By: Daniel Gross

The magic of AI-native apps
Or: deflationary 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 the advent of AI, and in particular, LLMs – Large Language Models – algorithms that have human-level writing and understanding capabilities.

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 the advent of LLMs and AI, 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.

And it is often the case that new kinds of use-cases – things that are hard to predict but obvious in hindsight, like Uber or YouTube – are of the largest economic consequence. If everyone has a phone in their pocket, we get Uber. What do we get if everyone has a remote universal worker?

Why is this happening now?

The “AI” of 2022 is very different from 2012. This can be traced to two particular breakthroughs, one in understanding language, the other in images.

Language

LLMs -- Large Language Models -- are machine learning models that are capable of completing language tasks with near-human ability. These models can write customer service emails, summarize contracts, and even write computer code.

LLMs trace their roots back to a breakthrough paper released in 2017, which introduced the concept of a Transformer, a new architecture that enabled massively parallelized training. This idea enabled the creation of large 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 running the internet's public text through NVIDIA chips (the A100 is a particular favorite). Every moment of compute time on these 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.

Commercialization

How do we make money here? Upwork does about $4B of "gross services volume", connecting demand to supply for tasks like data entry, simple design, and research. If a model can do $4B of GSV with 90% gross margins, how much is that worth?

Others have more specialized goals -- Copilot, released by GitHub, is an AI that writes code and cuts down development time by 55%. Japser.ai and Copy.ai focus on writing marketing text. Many others are working on verticals like law, tax, and other information goods.

The loftiest goal is to create AGI -- an artificial general intelligence. The idea is to make a single model capable of doing all human tasks and even more. (The tasks go beyond just language; DeepMind, a Google subsidiary, has spun out a division called Isomorphic Labs to focus AGI-like capabilities on drug discovery.)

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

Model size

Today's language models are very large and expensive to make, but it's unclear if bigger is better. GPT 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.

Other top-tier researchers believe that the Transformer architecture itself is wrong, and that intelligence is achievable through very different approaches (like training on video instead of text).

Images

In 2021, Emad Mostaque, a hedge-fund manager in London acquired a $10M NVIDIA training cluster on Amazon. Mostaque donated time on 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. 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. These companies have led to a cambrian explosion of generated images across the web.

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. (Lexica is a universal feed and search engine for generated images.)

Politeness problems

Another distinction is moral underwriting. Since DALLE-2 is owned by OpenAI, it maintains careful standards around the types of images it produces, filtering out results considered sexist or pornographic. Stable Diffusion, on the other hand, is completely open source like Linux and not subject to editorial constraints. This will be an important distinction to watch with AI, 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.

Unlike text, image models seem much cheaper to train. Stable Diffusion cost only about $600,000. And as of this month, it's small enough to run on a laptop.

Commercialization

One obvious application is around helping creatives prototype ideas -- animators, game designers, and other Adobe users can paint storyboards in seconds instead of days. As technology improves, animated shorts and movies will further accelerate this.

A more novel idea is a new kind of social network. Since the invention of the cable, mass media has been evolving to become more personalized -- from TV, to Netflix, then TikTok. Generated content might be the next evolution. This kind of world building could be very popular with children -- instead of watching bedtime stories with our inner eye, we might start watching it with a naked eye, generating movies by feeding text prompts into computers.

Early innings

The AI revolution is still in its infancy. Many things remain unclear – do language models generalize, or do they remain experts in narrow domains? Do the benefits of AI accrue to incumbents, or do startups stand a chance? What happens to existing jobs? Do those practitioners become more productive, or must they change occupation? There's a lot we don't know, and a lot we need to build.

New kinds of software

One of the least discussed questions in AI is interface. Do AI-native apps look like today's apps, or is it fundamentally different? If computers were designed to be operated by two people simultaneously, with a copilot suggesting actions to the captain, would things look materially different? AI-native companies might break the 60-year old interface paradigm rut we've been stuck in; not much has changed in GUIs since Xerox PARC.

Geopolitical consequences

As intelligence grows, LLMs will increasingly catch the eye of governments. Once AI 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.

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.

The trades?

So how does one make money here? Where does margin accrue in the AGI supply chain? 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. Deeper in the stack, NVIDIA seems to have a decent grasp given the CUDA/Pytorch lock-in, but who knows where that might go. I'm not an expert on that, there are other factors in the story (crypto, etc).

Of note, AGI and biopharma are similar in many ways, but there is one key distinction: secrets don't last long in software. Once a company figures out a dark magic training secret, it usually becomes public knowledge in months.

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 around the world that are making very powerful & compelling products. The future is finally coming. If the last decade was about software eating the world, the next one is about AI powering the software.

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