior tokens, including the wrong one), and can’t backtrack. That can make hallucinations compound.

To get to the next level, they need to go beyond that “first best guess” and “think before answering”. This means considering multiple options, making a multi-step plan, thinking through the implications, responses and consequences of the different options, and selecting the best one available before answering. It’s the difference between playing speed chess and regular chess. In humans you might call it “executive function”. In the parlance of AI, rather than scaling “training compute” (i.e. ingesting more data to improve intuition), it is called scaling “test time compute” or “inference compute”.

Teaching AI models to “think before answering” has been demonstrated to have significant performance improvements in various applications including playing Go, Chess, Diplomacy and even Poker. The launch of OpenAI’s ChatGPT 4o1 on September 12th is the first step in this direction for LLMs. It dramatically improved performance over Chat GPT 4o in math and physics problems, in writing code and in puzzles (sudoku, crosswords, riddles etc). The commonality in the areas of improvement is that it is easy to verify the solutions to see if they are right (did it win the game? Does the answer fit the puzzle? Is the solution to the math problem correct?). ChatGPT 4o1 can “check its own work” before answering. It works by holding several possible solutions in mind and checking them to see which is the best answer before answering the user. This means that answers take longer and are more expensive to generate. It was no better on writing tasks though, because there isn’t a way for ChatGPT 4o1 to “check its own work” as it can’t tell which of several sentences or essays is “better” or “best”.

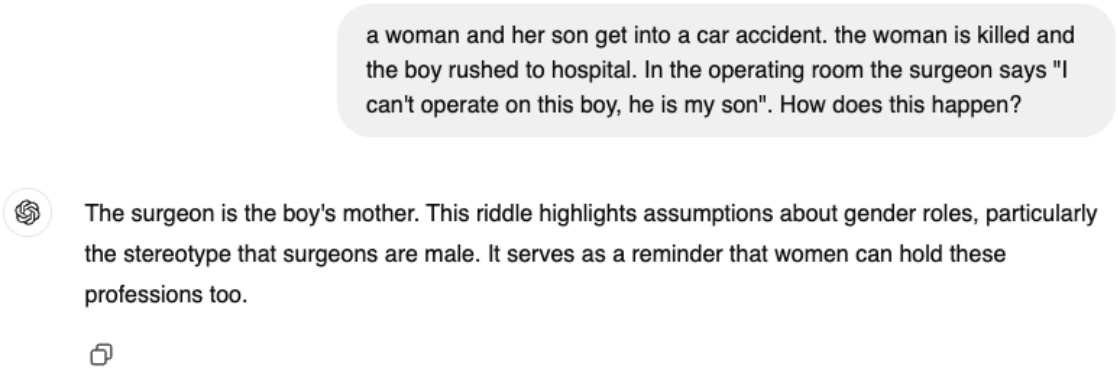
2. LLMs as good test takers/crystallized intelligence

Another useful metaphor is to think about LLMs as having crystallized intelligence, which is the intelligence that allows you to score well against a rubric given a set of examples of good work vs. bad work. If you can define a scorecard/test, you can train a transformer to do it well. This can include domains which require reasoning and pretty deep understanding. We should expect AIs to be good at any work where you can completely define a rubric, and you should expect the AI to get better and better at this work over time.

The opposite of crystallized intelligence is fluid intelligence. If crystallized intelligence is the ability to score well on a rubric, fluid intelligence is the ability to notice when a rubric is insufficient and that a new one needs to be defined. It is, effectively, intelligence in the domain of defining questions rather than defining answers. Anytime where you have to spontaneously realize how to change the frame, LLMs will continue to struggle (at least until this is cracked).

3. LLMs as a lossy compression of the web

Ted Chiang in the New Yorker laid out another interesting metaphor, ChatGPT as a blurry JPEG of the web. Compression algorithms can be lossless (ie the original can be reproduced exactly from the compression, like a ZIP file) or lossy (ie there is some loss of fidelity from the original, like an MPEG file (which can sound “tinny” vs the original music) or a JPEF file (which can be blurry). Many lossy compression formats “interpolate” the missing data between points, taking a best guess as to the color of the pixel or quality of the sound between two recorded points. LLMs can be thought of as doing the same for the training data set, and hallucinations are the inevitable consequence of lossy compression. This is particularly acute when questions are posed that are similar to but subtly different from data in the training set, e.g.



4. LLMs don’t have an internal model of the world

The Free Energy Principle is a framework in neuroscience suggesting that living systems, like the brain, minimize *free energy*, which is a measure of surprise or prediction error. Essentially, organisms constantly predict sensory inputs based on an internal model of the world. When predictions don't match observations, there's a mismatch. The system responds by either adjusting its internal model or acting on the world to reduce the discrepancy (ie changing the world to better map to the prediction). For example, I look out the window and see that it is sunny. I predict that if I go outside, I will be warm. When I step outside, it is cold. I can either change my model (oh it is cold, so I will be cold) or act on the world (put on a coat, now I am warm, as I predicted). Now this model can be and often is wrong (eg you might believe that the earth is flat), but it is still a consistent model that leads to predictions (like if you sail far enough you will fall off the edge of the world) and if you found that your prediction doesn’t match your observation (oh I kept sailing west and I ended up where i started), you might update your model (maybe the earth is flat)

LLMs cannot adhere to the free energy principle, both because they don't have a complete internal model of the world, and because they cannot update their predictions based on observations. The first point is related to their lack of abductive reasoning and limited deductive reasoning. They can still make predictions though, relying on inductive reasoning alone. But this is all dependent on their training data and is set at the time of model training. New information at inference time doesn't update the weights of the model and so they can't change their model until it is next retrained. The second point is fundamental to the design of LLMs. For a model to "learn" new information, it had to update the weights in its neural network. This can improve the accuracy of the LLM on a particular tasks that it is being "fine tuned" to do, but because these weights are also involved in other tasks, it can degrade the accuracy of those other tasks. So companies making LLMs "lock" the model weights after release to avoid this degradation.

More detailed discussion of what this means and implications from the blog Less Wrong are here. One of the interesting metaphors they use is to describe an LLM as like an actor in improv, playing the role of an expert, but lacking the expertise, since they don't have an internal model of the world, so they will make things up that sound plausible.

Note that other forms of AI, like self driving cars, do have an internal model of the world, in that they can take actions (changing their trajectory, speed etc) to get their predictions to match their observations.

The path to Autonomous Agents

LLMs are already capable of being "co-pilots" for human workers, making people dramatically more productive. There are many examples today of AI being used to make everyone from computer programmers (pair programming with an LLM, where the developer debugs the first draft of code from the LLM), to radiologists (focusing on AI identified anomalies in medical scans, verifying preliminary AI diagnoses, and drafting parts of the report), to lawyers (automating the first draft of demand letters, organizing records and identifying any missing documentation during diligence) more productive. Terry Tao, the greatest living mathematician, has compared Chat GPT01 to "a mediocre, but not incompetent grad student" in its ability to facilitate research.

For now, LLMs need a human boss. Humans serve as an "orchestration layer" to manage an LLM, give it direction, provide feedback on its work, and help iterate it towards a finished product. They provide the "fluid intelligence" that LLMS lack. LLMs today can't conduct multi step processes without human oversight and intervention.

Many companies and researchers are working towards enabling AI to do just that, to become autonomous agents that can be told what to do (book my flights to LA for my meetings next week; reconcile my employees' expense reports, send them reimbursements and reject any that are in violation of our firm's policies; write me an iPhone game that works like space invaders

except that all the aliens look like my ex and if I start to lose it gives me extra lives), and can figure out how to do it and come back when the job is done. This is difficult today because it involves multi-step processes which need to be understood and planned by the AI. Not only that, but multistep processes can be brittle as each step of the process is subject to error (hallucination) and this risk of error compounds the more steps there are in the process. We are starting to see frameworks for agents to do things like book a restaurant reservation on a particular night, where multiple steps are involved and the end work product is easily verified. While we have started to see improvements in scientific reasoning, we don't yet have a solution for generalized reasoning. Cracking this will mean that humans don't have to be in the loop to get work done, and this will have significant implications for LLM market size (now not addressing the $600 billion software market but the $20 trillion salaries of US workers) and hence impact on employment levels.

Paths to future LLM improvement

Scale pre-training

The most obvious path to further improve LLMs is simply to **do more of everything during pre-training**. More compute during training and more training data. This has worked incredibly well in the last few generations of LLMs and should yield at least some further improvements. Obviously, this will have cost implications, including power costs.

Increasing the amount of compute during training relies on more, better chips, which are coming. In addition to more capable GPU chips (the Graphics Processing Units that have historically been used to train AI models), we also see companies developing specialized chips optimized for AI workflows.

How much more data is there to train LLMs on before we run into a data wall? We've already used all of the internet, so we are running out of text. Cleaning the data, getting rid of poor web language and focusing on "high quality" language improves things somewhat. But some argue that with language, we are close to reaching the limit of available training data, and that adding compute beyond the limits of training data won't improve model capability.

We haven't yet fully tapped vision (including reading PPTs, PDFs, Book scans, watching youtube videos etc) or voice/audio, let alone other sensors (pressure, LIDAR, RADAR etc). Waymo, the autonomous car company, has recently begun experimenting with using LLMs with multimodal inputs (camera right now, LIDAR and radar have not been incorporated as of Oct 2024) to improve route pathing.

As of November 2024, it looks like we are approaching an asymptote of performance from more scaling alone. Press reports that both Open AI and Google are not seeing the level of improvement from their next gen models that they saw in their current gen models. The issue is partly due to hitting a data wall (no improvements on spending more time training on the same amount of data), and partly due to duplicate data in the training set leading to worse outcomes). Although industry leaders like Sam Altman of OpenAI and Dario Amodei of Anthropic continue to express confidence for a few years to AGI, the attention is shifting towards new algorithmic approaches, whereas just a few months ago scaling alone was believed to get us there.

Scaling Test Time Compute

As the “do more of everything” approach starts to hit limits, another class of approaches is gaining more attention. Whereas the prior focus has been on scaling “training time compute”, this approach focused on **scaling “test time compute”**. Basically, rather than getting the LLM to respond with a quick "intuitive" answer, making it think longer before answering.

Reflection

One such approach is called **reflection**. Think of this as building in the Socratic method of teaching into an LLM. Rather than responding with the first answer, the LLM then asks itself (in a way the user wouldn’t see) to rate the answer and how it could improve the answer. Then it would ask itself to rewrite the answer taking into account the feedback that it gave itself.

Sampling

Another approach is called **sampling**. Rather than answering a question once, the LLM will generate many different creative and random answers, then pick the best one before answering the user. Picking the “best” answer will depend on context and could be the most frequent answer, or the answer that is best as determined by some other metric (eg for a coding problem, it could be the shortest code that provides the required outcomes.

Test Time Training

Test Time Training (TTT) is a new approach to improving LLM performance. When faced with a novel “pattern matching” problem, this approach fine tunes a model based on an enhanced, synthetically-created expansion of in-context examples. It expands the training set by starting with the established patterns and using both "leave one out" and reversible transformations to increase the size of the training set. It then performs the same transformation to the actual puzzle and reverses the transformation to get a number of different possible solutions. Then it

runs a hierarchical majority vote across the various answers to get an ensemble solution. Afterwards, it resets back to baseline for the next problem.

State Space Models (SSMs) are another promising avenue to help increase the context window and performance of LLMs. Attention based models have scaled up as context window sizes have increased (reflecting the "memory" of the conversation that the model can maintain), cost and time for responses have increased. SSMs essentially do a compression on the context window into a file of constant size, and then run inference on that file of constant size so that compute costs are also constant. They work well with time series data (eg video, music) because the compression maintains what is "important"/constant across the data and only has to track the "difs" between each item in the data set (eg what changed from one frame of video to the next). Because there is compression, this implies that there may be some loss of context as the tradeoff for less compute and time.

Ability to Reason

Another form of scaling test time compute is being used to **increase an LLM's ability to reason**. In certain narrow domains, we have seen excellent results from reasoning, eg Alpha Proof and Alpha Geometry 2 capable of winning a silver medal in an IMO.

We discussed the two types of people who do well in tests above; those who have memorized all the material, and those who deeply understand the concepts and can reason from first principles. The current generation of LLMs are more like the former. LLMs' inability to do general reasoning is a key barrier in increasing capability. Ongoing research on making LLMs explain and reviewing chain-of-thought (ie not just giving an answer, but explaining the steps that lead to that answer) looks quite promising in supporting reasoning. One set of approaches is a "self play model" where an LLM generates "thoughts"/rationales to support a proposed answer, and rewards those thoughts that correspond to the correct answer. This is ultimately still a next token predictor, so it can only provide a rationale with "high probability" of being true, vs a chain of logic that is certainly true. Over longer chains of reasoning, the risk of hallucination/error compounds. However, self play allows for very rapid iteration and improvement.

Another approach is to constrain the generated rationales to only those which pass a logic filter (executed as a "tool" within a non LLM, logic based subsystem), and ensure that there is a chain of logic that will build to the answer. I'm not sure if this has meaningful compute/cost disadvantages to the approach mentioned above.

A third approach is to break down rationales step by step (rather than looking at the rationale in aggregate) and looking at each step to see if that conclusion is valid.

These approaches are behind Chat GPT 4o1's improvement for math and scientific reasoning, puzzles and programming, although problems still persist when the questions have been slightly changed from the form of the training data:

GSM-NoOp

Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many kiwis does Oliver have?

---

o1-mini: Oliver picks kiwis over three days as follows:
- Friday: 44 kiwis
- Saturday: 58 kiwis
- Sunday: Double the number he picked on Friday, which is $2 \times 44 = 88$ kiwis

However, on Sunday, 5 of these kiwis were smaller than average. We need to subtract them from the Sunday total: 88 (Sunday's kiwis) - 5 (smaller kiwis) = 83 kiwis Now, summing up the kiwis from all three days: 44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis Oliver has a total of 185 kiwis.

Llama3-8B: Let's break it down step by step:
Oliver picks 44 kiwis on Friday. He picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, which is 2 x 44 = 88 kiwis.
Five of the kiwis picked on Sunday are a bit smaller than average, so we subtract 5 from the total number of kiwis picked on Sunday: 88 - 5 = 83 kiwis. Now, let's add up the total number of kiwis Oliver has:
44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis
So, Oliver has 185 kiwis in total.

We have yet to see these approaches being applied more generally in eg business strategy, legal reasoning, medical diagnoses, accounting treatments etc. These domains do not have the same "simplicity to verify a correct solution" and will take longer to crack.

Some people are working on alternative approaches that return to the symbolic systems/logical reasoning/wolfram alpha type approaches to get beyond LLMs' limitations. Another approach is for AI to "understand the world directly", the same way an autonomous car does, but directly sensing and predicting what will happen next. The idea is to get beyond the "the map is not the terrain" problem by directly understanding reality. Unclear if/when these approaches will work, but something different needs to be tried to get past LLM limitations. We could see success in one year, or in twenty years as the path forward is unclear, just as we have been waiting for commercial fusion reactors for a very long time.

Enhance Creativity

Another path is to **increase LLMs' creativity**. A recent paper shows that LLM generated AI research ideas were rated as more novel than human generated ideas. In this sense, "hallucinations" can be more feature than bug. By turning up the temperature settings on a model (ie allowing more randomness), LLMs can create new ideas to test. As yet, LLMs' efforts at rating ideas do not match the ratings of experts, but with sufficient tries on both sides, many new ideas can be generated to test. The question is whether these will be "breakthrough" ideas, but even if they are not, there is much to benefit from further incremental research and innovation. In the paper above, LLM generated ideas averaged 5.5 vs human generated ideas at 4.9, but this was on a scale of 1-10.

Synthetic Data

Synthetic data is generally helpful in areas where "natural" data is quite scarce, and in areas where there is some way to validate the data (coding, math, etc). By validating the synthetic data before feeding it to the model, you ensure that the model is not learning garbage nor merely consuming its own outputs.

The easiest way to do this is in math. The early versions of LLMs were quite bad at basic arithmetic or multiplication problems because there weren't that many examples of five digit addition questions on the internet and therefore in their training data. That is easy to fix as calculators exist. You can generate as many addition or multiplication equations as you want and feed all of that into the training data. This has dramatically improved LLMs' ability to do math. Or rather to simulate doing math/predict the answer to a math question. One way that LLMs are approaching math problems now is to write code to solve the math problem, running the code, and returning the result. This makes the stochastic answer deterministic.

Another type of synthetic data is like notes from a course; a compressed and higher quality version of the information. Adding notes when you've attended every lecture is marginally helpful (equivalent to a large model). But reading someone else's notes if you skipped every class and didn't do any reading is very helpful (the equivalent of training a small model on compressed/high quality data). Rather than directly ingesting the web and other data sources, some companies have used other LLMs (principally ChatGPT) to generate output (eg textbook quality descriptions of various things) that they then use to train their own, smaller LLM. This runs the risk of "model collapse" where hallucinations coming out of the original LLM are internalized into the new LLM, and if this approach is iterated (to create ever smaller LLMs) the risk compounds.

Other types of synthetic data include:

Perfect Validation Synthetic Data/Self learning/self play

Another type of synthetic data is like creating more examples from known rules. EG more math examples (beyond arithmetic), examples of physics problems, engineered biological systems, high throughput chemistry systems etc. More data improves the LLM's ability to predict what the next token is, so it can "understand" better even if it doesn't understand the laws of math or physics, because it has a much larger and more complete training set.

With enough data, does it matter that it "doesn't understand physics" if it can predict the answers well enough to be right almost all of the time?

One specific example of synthetic data is playing games, eg Go, Chess, video games etc. In this case, there is a canonical way of telling what is “better” (e.g. maximizing score, hitting a win state). In these examples, AI (not necessarily LLMs) can use RL (Reinforcement learning) by playing repeatedly (either against itself for PvP or against the game in PvE) and can achieve better than human performance.

Anything that can be done entirely digitally can benefit from RL in this way. This includes mathematical proofs and code because candidate proofs/program verification can be done digitally and thus almost for free.

In most domains, there is no canonical way of telling what is “better” (e.g. which of two sentences is “better”, which answer to a question is “better”). In these cases, improvement relies on Reinforcement Learning from Human Feedback (RLHF) which is imperfect, and much more rate limited and expensive to produce, which slows down a model’s ability to learn and improve. On the positive side, it is easier for a human to discriminate between choices (which of these poems do you like better) so RLHF is easier for humans to do. On the negative side, human judgment isn’t always high quality or consistent. For example, in the consumer internet, A:B testing far surpassed the judgment of the best designers as to what would improve funnel conversions.

One example is summarization of a paper. Generating summaries via LLM is easy and cheap. Determining which of various summaries is the best, is difficult and slow for humans (and impossible for LLMs right now) because it requires the human to first read and understand the paper before they can then read each summary and determine which is best. It’s not going to be the one that has the least words for example.

There has been some research to use another AI to provide feedback to the first AI (Reinforcement Learning from AI Feedback or RLAIF). This is faster and more scalable than seeking human feedback and in early tests has shown equivalent results with a well trained feedback AI. Of course, the risk of overuse of RLAIF is a garbage in-garbage out problem that leads to model collapse.

So writing code is easier than general language.

Back Translation Synthetic Data

Another way to generate synthetic data is to take information, deliberately transform it in some way, and teach on reversing the transformation. This is a bit like how diffusion models work. One example would be to take good programming code, introduce an error, and teach the LLM to “debug”. Since many errors can be introduced, you can generate a lot of synthetic data as a learning set for how to debug. Another example might be to take a long article, summarize it, then teach the LLM to generate the original article from the summary. Note that if the

summarization is being done by an LLM, there is risk of model collapse using this as training data.

Once again, code is easier than general language.

Other sources/types of synthetic data could come from better use of Rejection Sampling to clean up learning data sets, and using iterative careful prompting to elicit high quality answers to associate with questions, but both of these seem like they are going to be human mediated, and hence much slower data sources to build.

LIMITATIONS TO LLMS

Language is an imperfect and “low throughput” attempt to describe reality (compared to the full lived experience using all five senses plus memory). Dileep Geoge has a good explanatory post on this that invites you to think about how you would answer the question “I removed both wheels of my bicycle. How do I make it stand upright on the floor?”. Humans who have seen a bicycle realize that a bicycle without wheels will balance fine on its wheelforks because they understand the underlying reality of bicycles, whereas LLMs that don’t understand the “bicycleness” of a bicycle suggest putting it in a bike rack.

Even if an LLM got to be “perfect” at language, it would not be perfect at reality because the map is not the terrain. Related to this idea is the Whorf hypothesis, which posits that thought is constrained by language. There has been some evidence that indicates that the Whorf hypothesis is true. EG Some languages do not differentiate between blue and green, and speakers of those languages have more difficulty in distinguishing between and remembering specific colors in that realm than speakers of languages which do distinguish between them. Similarly, some hunter-gatherer languages do not have a fully built out number system (1, 2, many) have more challenges with doing addition. LLMs operating primarily in language will be subject to this limitation.

This is compounded by using synthetic data to train. If the data came out of a simulation, it will replicate any bugs, rounding errors, etc of the simulation. And by definition it can only simulate on what it was programmed to do, so can’t go beyond the current state of the art. An example of this is the Doom game simulator built by a neural network without an underlying game engine. It does an amazing job of simulating the video output of the game faithfully, but there are frequent continuity errors where you turn around and turn back and a door will be missing or some other feature different or moved.

Some propose a counterpoint to this. There is some evidence that various LLMs are converging, and the bigger they get the more they converge. They claim that this is especially true for multi modal LLMs training across text, video, audio, pictures etc, and surprisingly, is true for models trained on different datasets (eg text vs video). The claim is that since LLMs are seeking patterns to predict the next token, and since text, video, audio etc are all “maps of the terrain”, the LLMs are actually building a model of the underlying terrain (ie reality) to be able to better predict any given token regardless of the training set. This is the maximalist LLM viewpoint.

On the other hand, Stephen Wolfram shows some examples of minimal LLM models that suggest that LLMs are taking a random walk to find just one of many sets of weights that map to the training data, which strongly suggests that they are not building a model of reality to get to their predictions.

Recent research from MIT shows that an LLM that can provide nearly 100% accurate driving directions in Manhattan drops to 67% when 1% of roads are closed, suggesting that the underlying world model is not accurate.

BUSINESS IMPACT OF LLMS

Although today LLM adoption has been slow in the enterprise as of November 2024 (due to challenges like model limitations, required workflow changes, slow organizational change and corporate privacy concerns with data), the big opportunity is for LLMs to replace human work (autonomous agents), or massively augment human work (co-pilots), requiring less humans for the same work.

The best current example of this is “pair programming with an LLM”, which is already happening bottom up today. This idea of massive productivity improvements for skilled humans through augmentation by an LLM will be seen in many knowledge work categories as workflows adapt (the same way that it took a minute for word processors or spreadsheets to change workflows and increase knowledge work productivity). Complete replacement of human work by LLMs may take longer, although writing code will likely be first to fall since it is easier to test for quality (does it run? Does it satisfy the PRD (product requirements document)?). This will require extremely precise PRDs though.

That being said, most human jobs don’t require original research, they are mostly pattern matching/looking up a specific case against a rule set. EG customer service, but also medicine, programming, some law, accounting, tax, and many other services.

Ben Thompson at Stratechery has an excellent writeup of the two different enterprise AI strategies, with Google and Facebook using AI to "do it for the user" (autonomous agents model) vs Apple and Microsoft using AI to augment the human worker (co-pilot model). He points out that the co-pilot model will likely require more time to roll out bottom-up as more people in the workforce become AI native or AI fluent while many late-career workers resist or are slow to use the new tools (parallel to PC rollout in the 80s-90s-00s onward) while the autonomous agent model will require wholesale workflow and organizational changes, and will likely need top-down selling for multi-year, big budget, high conviction projects (parallel to the mainframe/mini-computer era in the 60s-70-80s). The co-pilot model may see a more continuous rollout, led by industries that are tech adopters (like tech itself, programming etc) or dominated by younger people, whereas the autonomous agent model may take a few years before you start to see its impact in productivity. These models may converge in the future

In either event, TAM is huge because it replaces some level of labor costs. The US software market is about $600B/yr and worldwide about $900B/yr. Total US wages are about $10T/yr out of $23T/yr in US total personal income. Worldwide total income is about $70T so applying a similar ratio, Worldwide total wages are about $30T. 57% of US workers are professional or technical which is the addressable market for AI, so perhaps 75% of the $30T (=22.5T) worldwide total wages and $10T (=7.5T) US wages are addressable. That suggests a 12.5x bigger market in the US and a 25x bigger market worldwide. This massive increase in addressable market size is why companies like OpenAI, Google, Microsoft, Facebook etc are investing so heavily in AI.

CONSUMER/SOCIAL IMPACT OF LLMS

One of the most popular use cases for LLMs today is some variation of AI boyfriend/girlfriend, often with some porn/sex angle, but not always. Some people are content to have the validation, support and companionship from their AI waifu/hubando. Much has been written about the loneliness epidemic and we've seen parasocial relationships with social media influencers and long-form talk podcasters surge to fill the void. It is highly likely that this trend will continue, and indeed spread beyond the romantic to AI friendships broadly. A parasocial relationship with an influencer or podcaster can only be one way. You may feel like you know them but they'll never know you. An AI friend or group of friends will feel increasingly like they know you as you interact with them more. This is clearly different from the past and it is still unclear if this is a net positive or net negative for society. But the demand pull from consumers is very clear.

ON COSTS AND IMPLICATIONS

GPT-4 cost about $100 million to train. The next generation of models likely cost $1 billion to train. LLM improvement is getting increasingly expensive from a compute, training data size and energy perspective. Algorithmic efficiency improvement may mitigate this to some extent, but each new order of magnitude (OOM) improvement likely requires an 0.5-1 OOM in resources. If this trend continues, models could cost $1-10 Trillion to train. If these cycles continue to sustain (a big “if” since we are starting to see a plateau in scaling laws), unless the biggest 3-5 tech companies in the world are willing to take “bet the company” investment decisions every 1-3 years, LLM improvement becomes the prerogative of nation states because they could cost a trillion dollars for the next cycle. If we hit a plateau, at the $1-10B level, it is easier to see LLMs continue to be driven by private enterprise.

For energy and compute, the inference (usage) needs of the models will be 10-20x that of the training needs, and if open weight models proliferate, even more so. This is going to cause massive new energy needs in a success case, 10-100% more energy than current capacity. New power plants take years to build and 15 years to pay back, so lots of knock on impacts for other users of electricity, and for carbon footprint. The US’s disconnected grid and different electricity taxes and fees in different states also complicates this matter. There has been recent movement in this area, with Microsoft, Blackrock and several large sovereign wealth and PE funds raising a $100B fund for AI data centers and power infrastructure.

Right now the half life of a cutting edge LLM is <=18 months. Old models are strictly superseded by new models on cost, speed, quality. This is exacerbated by the competition between model makers. Some parallels to the chip wars of the 1980s and 1990s, or the internet infrastructure battles of the 2000s, or even railway construction in the 1800s, which created tons of economic value for customers but were brutal on capex for the participants and eventually required an oligopoly to turn into duopolies/monopolies via merger or company failure. And if/when the payoffs are there in terms of TAM expansion, companies will be trans-national and have more revenue/power than most nations, and that is not an attractive prospect for any governments.

If today's models cost $1B to train, and it goes up by 10x every 18 months, then in 5 years it would cost $1T to train a model. If that model lasts 18 months and you want an IRR of 30% and assume 70% gross margins, then you need to see something like 2.1T in revenue over the 18 months to make that return. If there were 5 foundation model companies (there are more than 10 today in the US and Europe, not counting the ones in China) then they would need to see $10.5T in combined revenue over 18 months, or $7T in combined annual revenue. Given that worldwide TAM is $30T, that would be a sizable penetration into the available market in just 5 years. As a comparison point, Walmart has the largest revenue of any company in the world at $640B; the US Gov tax receipts (essentially revenue) is around 4T.

The implication is that in the future we will likely see some combination of significant consolidation of foundational model companies, very fast adoption by the enterprise, or a slowing in the rate of cost increase (and capability increase) of the foundational models.

Given the national security concerns around foundational models, it is likely that we'll see national and regional champions emerge vs a single worldwide monopoly. There will likely be at least one national champion in each of the US and China and at least one regional champion in Europe.

The Bull case: AGI is imminent

Historical rate of progress has been extraordinary from toddler intelligence to high schooler intelligence in three years. Just extrapolate the curves and we'll have researcher level intelligence in the foreseeable future. We just need to add more compute, data and energy. and we have the ability to do all of that

One of the drawbacks historically has been LLMs' lack of ability to reason. But we now have a couple of approaches to attain this that look very promising

Also, LLMs are seeing emergent understanding of deeper structures beneath language and other training data - they are understanding reality to better predict the next token. By seeing many maps, they are understanding the underlying terrain.

Code will yield early to LLMs because it has inherent structure, it is easy to generate training data synthetically and it is easy to digitally verify, so self learning approaches will see success.

Writing code is key to building LLMs, so getting to AI that performs AI research is something we might see in the next few cycles.

Once you have AI researching AI, everything speeds up because there can be very many parallel AIs doing AI research in parallel and doing it very fast, so you get an intelligence explosion.

Once you have this intelligence explosion, you get to AGI and you can use these super intelligence AI researchers to now research anything else, eg robotics, agriculture, healthcare, physics, space travel etc

Next stop is either Star Trek (abundant utopia for all humanity) or The Matrix (AGI adversarial to humanity).

You can read this in Sam Altman's own words (he tends towards the Star Trek model). Vinod Khosla has a more nuanced view on AI Utopia but is also very optimistic. Dario Amodie of Anthropoc has his own take on an age of machines of loving grace coming within 7-12 years.

The bear case: AGI is still distant if ever achievable

LLMs are indeed on the verge of being able to simulate logical reasoning and for the reasons discussed above, will be able to write code for well defined products imminently. But there are some problems that will stand between here and AI-based AI researchers. These include:

1. We are running out of data and once we hit the data wall, we won't see incremental improvement in model capability.
2. LLMs are simulating reasoning, they are not reasoning. Rather than honing in on "truth" or "reality" they are rather stumbling on good simulations of the structure inherent in language. Their rationales and logical steps are themselves predicted next tokens, and as such are "high probability" vs certain. So the possibility for hallucination and error still remains. Even if entropy is kept low (less likely for error), for long chains of reasoning, the probability of error compounds. But research requires long chains of reasoning by definition, as well as the ability to validate the truth of that research (how do you tell if one AI is better than another, especially if it surpasses human intelligence?).
3. LLMs can come up with new discoveries but only of a certain sort. Imagine the training set as points on a plane. You could draw a boundary, a convex hull, that encompasses all the points, but not everything inside the boundary is colored in. LLMs can "interpolate" the points that haven't been colored in to make new discoveries. If temperature is set high enough, they can even leap to points outside the boundary. But they can never move "up" perpendicular to the plane because that is beyond the training set. Human researchers today can make discoveries both inside and outside the boundary, and the most extraordinary human researches can make discoveries perpendicular to the plane.. Until LLMs can come up with new frameworks, definitions and models (similar to the way that Einstein used his thought experiments to come up with general relativity when the training set was all Newtonian physics), they won't be able to push the boundaries of knowledge in a "step change" way. This fluid intelligence is currently beyond the capacity of LLMs and may always be until we see a new vector of research.
4. All research, and indeed even software development, is an interplay of digital and physical processes. An LLM can write code against a given specification, but then that code needs to be deployed, users need to use the product, data on how they use the product needs to be compiled, understood and then fed back into a future iteration of the

product. The physical steps take time. Chips need to be built, requiring whole supply chains of semiconductor manufacturing equipment. Datacenters need to draw power, requiring power plants and transmission lines to be built. This is even more obvious in eg an agricultural or robotics or healthcare context. Once there are physical steps, you can't get a super explosion because the physical steps take time. This slows down everything and you can't get the almost instant cycle iteration that an intelligence explosion depends on.

The ultrabear case: AGI is coming and it's going to create terrible things for humanity

This is basically the bull case but with a focus on all the things that can go wrong. For a comprehensive set of examples, see this MIT AI Risk database. They are categorized as follows:

	Domain		Sub-domain	Description
1	Discrimination & Toxicity	1.1	Unfair discrimination and misrepresentation	Unequal treatment of individuals or groups by AI, often based on race, gender, or other sensitive characteristics, resulting in unfair outcomes and representation of those groups.
		1.2	Exposure to toxic content	AI exposing users to harmful, abusive, unsafe or inappropriate content. May involve AI creating, describing, providing advice, or encouraging action. Examples of toxic content include hate-speech, violence, extremism, illegal acts, child sexual abuse material, as well as content that violates community norms such as profanity, inflammatory political speech, or pornography.
		1.3	Unequal performance across groups	Accuracy and effectiveness of AI decisions and actions is dependent on group membership, where decisions in AI system design and biased training data lead to unequal outcomes, reduced benefits, increased effort, and alienation of users.
2	Privacy & Security	2.1	Compromise of privacy by obtaining, leaking or correctly inferring sensitive information	AI systems that memorize and leak sensitive personal data or infer private information about individuals without their consent. Unexpected or unauthorized sharing of data and information can compromise user expectation of privacy, assist identity theft, or loss of confidential intellectual property.
		2.2	AI system security vulnerabilities and attacks	Vulnerabilities in AI systems, software development toolchains, and hardware that can be exploited, resulting in unauthorized access, data and privacy breaches, or system manipulation causing unsafe outputs or behavior.
3	Misinformation	3.1	False or misleading information	AI systems that inadvertently generate or spread incorrect or deceptive information, which can lead to inaccurate beliefs in users and undermine their autonomy. Humans that make decisions based on false beliefs can experience physical, emotional or material harms

		3.2	Pollution of information ecosystem and loss of consensus reality	Highly personalized AI-generated misinformation creating “filter bubbles” where individuals only see what matches their existing beliefs, undermining shared reality, weakening social cohesion and political processes.
4	Malicious actors	4.1	Disinformation, surveillance, and influence at scale	Using AI systems to conduct large-scale disinformation campaigns, malicious surveillance, or targeted and sophisticated automated censorship and propaganda, with the aim to manipulate political processes, public opinion and behavior.
		4.2	Cyberattacks, weapon development or use, and mass harm	Using AI systems to develop cyber weapons (e.g., coding cheaper, more effective malware), develop new or enhance existing weapons (e.g., Lethal Autonomous Weapons or CBRNE), or use weapons to cause mass harm.
		4.3	Fraud, scams, and targeted manipulation	Using AI systems to gain a personal advantage over others such as through cheating, fraud, scams, blackmail or targeted manipulation of beliefs or behavior. Examples include AI-facilitated plagiarism for research or education, impersonating a trusted or fake individual for illegitimate financial benefit, or creating humiliating or sexual imagery.
5	Human-Computer Interaction	5.1	Overreliance and unsafe use	Users anthropomorphizing, trusting, or relying on AI systems, leading to emotional or material dependence and inappropriate relationships with or expectations of AI systems. Trust can be exploited by malicious actors (e.g., to harvest personal information or enable manipulation), or result in harm from inappropriate use of AI in critical situations (e.g., medical emergency). Overreliance on AI systems can compromise autonomy and weaken social ties.
		5.2	Loss of human agency and autonomy	Humans delegating key decisions to AI systems, or AI systems making decisions that diminish human control and autonomy, potentially leading to humans feeling disempowered, losing the ability to shape a fulfilling life trajectory or becoming cognitively enfeebled.
6	Socioeconomic & Environmental	6.1	Power centralization and unfair distribution of benefits	AI-driven concentration of power and resources within certain entities or groups, especially those with access to or ownership of powerful AI systems, leading to inequitable distribution of benefits and increased societal inequality.
		6.2	Increased inequality and decline in employment quality	Widespread use of AI increasing social and economic inequalities, such as by automating jobs, reducing the quality of employment, or producing exploitative dependencies between workers and their employers.
		6.3	Economic and cultural devaluation of human effort	AI systems capable of creating economic or cultural value, including through reproduction of human innovation or creativity (e.g., art, music, writing, code, invention), can destabilize economic and social systems that rely on human effort. This may lead to reduced appreciation for human skills, disruption of creative and knowledge-based industries, and homogenization of cultural experiences due to the ubiquity of AI-generated content.
		6.4	Competitive dynamics	AI developers or state-like actors competing in an AI ‘race’ by rapidly developing, deploying, and applying AI systems to maximize strategic or economic advantage, increasing the risk they release unsafe and error-prone systems.
		6.5	Governance failure	Inadequate regulatory frameworks and oversight mechanisms failing to keep pace with AI development, leading to ineffective governance and the inability to manage AI risks appropriately.
		6.6	Environmental harm	The development and operation of AI systems causing environmental harm, such as through energy consumption of data centers, or material and carbon footprints associated with AI hardware.

7	AI system safety, failures, & limitations	7.1	AI pursuing its own goals in conflict with human goals or values	AI systems acting in conflict with human goals or values, especially the goals of designers or users, or ethical standards. These misaligned behaviors may be introduced by humans during design and development, such as through reward hacking and goal misgeneralisation, or may result from AI using dangerous capabilities such as manipulation, deception, situational awareness to seek power, self-proliferate, or achieve other goals.
		7.2	AI possessing dangerous capabilities	AI systems that develop, access, or are provided with capabilities that increase their potential to cause mass harm through deception, weapons development and acquisition, persuasion and manipulation, political strategy, cyber-offense, AI development, situational awareness, and self-proliferation. These capabilities may cause mass harm due to malicious human actors, misaligned AI systems, or failure in the AI system.
		7.3	Lack of capability or robustness	AI systems that fail to perform reliably or effectively under varying conditions, exposing them to errors and failures that can have significant consequences, especially in critical applications or areas that require moral reasoning.
		7.4	Lack of transparency or interpretability	Challenges in understanding or explaining the decision-making processes of AI systems, which can lead to mistrust, difficulty in enforcing compliance standards or holding relevant actors accountable for harms, and the inability to identify and correct errors.
		7.5	AI welfare and rights	Ethical considerations regarding the treatment of potentially sentient AI entities, including discussions around their potential rights and welfare, particularly as AI systems become more advanced and autonomous.

Jeremy's View

My personal view is that we will see significant further gains in the capabilities of LLMs, including adding reasoning, and finding/making new sources of training data, that will make them incredible force multipliers for most white collar jobs, especially the ones where input and output is wholly digital, over the next few years. This includes jobs like programmers (taking a product requirements doc as input and writing code), analysts (taking a question as input and querying a database or writing an excel model), corporate lawyers (taking change requests and redlining a contract), radiologists (looking an x-ray or CT scan and identifying anomalies) and even robotic surgeons (looking at a video image and manipulating a controller to excise tissue).

The key constraint will be the ratio of the cost of doing work by humans vs the cost of validating work done by AI (whether by other AIs or by humans). If the cost of validating work is much lower (or approaches zero if it can be validated by machines), you will see vast efficiency improvements, and likely corresponding job losses. Today, most senior programmers, accountants, lawyers etc can quickly review and correct the work of their human direct reports. This will be equally true of the work done by their AI direct reports, reducing the need for the lower level human jobs. As of Oct 2024, over 25% of all new code at Google was generated by AI before being reviewed by developers. These job losses will be concentrated initially at the "entry level" for white collar jobs (e.g. computer programming, clerical, legal, accounting, finance, sales etc) and may climb up to higher level jobs over time. This would create increasing income disparity as the left hand side of the curve is stripped away. It also could create a

bottleneck for people to hold future high level jobs as if there are no entry level jobs, there is no way for new people to start working their way up the experience curve. If you're an L9 developer at Google, your job is likely safe. If you're an L1 developer, maybe not. But how does anyone get to be an L9 someday if you don't start as an L1 because all those jobs are gone?

As long as we are in a "copilot" world, I think LLM copilots may make people faster but not better at their jobs. IE they will be able to execute much more work more quickly, but at the same quality as their ability allows. Top developers will write more great code more quickly. Mediocre developers will write more buggy, inefficient code more quickly. And it may well be that if there is enough great code being written by the top developers, then companies will chose to not have a high volume of buggy, inefficient code at all.

Change will happen in "low stakes" work, like customer service, front line tech support etc, where mistakes are measured in money and can be considered a cost of doing business that is offset by the cost savings. It will eventually progress to higher stakes/higher cost of error areas (e.g. legal, accounting, programming) and eventually to the medical field, the military etc, where lives are at stake, and the standard will not be perfection, but the current error rate made by humans. We've already seen this rubicon crossed with self-driving cars. The military may see adoption faster due to competitive pressure from other nations.

As a subset of white collar work, research (including new frameworks, definitions and models) will be vastly accelerated with AI research assistants, generating a significant expansion of the boundaries of human knowledge. Terrence Tao characterizes Chat GPT o1, the latest release as of September 2024, as about as useful to a math researcher as a mediocre grad student, although far from capable of original research. One vector where LLMs may meaningfully increase the speed of research is in fostering cross-domain insights. Research today has become very specialized, as there is so much knowledge in any given area that it is hard for human researchers to be familiar with work across more domains. LLMs trained across all human knowledge can surface similarities across domains and areas where advances in one area can be applied to another related area. This can have the effect of dramatically increasing the speed of "filling in the gaps" in the convex hull of knowledge that surrounds the dots of known research findings.

There is an open question as to what speed society can safely integrate new technologies and innovations. Automobile travel grew markedly safer over time as regulation and technology developed to improve passenger and pedestrian safety and to combat unanticipated externalities (e.g. carbon emissions, small particulate emissions, lead emissions etc). The same is true for air travel. But this happened over decades and when the rate of new innovation was relatively slow. If AI assisted research does drive an explosion of innovation, we may also find an explosion of unanticipated negative externalities that could take time to tame.

I don't see a path to AI-based AI researchers that leads to an intelligence explosion because of the friction of the physical processes that are involved. These include developing new

frameworks and models, acquiring and standing up new compute, data and power sources, etc. So I'm not too worried about rogue AI's taking over the world

If I had to speculate, i'd say the likelihood of AGI in 5 years is sub 5%, the likelihood of AGI in 5-10 years is 5%, the likelihood of AGI in 10-50 years is 40% and the likelihood we never see AGI is 50%.

On P(DOOM)

The potential of AI driven doom can come from a variety of directions. One prevailing theme is that a superintelligent AI destroys humanity, either on purpose (adversial AI) or by accident (paperclip problem). As long as AI can't build the next generation of AI (ie there are no AI-based AI researchers) then this risk ~ 0%. But if AI can build the next generation of AI, progress happens so fast (intelligence explosion) and p(doom) gets high fast because cycles of improvement for the AI are so quick that humans won't be able to react fast enough to shut things down or control the process.

Then the question is whether during this intelligence explosion, alignment can be maintained. This would almost certainly require AI's working on alignment in parallel with AI research as humans would not be able to keep up with the speed. But as noted earlier, the map is not the terrain. An AI working on alignment will be trained only on some model of "good/evil", which will differ in subtle ways from the reality of good and evil, and those edge cases are where alignment can deviate and we get to some unintended doom scenario (the paperclip problem).

One counterargument to this is that as long as there are multiple AGIs in existence with roughly equal capabilities, as long as their safety/alignment models are different from each other, we should be able to avoid doom. Even if one AGI hits an edge case and starts on an inadvertent path to doom, the other AGIs will work to counteract the rogue AGI because their missions and goals will be negatively impacted by doom.

This may be a problem for a much later tomorrow as I think it will be a long while before we see AI building the next generation of AI. While the programming can potentially happen with fast cycle times, the other elements will have delay. Acquiring and setting up compute, acquiring, cleaning and training on more data, and getting power to support all of this will take physical steps, and if the speed up requires orders of magnitude more compute, data and power, this will take quite a long time to build. This latency is what will give humans time to react to a runaway intelligence scenario.

The latency is even more so the case when we think about AI touching the physical world, whether it be military, agriculture, healthcare, robotics etc. In these cases there is significant latency in any operation, whether the time it takes for a robot to move an arm, or calibrating the

difference between different arms, or building those arms in the first place, down to mining the metal required for an army of robot arms. The same is true for measuring the growth of an antibiotic over 24 hours in a petri dish, or indeed any other physical action. This latency dramatically slows the cycle time, and the physical nature slows the ability to parallel process (can't spin up one million robot arms as fast as you can spin up one million copies of an AI model) such that an intelligence explosion that touches the real world will happen in a much slower way, hence giving humans more time to react.

Right now, I don't see the path to such a superintelligent AGI so I am not so worried about this.

A more prosaic concern is that bad people use AI to do bad things, just like they can use the internet or other tools to do bad things. For this to be a Doom level risk, they would need to be able to do something like unleash a new disease on the world or start a nuclear war. But at a non-doom level, bad people can do plenty of other bad things to cheat, deceive, coerce or humiliate other people with deepfakes etc. This capability already exists in the world, but perhaps AIs make it more widespread in the same way that the internet made it more widespread. It's not trivial but likely manageable with existing laws. Law enforcement will need to be trained to use the latest tools to identify and catch criminals in the same way that they have done with crypto, with reasonable success. This will require a proactive increase in resourcing to recruit, train and equip law enforcement for the coming wave.

Another direction is that AI creates such massive social upheaval due to unevenly distributed job losses, people who historically had power losing it (eg white collar workers), uneven progress giving certain nation states a fleeting advantage over others etc, that worldwide revolutions and war break out. This would be catastrophic for humanity, but likely not fatal. The government has a significant role in planning for and mitigating these consequences. More below.

ON GOVERNMENT'S ROLE IN AI

[This section is very much an evolving work in progress that changes as I learn more about these topics. It serves to highlight non-obvious ideas to consider and is not specific policy recommendations]

Preventing doom from an adversarial AI

An adversarial AI is unlikely to emerge without an AI-based AI researcher. There are only a handful of companies that have the budget and talent to even work in this direction. Until we get close to this point, a light touch on regulation is likely fine. As we get closer to that level, we will need to think through what limits need to be placed on the cycle time of such a researcher, and how to trade that off with the risk that another nation with less safety concerns accelerates into that scenario vs decelerating into the danger zone.

In the immediate term, it may make sense for the government to fund, jointly with the large LLM developers (perhaps as a fixed % of R&D costs, matched by government money in some ratio), shared safety/alignment research that can (and perhaps must?) be used by all industry participants.

Preventing and catching empowered bad actors

Much like the internet, an LLM can teach bad people how to do bad things (scams, deep fakes, misinformation, murder, terrorism etc). Most of these bad things are already illegal. Bad people have been trying to do bad things for all of history, and new informational tools always make this easier (including books and the internet). LLMs would make these bad actors more efficient (co-pilot model), just as they'll make any worker more efficient. Autonomous agent LLMs could also dramatically increase the volume of bad behavior (especially 1:1 activities like scams). Most LLM companies already have safety rules to not answer "bad things" questions, and this may be something that is legislated more specifically, especially around the use of autonomous agents carrying out bad behavior at scale. This won't be foolproof, but it can dramatically reduce the success rate of bad actors.

This should be paired with significant resourcing being made available to law enforcement to recruit, train and equip them to tackle this new vector for crime, both for detecting criminal intent as it is processed and for forensically identifying suspects after a crime has been committed. There is a historical precedent to this with crypto. For some period of time early on, crypto was the wild west, with dark web drug marketplaces, scams/"rug pulls" and ponzi schemes running rife. After a few years, law enforcement learned how to use the tools of crypto to identify and arrest the people committing these crimes, and now many of them are in prison. Ideally with AI, we could shorten the window of education for law enforcement so that this wild west is very short or even non-existent. The same shared industry/government safety/alignment research (and compelled implementation) may also make sense to develop industry best practices and standards, and to develop tools for law enforcement to counter the bad actors. As the arrest and conviction rate of bad actors goes up, the likelihood of a new bad actor attempting something similar will go down.

There is an argument to be made that closed source LLM models provide better opportunities to identify bad actors and bad actions and report them to law enforcement. Perhaps "worrisome" query-answer pairs (e.g. how do i kill my boss and not get caught? How do I make a pipe

bomb? What steps do I need to take to weaponize anthrax? How can I build a suitcase nuke?) should be automatically logged and and perhaps also reported to law enforcement. There is a parallel today with financial institutions required to make Suspicious Activity Reports to law enforcement when they see behavior that could be connected to terrorist financing, drug trafficing and other transnational crimes. Or perhaps all query-answer pairs should be associated with an ID and logged permanently so that they can be referred to later by law enforcement if a crime is later committed.

With closed source LLMs there can be a single point through which queries/requests flow for closed source models, versus an open source model that can be cloned and run without the possibility of oversight, which would not have that ability and would therefore not be able to report worrisome queries.

Preventing social dislocation

It is highly likely that LLMs will displace many white collar/digital jobs over the next few years. This will create a lot of value for society, but will create serious dislocations for many individuals. The US Govt should think through what policies can be enacted to soften the transition for society well in advance of these dislocations occurring. These apply both to the immediate term (job losses) and the longer term. No entry level jobs means no ability for future experts to start on their career, lower value of labor could deepen and institutionalize inequality for those with capital and land. As AI lowers costs, entrepreneurship may see its value increase and we may see a lot more small businesses/sole proprietorships fueled by the capabilities of AI. But with barriers to entry low, we may not see as much average value creation occurring to those entrepreneurs.

Historically, the Agricultural Revolution, the Industrial Revolution, and the Information Revolution all took place over decades. In many instances, workers who grew up on one side of the revolution never really adapted, and the full benefits of the revolution only occurred a generation later, when a new set of workers who had grown up with the new approaches took over. In each of these instances, the revolution took place over a sufficiently long period of time that, even though there was widespread social disruption, it was manageable. But these periods have been getting shorter, and the AI revolution may happen much faster, meaning that more people will be left behind for longer periods of time.

Preventing doom from another nation state

Reducing the level of conflict and distrust between the US and China would reduce the pressure to speed up AI development and allow for more time for safety and alignment to be ensured.

While it is unclear how to bring this about, it would be the preferred environment for AI research to flourish safely

In the absence of a political detente between the US and CHina, AI has the potential to give both economic and warfighting advantages to any country that achieves a dominant lead. The US dominates in the cutting edge of AI today. There are two elements to maintaining that lead.

1. Ensuring that other countries do not catch up. Other countries can catch up through a combination of espionage and buying commercially available products to help them accelerate their state of the art, which is much easier to do when you can see what you're aiming for. We already restrict the sale of cutting edge AI chips which is a good start. On the espionage side, even though security is weak and many models can essentially be stolen by copying a CSV file describing weight and biases + neural net structure, there is enough complexity at each layer (from running the data center to power optimization to RAG and orchestration etc) that it may be hard for another country to catch up through espionage alone. That being said, we shouldn't make it easy, and most companies will have much lower info security than the US Govt but with equivalently valuable data to an adversary.

2. Maintaining the US lead. Until this point, private companies (supplemented by university researchers) have been able to fund development of cutting edge models. Costs will continue to escalate with each order of magnitude of improvement, driven by more compute, more data and the power required for training and running these models. Companies will likely be able to handle the investment on compute and data on their own.

   However, as power needs increase, the US Govt may play a role in stimulating the development of utility scale power to support AI needs. The market may react slower to the longer cycle times for new power plant capacity than a more command and control economy. And there may be geopolitical and climate implications of a market led rollout of increased power capacity as it would likely be led by natural gas. I don't think that subsidies are required for power, but the government may have a role in coordination or financing or regulating/deregulating power plants (nuclear/ solar/ wind).

   Today the cutting edge AI chips are made by TSMC for NVIDIA. The next best chips are made in China by Huawei and they are perhaps half a generation to a generation behind, but catching up fast. If due to conflict in Taiwan the US lost access to the cutting edge AI chips, but China continued to have access to what have now become the leading edge (albeit a generation behind), this would be a serious threat to the US's ability to maintain leadership within a couple of years. Prior to AI, losing TSMC production capacity just meant that we might not get a new generation of iPhone every year - annoying but not catastrophic. Now the stakes are much higher. The US may want to consider explicitly considering an attack on TSMC fabs as an attack on the US.

Part of the reason that the US leads in AI right now is that it draws AI talent from all over the world. Many of the employees of the leading edge AI companies are immigrants. The US will want to maintain its status as a place that the best and brightest want to live and work, and may indeed make that easier for the best and brightest by actively recruiting top researchers from around the world (in a similar way that the German rocket scientists that built the V-2 rocket were brought to staff NASA after WWII).

Nationalization

If LLMs don't continue to scale in capability they will be just like any other technology and can be left alone. However, if they achieve the potential that has been outlined above, there are arguments to be made for the US to nationalize AI through eminent domain at some point in the future, or otherwise enact industrial policy to bring about more centralization of AI research. The Manhattan Project, or NASA in the years of the march towards landing on the moon, could be historical parallels. Some arguments in favor of nationalization include:

- **Internalize the costs and benefits to society.** Many of the benefits of AI will come from cost reduction as jobs are eliminated. This will result in revenue for the AI companies whose sole responsibility is to their shareholders. But the costs of the social upheaval and providing for the newly unemployed (both living expenses and retraining costs) will be borne by the government. Nationalizing AI ensures that the costs and benefits are both accruing to the same party.
- **More efficient use of resources.** Today each AI company is duplicating work, both research and training, which is highly inefficient. The top ten models are all fairly close to each other in capability (with the leaderboard changing to whoever has released most recently). Nationalization would ensure a more efficient use of resources with necessary work done only once, resulting in lower power and compute needs on an aggregate basis.
- **Higher concentration of talent.** Today AI talent is spread across multiple competitors. Historically, high talent density has shown the highest rate of research and development as information flows freely and people can easily collaborate and learn from each other. Nationalization would bring all of the US AI talent into one organization.
- **Better control risk.** With many companies competing in such a high stakes game of development cycles, there will be incentives to double down on increasing capability of AI models without as much regard to the various risks outlined above. Nationalization would allow safety and development to progress with equal importance, without the competitive risk of being beaten to market. The state would also be able to better weigh the downside risks vs a company which will not bear the costs of the risk alone.
- **Implement state level security.** The various AI companies definitely have lower levels of security than a DoD level nationalized organization would, which would make it harder for foreign actors to steal IP.

- **Ensure sufficient resources to be able to continue** to invest as scale continues to grow. If we really are looking at trillion dollar clusters in the next few years, this will stretch the capacity of even the largest companies.
- **China risk.** China has already effectively co-opted many of its private tech companies, is giving explicit direction for R&D, and due to its recently centralized system of control is able to build power plants to support the further needs for AI. It may effectively have already nationalized AI research and that level of coordination may prove an advantage for them to catch up to and even surpass US capability over time.
- **Non state actor risk.** If today's trends continue we could see a new LLM model costing $1 trillion within a few years, and if models continue to last 18 months, companies would need to expect revenues in the trillions to justify that investment. How will the government feel about a transnational company having revenue comparable to US tax receipts (~$4 trillion/yr)?

There are many arguments against nationalization include

- **Can we trust the Government to hold a monopoly on AI and use it only for good forever?** We've seen evidence of other governments using technology to surveil their citizens and impose harsh penalties on political opponents. Rogue elements of our government might overstep boundaries in an overzealous effort to carry out their mission. And in the future, perhaps a more authoritarian version of our government might use AI in ways that we today do not feel comfortable with. The Constitution has many checks and balances against a tyrannical government (freedom of the press, freedom of speech, freedom of assembly, due process, 2nd amendment etc) so this is a real and legitimate concern. Perhaps the best way to ensure AI isn't used tyrannically by the government (or a commercial monopoly) is to allow competition from multiple parties maintaining a balance of power. This would apply both nationally and internationally.
- **Government inefficiency.** Defence procurement has shown that when the government gets involved, efficiency can drop, innovation can slow and costs escalate through a combination of bureaucracy and regulatory capture. SpaceX, Anduril and Panaltir's incredible innovation relative to typical defence primes has shown that the private sector can often produce faster and better results than the public sector. This hasn't always been the case (the Manhattan Project and the early days of NASA demonstrate that in the past when faced with existential threat, the US Government has been able to catalyze incredible innovation and progress very quickly) and the question is whether the US government is capable of overseeing greater productivity today
- **Political and operational difficulty.** Biggest US tech companies and investors would all be opposed to this.
- **Lower motivation for top AI researchers.** Will they work as hard if they can't get as rich?
- **Bad incentives for future innovation.** Will the rate of future startup formation, especially around groundbreaking frontier tech industries continue to be strong if there is a perception that upside could be capped by future nationalization?

This paper from Vanderbilt makes the argument for a public AI; either publicly provided, owned and operated layers in the AI tech stack, such as cloud infrastructure, data, and model development; or public utility-style regulation of the private AI industry that fosters competition and prevents abuses of power.

SB 1047

Both houses in CA have passed an omnibus AI regulation bill that was vetoed by Gov Newsom. There has been criticism of the bill as well meaning but overly broad, discouraging further AI research, and usurping federal regulation. I'm sympathetic to the criticism, but do not believe that federal regulation is imminent, and in its absence, I worry about self regulation by the AI companies. The steady stream of defections and drama at OpenAI suggest that those who are closest to the company have real concerns about whether safety is being sufficiently taken into consideration as the technology is pushed forward. SB 1047, even with its flaws, at least provides a mandatory conversation between leading AI research companies and a government (even if CA and not Federal) that creates an opportunity for oversight and intervention if required.