- Until recently, the history of technology could be encapsulated in a single phrase: humanity's quest to manipulate atoms. From fire to electricity, stone tools to machine tools, hydrocarbons to medicines, the journey is a vast, unfolding process in which our species has slowly extended its control over atoms
- At root, the primary driver of all of these new technologies is material – the ever-growing manipulation of their atomic elements
- Then, starting in the mid-twentieth century, technology began to operate on higher level of abstraction
- In the decades after WW2, scientists, technologists, and entrepreneurs funded the fields of computer science and genetics, and a host of companies associated with both
- Eventually, the technologies matured and gave us everything from smartphones to genetically modified rice. But there were limits to what we could do.
- Those limits are now being breached. We are approaching an inflection point with the arrival of these higher-order technologies, the most profound in history.
- Industry research output and patents soared.
- In 1987 there were just ninety academic papers published at Neural Information processing systems, at what became the field's leading conference.
- By the 2020s there were almost 2 thousand
- In the last six years there was a six fold increase in the number of papers published on deep learning alone, tenfold if you widen the view to machine learning as a whole.
- AI really isn't emerging anymore. It's in products, services, and devices you use every day. Across all areas of life, a raft of application rely on techniques that a decade ago were impossible
- AI is already here. But it's far from done
- A big part of what makes human intelligent is that we look at the past to predict what might happen in the future.
- In this sense intelligence can be understood as the ability to generate a range of plausible scenarios about how the world around you may unfold and then base sensible actions on those predictions.

Language Learning Models

- LLMs take advantage of the fact that language data comes in a sequential order.
- Each unit of information is in some way related to data earlier in a series
- The model reads very large numbers of sentences, learns an abstract representation of the information contained within them, and then, based on this, generates a prediction about what should come next.
- The challenge lies in designing an algorithm that "knows where to look" for signale in a given sentence.

- What are the keywords, the most salient elements of a sentence, and how do they relate to one another? In AI this notion is commonly referred as “attention”.
- When a LLM ingests a sentence, it constructs what can be thought as an “attention map”.
- It first organizes commonly occurring groups of letters or punctuation into “tokens”, something like syllables, but really just chunks of frequently occurring letters making it easier for the model to process the information.
- It’s worth noting that humans do this with words of course, but the model doesn’t use our vocabulary.
- Instead, it creates a new vocabulary of common tokens that helps it spot patterns across billions of billions of documents
- In the attention map, every token bears some relationship to every token before it, and for a given input sentence the strength of this relationship describes something about the importance of that token in the sentence.
- In effect, the LLM learns which words to pay attention to.
- These systems are called transformers
- Much of AIs progress during the mid-2010s was powered by the effectiveness of “supervised” deep learning. Here AI models learn from carefully hand labeled data
- Today’s LLMs are trained on trillions of words.
- Transistors are getting so small they are hitting physical limits: at this size electrons start to interfere with one another, messing up the process of computation.
- While this is true, it misses the fact that in AI training we can keep connecting larger and larger arrays of chips, daisy-chaining them into massively parallel supercomputers. There is therefore no doubt that the size of the large AI training jobs will continue to scale exponentially
- Researchers meanwhile see more and more evidence for “the scaling hypothesis” which predicts that the main driver of performance is, quite simply, to go big and keep going bigger.
- Keep growing these models with more data, more parameters, more computation, and they will keep improving – potentially all the way to human-level intelligence and beyond
- AI researchers are racing to reduce costs and drive up performance so that these models can be used in all sorts of production settings.
- In the last four years, the costs and time needed to train advanced language models have collapsed.
- AI systems now help engineers generate production ready code
- In 2022, OpenAI and Microsoft unveiled a new tool called Copilot, which quickly became ubiquitous among coders.
    - One analysis suggest it makes engineers 55% faster at completing coding tasks, almost like having a second brain on hand
    - Manu coders now increasingly outsource much of their mundane work, focusing instead on knotty and creating problems
- In the words of an eminent computer scientist “It seems totally obvious to me that of course all programs in the future will ultimately be written by AIs, with humans relegated to, at best, a supervisory role”.

- Anyone with an internet connection and a credit card will soon be able to deploy these capabilities-an infinite stream of output on trap