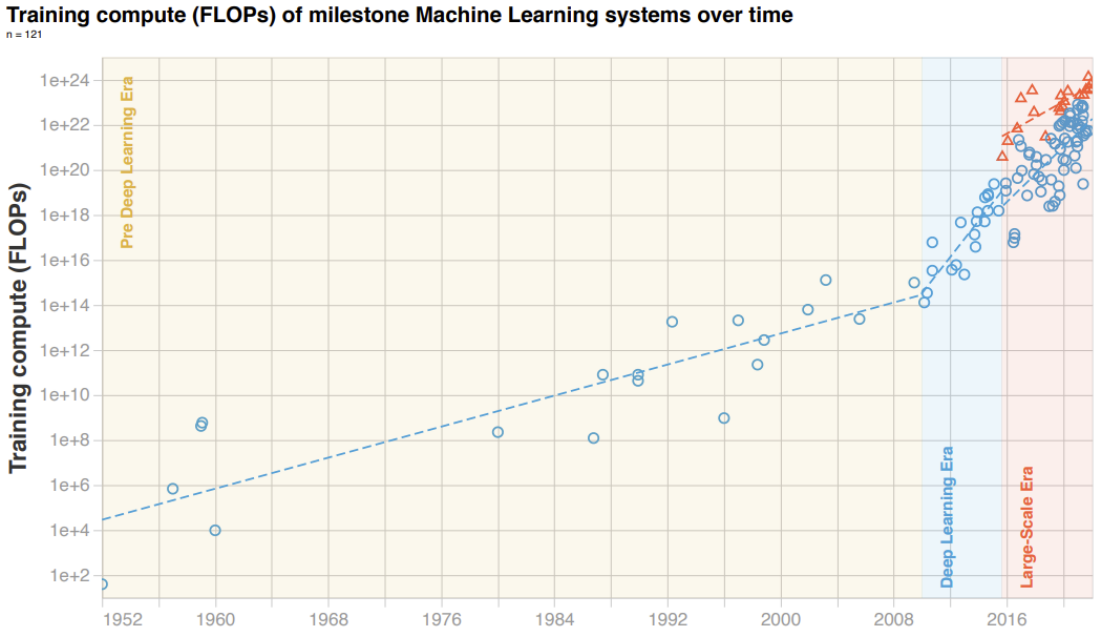
This is an important chart:



.
Engineers have been trying to create real AI for 50 years with limited success. Something changed recently.

In 2017, researchers from Google released a paper which introduced the concept of a Transformer, a new architecture that enabled massively parallelized training. This idea enabled the creation of LLMs – large language models – like GPT-3. GPT-3 was created by OpenAI and is considered by many to be state of the art at understanding language.

These complex models are created by taking the Internet's text, shaping it into a transformer, then running it through NVIDIA chips. Every moment of compute time on these A100 chips costs a few pennies, and over a period of months the numbers add up. OpenAI hasn't released exact numbers, but it is assumed it took tens of millions to train.

The ultimate goal is to create AGI -- artificial general intelligence. The idea is to make a single model capable of doing all human tasks and even more.

OpenAI seems to be the current leader, but it is not alone. Other startups like Anthropic (a spin-out from OpenAI), Character.ai and Adept.ai (two companies by the inventors of the Transformer), Conjecture (started by an open source AI community) have raised between $10M to $580M to create AGI models.

Why this matters, or: defilationary software is finally here!
11 years ago, Marc Andreessen penned his seminal piece about software eating the world. Andreessen's thesis was that every industry would be disrupted by software, and that has certainly been the case. He closed the piece saying "I know where I'm putting my money", and he very much did – his firm,

Andreessen Horwitz, has since amassed $35B in AUM and returned billions through investments in companies like Coinbase, Asana, and Airbnb.

Software ate the world. And yet, the world hasn't gotten much better?

Instead of faxing we email, and instead of calling we text, but where are the actual productivity gains? Has software materially changed our lives, like electricity or the 747, or is it all marginal? It's unclear how it improved our lives in the way Ford or Edison did.

All of this might change very soon with LLMs. Looking back, 1990-2020 might be viewed as a kind of interstitial era for software, like early films without sound or television without cable. We think software is impressive but maybe in hindsight, digitizing reality will be viewed merely as a *delivery* mechanism for AI. The keyboard is not that much mightier than the pen… until AI comes into view.

Take for example a day in the life of a lawyer, which hasn't changed much since 1970. They communicate with clients, write documents, send documents, argue on the phone, sign paperwork, etc. DocuSign is marginally better than fax, but not dramatically so. With LLMs, things become very different: a single lawyer is infused with the power of 100 paralegals. They become an editor, not a creator, supervising the output of a computer model instead of doing the work. A single person becomes five or ten times more productive.

It's not just law. Every producer of information goods – software engineers, pharma scientists, accountants, and project managers suddenly become ten times more capable.

If the last decade was about software eating the world, the next one is about AI powering the software.

Stable Diffusion

We'll now make a small detour to discuss AI art which has taken the world by storm.

In 2021, Emad Mostaque, a hedge-fund manager in London acquired a multimillion dollar NVIDIA cluster on Amazon. Mostaque changed the world by donating his cluster to researchers in the University of Munich. These researchers trained and released Stable Diffusion, an open source model that can turn a series of words into an image, which is the underlying algorithm behind most AI art today. It is completely free, and took the open source world by storm, leading to successive developments around generating video, 3D images, all from a single prompt of text.

OpenAI also announced DALLE-2, its own closed-source image model. Midjourney, a very popular Discord server, also created its own image model with skyrocketing popularity.

Interestingly, the images differ in quality and style. DALLE-2 was trained on stock images and tends to produce very realistic photos, while Midjourney and Stable Diffusion were trained on a broader corpus and produce more artistic images.

The obvious application is around helping creatives prototype ideas – Adobe users can paint storyboards in seconds instead of days.

This will also create new kinds of entertainment feeds like Lexica.art. Since the invention of the cable, mass media has been evolving to become more personalized -- from TV, to Netflix, then TikTok. Generated content seems like the next logical step.

Model size
It is important to understand if useful intelligence requires a large model which takes $10M to train, as that drives which kinds of companies will dominate in the market.

In language tasks, this is an open question. Transformer zealots believe this to be true, and are buying every morsel of NVIDIA they can get their hands on. However, last April, DeepMind released "Chinchilla", a model achieving GPT-3 like performance in a fraction of the size. The Chinchilla paper proved that GPT-3 wasn't trained with enough data, sort of an oversized suitcase with too few items inside. It's possible today's models are much larger than they need to be.

Images models are cheaper to train. Stable Diffusion cost only about $600,000. And as of this month, it's small enough to run on a laptop. This will drive different margin dynamics in the image market; winners will dominate through distribution, not model development.

Moral underwriting
Another unique aspect in AI is that companies are assumed responsible for model output, which can sometimes be offensive.

For example since DALLE-2 is owned by OpenAI, it won't output results considered sexist or pornographic; it's been neutered to produce results that satisfy OpenAI's editorial regime. Stable Diffusion, on the other hand, is completely open source like Linux and not subject to those constraints.

This will be an important distinction to watch, since black-box models will inevitably produce societally unacceptable information goods (just like humans). Companies might need to tame models to project and speak politely, whereas open source won't, and tame models might be… too tame.

Geopolitical consequences
As intelligence grows, LLMs will increasingly catch the eye of governments. Once models start finding cyber exploits, they will become a must-have for any developed nation. Who wins? If more exploration and creativity is required to make AGI, then the West might have the upper hand; it seems far more able at exploration. China on the other hand is fantastic at focus. If Transformers are "the answer", China might win (they are winning some AI leaderboards today).

If one logically plays things out, it would seem AGI results in a huge fight for TSMC as the unique link in the supply chain. The White House seems curiously aware of this, and recently banned NVIDIA A100 and H100 (GPUs made specifically for AI workloads) sales to China.

Optimistically, this sort of competition will yield massive spillover benefits for humanity at large – miracle drugs, cognitive prostheses, and abundant energy. Pessimistically… who knows.

Investment
From Zeiss->ASML->TSMC->NVDA->MSFT->OpenAI->Datasets->ApplicationCo, which is the right part of the stack to invest in?

Complex integration points (famous x86 vs memory Intel pivot) and controlling distribution tends to be a good idea. Top of the value chain startups might win – in consumer, new kinds of feeds and networks.

In enterprise, new kinds of AI-native SaaS that have specialized models for particular use-cases. Copilot, released by GitHub, is an AI that writes code and cuts down development time by 55%. Japser.ai focuses on copywriting text and is now a unicorn startup.

Others are working on verticals like law, tax, and other information goods. Upwork, which does about $4B of "gross services volume", connecting demand to supply for tasks like data entry and research. If a model can do $4B of GSV with 90% gross margins, how much is that worth?

Then there are those selling picks and shovels like Forefront or HuggingFace, attempting to become the AWS of this market. They provide model hosting and "fine-tuning" – teaching models specific tasks for bespoke solutions in the enterprise.

Deeper in the stack, NVIDIA seems to have a decent grasp given the CUDA/Pytorch lock-in, but who knows where that might go. Particular kinds of training data will also be important.

One thing that is important to note: unlike pharma, secrets don't last long in software. Once a company figures out a dark magic training secret, it usually becomes public knowledge in months. This might provide a challenge for the "foundation model" companies. We have a few more thoughts here that we will keep to ourselves for now.

Couldn't come at a better time
Fertility rates are declining, people don't want to work, inflation is rampant, and human capital seems to be on edge. We're in dire need of deflationary goods; productivity multipliers like AI couldn't have come at a better time. The good news is the spirit of technological invention is alive and well in startups; I'm in touch with founders every day that are using this technology to make very powerful & compelling products.

We spent a decade building the grid, and we're about to flip on electricity.

- DG