Dear Partners,

After a lot of research, we have a lot to say – 40+ pages – about the emerging field of Artificial Intelligence (AI), and its impact on the business and the investing world. But first, a performance review. As we discussed throughout the year, performance in 2024 remained tepid and certainly not in line with our goals or historical achievements.

	ACM Gross Returns	Gross less mgmt fee (Semi-Net)	S&P 1000 Total Return Performance	Semi-Net Vs Benchmark	Net Excl Clawback (Individual-Year)	Net Incl Clawback	Net vs Benchmark
2016-01-08	69.5%	67.0%	31.2%	35.8%	56.3%		25.1%
2017	-1.8%	-3.5%	15.3%	-18.8%	-3.5%		-18.8%
2018	21.7%	19.9%	-10.3%	30.2%	10.8%		21.1%
2016-2018 Cumulative	102.6%	93.2%	35.7%	57.5%		76.0%	40.3%
2016 - 2018 Annualized	26.5%	24.6%	10.7%	13.8%		20.7%	10.0%
2019	45.5%	42.9%	25.2%	17.7%	37.6%		12.4%
2020	4.6%	3.0%	13.0%	-10.0%	3.0%		-10.0%
2021	53.4%	51.3%	25.4%	25.9%	43.5%		18.1%
2019 - 2021 Cumulative	133.5%	122.7%	77.4%	45.3%		109.1%	31.7%
2019 - 2021 Annualized	32.7%	30.6%	21.1%	9.5%		27.9%	6.8%
2022	-19.5%	-20.9%	-14.0%	-6.9%	-20.9%		-6.9%
2023	35.3%	33.7%	16.4%	17.3%	28.5%		12.1%
2024	4.3%	2.8%	12.3%	-9.5%	2.8%		-9.5%
2022 - 2024 Cumulative	13.6%	8.7%	12.5%	-3.8%		8.7%	-3.8%
2022 - 2024 Annualized	4.3%	2.8%	4.0%	-1.2%		2.8%	-1.2%
2016 - 2024 Cumulative	437.2%	367.8%	170.8%	197.0%		271.4%	100.6%
2016 - 2024 Annualized	20.5%	18.7%	11.7%	7.0%		15.7%	4.0%

Since we're not seeking new clients at the current time, these returns are for discussion purposes only; please check your individual account statements for your specific returns, since individual client account returns can vary due to fee structure, timing of deposits, and so on. Returns reflect Askeladden Capital Partners LP returns when it was in existence; gross returns now reflect the aggregate performance of client accounts, with fees applied as if they paid the standard fee structure. When in doubt, we try to use the more conservative approach for calculations (i.e. reporting lower rather than higher returns.) I'm always happy to provide clients with raw data as necessary for their own analysis. I could go on with disclaimers (we used to have a page), but again, we're not marketing.

With that out of the way: to put it bluntly, our returns have sucked recently (i.e., over the past 3 years). We have, extensively, discussed the extent to which this is partially our fault. We made some mistakes that we shouldn't have; had we avoided those, our performance would at least be *materially better.* Conversely, the market environment has been brutal for small-cap value; many large companies such as Starbucks (SBUX) trade at absurd and unjustifiable multiples despite poor near-term results and a tepid medium-term outlook, while many small caps generating real free cash flow with clearly bright outlooks are left for dead.

There are only so many times I can keep saying this, so I don't plan to keep beating a dead horse and talking about it at length. As I wrote last quarter, eventually, value wins.

Below is a table of contents, then the meat of the letter.

Table of Contents

An AI Deep Dive: The Why

The topic of today's letter is artificial intelligence. This is not one of my short letters; in fact, it is by far the longest in recent memory. But the length is intentional because of the importance of the topic, and the level of depth/nuance in which I'd like to discuss it.

While we do get into quite a few technical details, this is a business case analysis; I am focusing on the technology itself only to the extent that it's important. Turns out some aspects of the technology are critical to understand, so we do spend a fair bit of time on some technical concepts. (It helps that my wife, a PhD in applied mathematics, actually did her dissertation research in an area that could be considered AI-adjacent, or utilizing underlying techniques that also have applications in AI. Along the way, I picked up on some things from many years of discussions.)

This is an investment-focused rather than employment-focused analysis, except to the extent that the two intersect, which is of course considerable. Similarly, while I am publishing this letter for public consumption, I am not trying to become or claiming to be any sort of expert or a "thought leader" on AI; this is simply me synthesizing my views on the topic and communicating them to myself, my friends, our investors, and anyone else who might find them interesting, especially since I have a different perspective than most. Future letters will cover different topics, although our portfolio commentary will include increasing focus on AI-driven opportunities and risks, alongside other fundamental factors.

Anyway. So far, I think most people's attitudes towards AI have tended towards one of two extremes:

- the hyperbolic, tech-bro AI maximalist take (*"AGI is around the corner, AI will put everyone out of a job in three years!"*)

- or the dismissive Luddite take (*"AI is a fad that nobody will remember in three years, we don't need to think about it, let's wait and see how it plays out."*)

As a value investor who focuses primarily on real cash flows that we can analyze in the here and now, my natural inclinations are the latter. i.e., skepticism towards whatever is the current "hot thing" – especially since, as we discuss later, a lot of former "hot things" have turned into more of "nothings."

However, given the potential of AI to fundamentally reshape many aspects of the world, I thought it was important to approach the topic with an open mind. Our job is to assess the world probabilistically, and expected values are the sum-product of probabilities and magnitudes. Even if you believe there is only a 5 or 10% chance that AI reshapes the world, the magnitude of the potential change is substantial enough to justify spending time on it.

What I learned surprised me to the upside. I'll give you a teaser – I think my research process will change more in the next three months than it has in the past ten years, and I'll be 2-3x as productive. Based on that personal experience, it is hard not to extrapolate potentially substantial impacts of AI on knowledge work writ large, though historical base rates as well as competitive factors suggest that the way it actually plays out in the real world is likely quite a bit less dramatic than the AI maximalist hype that's out there (your job is probably safe).

I will also mention that, taking pages from Philip Tetlock's *Superforecasting* and John Lewis Gaddis's *On Grand Strategy*, these are convictions held lightly – I plan to be nimble, open minded, and re-evaluate as new data arrives, particularly in a field evolving as quickly as AI. I'm trying to be a fox, not a hedgehog. Some of what I say here will undoubtedly prove wrong with the benefit of time. Which is fine, as long as I don't still believe the wrong thing when it becomes obviously wrong! (If I'm wrong, please shoot me an email and tell me why; I welcome and value feedback, particularly on a topic whose importance is of this magnitude.)

I think the *why* is as important as the *what* (but if you don't, please feel free to skip this section and head to Page 7 – it won't hurt my feelings!) Clients and long-time readers know that I'm very focused on learning from my mistakes; I will always make mistakes, but I'd at least prefer to make *new ones* – instead of the same ones over again. Let's start by reviewing two of the largest and most impactful mistakes I've made during my investing career.

Mistake 1: COVID

I was slow to recognize, assess, and respond to COVID. I spent approximately zero time thinking about it in February 2020, because considering the impact of a rapidly-spreading virus did not fit neatly into my little analytical box of researching investments from the bottom up. But it turns out that even the hottest restaurant in town, with the greatest ambience and the Michelin-starred chef, can turn into a ghost town due to external circumstances you never would have imagined when opening the place.

In the long term our portfolio worked out fine, but you can also walk away from a game of Russian Roulette unharmed. That is to say, there are counterfactual worlds in which things could have played out a little differently, or a lot, and the temporary – but very real – impacts of COVID – could have permanently impaired a number of our portfolio companies, notwithstanding that they entered COVID with robust balance sheets. There is simply no reasonable way for any business to prepare for your revenues dropping to near-zero, overnight, due to a non-insurable event.

Meanwhile, in the world that actually was, I was too quick to sell obvious "COVID winners" in favor of doubling down on the beaten-down "COVID losers." While, in aggregate, the latter bucket performed quite well over time, we would have had a lot less stress – and higher returns – had we constructed a more balanced portfolio that was less dependent on a single factor (i.e., the speed of the world reopening).

A final point is that – both from a fundamentals and market sentiment perspective – I spent a little too much time focusing on the way the world should be rather than the way the world was. Many (certainly not all) of my conclusions on COVID proved accurate over time, but in the vein of *"do you want to be right or make money*," it didn't really matter – or help – at least from an investing standpoint.

To be extremely, 100% crystal-clear, I am not suggesting leaving your brains behind and participating in obvious bubbles on the basis of the greater fool theory, but I am suggesting that if everyone around you is behaving irrationally, you shouldn't stubbornly believe that they will suddenly behave otherwise (i.e., I am treating it as more of a gating factor or risk to consider, in the context of fundamental value investing). Ironically, as someone who adores all the fun behavioral economics stories in Richard Thaler's excellent book *Misbehaving*, I spent too much time thinking about Econs and not enough thinking about Humans.

After COVID, I spent a lot of time thinking about that experience, and my mistakes, in the context of something I'd once read somewhere – but could never find again – about Lisa Rapuano's experience during the global financial crisis. I've referenced this discussion previously; with the help of Grok3, I finally found it again. Here's an excerpt from a 2011 interview with The Manual of Ideas:

> *At the end of 2008, I identified the drivers of what made us do poorly, and which ones could be improved upon and which ones couldn't. One has to be careful in a situation like this to distinguish outcome from process. If the bad outcome was a result of a bad process, you should fix, change and adapt. If the bad outcome is a result of bad luck, but the process behind decisions was solid, you really can't do much about that. Unfortunately, 2008 had elements of both for Lane Five.*
>
> *First, I'll tell you what we did not do. We did not decide to become more macro-oriented. I believe that macro forecasting is impossible, and the best one can do is try to understand what is actually happening around them, not forecast it. I watched lots of value investors divert hours of time away from valuation and stocks into macro speculation, and I did not and do not think that is productive. Our philosophy has been to understand, and I did not change that. I could have vastly improved my execution of that understanding in 2008, but that is not a change in process, that's just doing what you do better.*
>
> *We did stress-test some of our valuation techniques and found that with our long-term orientation we were too prone to "look through" things when they started to deteriorate. We re-emphasized our commitment to using multiple valuation tools, including ones more focused on two- and three-year time horizons in addition to the five to ten-year models.*

History may not repeat, but it certainly rhymes. There are multiple elements of this discussion that I think are particularly relevant today, and we will return to it a bit later.

Mistake 2: Underestimating Exponential Growth and Technology

My mistakes during COVID were, like the virus, acute; my mistakes pre-COVID were chronic. As I've discussed extensively in prior letters (and thus will only very briefly mention here), like many value investors, for a long time, I underestimated the power of growth and technology. It's rather ironic that value investing is based on the idea of compounding, but many of us used to ignore that factor when it came to revenue growth, pooh-pooh paying a somewhat higher earnings or cash multiple for a business with *much* higher growth rates underpinned by blindingly obvious secular trends. Early in my career, I instead spent (*wasted*) a lot of time researching stodgy businesses at nominally low multiples, consistently underestimating the medium to long term tailwind and impact of technology.

I eventually got smart enough to decide to try not to make any bets against obvious technological or cultural trends (for example, it is non-obvious to me that electric vehicles are intrinsically superior to combustion vehicles, but nonetheless it seems more reasonable to operate on the assumption that ICE vehicles are dinosaurs going extinct than assume that they have a long and healthy life ahead of them.) We're increasingly trying to position ourselves to *benefit* from technological trends, without paying up for them, by finding companies aligned with secular tailwinds that simply don't have the hype or excitement already baked into their share prices.

Why Now?

Let's start here: I've spent a lot of time thinking about *human* intelligence in my life, primarily due to personal bias. Intelligence was, is, and always will be the foundation for my competitive advantages in the world. I am not an athlete, nor a social butterfly, but I did start community college at 13 and finish my MBA by the time I was 20. I process and synthesize information very quickly. (As does my wife; did I mention she's a PhD?)

Intelligence is not the end-all, be-all in investing; as Buffett once pointed out, LTCM's team was comprised of some of the smartest people in the world (literally) – and they blew up. (And I'm not nearly that smart; multivariable calculus overwhelmed me in high school, and quantum mechanics was… traumatic.)[1] Too much intelligence can in fact be a bad thing; smart investors often seek out complicated or difficult investments because of the challenge, when sometimes the best investments are simple. Process, temperament, and many other factors come into play.

Nonetheless, intelligence certainly helps. More broadly, intelligence and cleverness – not physical strength – combined with culture (i.e. the ability to crystallize and build upon that intelligence from generation to generation, rather than reinvesting the wheel), is the major factor that has resulted in us being the planet's keystone species. While there are certainly other animal species that display intelligence and even culture in a recognizable form – worthwhile reads here are Peter Godfrey-Smith's *Other Minds* and Jennifer Ackerman's *The Genius of Birds* – it's fair to say that intelligence is a large part of what makes humans feel unique.

I've also spent a lot of time thinking about technology, at least in a personal rather than investing context. I like technology, and I have always tried to maximize my use of it. I don't like cleaning, so I have Roombas (plural) that mop and vacuum. Instead of replacing my old and underpowered gas stove, I bought two Breville Control Freak induction burners, which were substantially cheaper and vastly superior. (Seriously, they're certifiably awesome.)

And, as we'll discuss in some depth later, I don't like wandering around turning on and off lights, so almost all the lights in my house are motion-activated, and/or can be turned on and off using Alexa. And so on. As far as I'm concerned, automation of tedious, manual tasks can't come fast enough.

What I haven't spent a ton of time thinking about – until relatively recently – is the combination of the two, i.e. artificial intelligence (AI). This is because when it was first introduced publicly several years ago, AI, specifically the sort of "generative AI" that comprises LLMs, was mostly a toy.[2] (We'll return to broader, non-LLM applications of AI later.)

[1] Technically it was "Physical Chemistry 2," but I think the whole semester was about Schrodinger's Wave Equation, and electrons quantum-tunneling sort of broke my brain in a way I never really recovered from. How do the electronics just vanish and reappear??? Anyway, the first day of class, the professor walks up to the whiteboard and says, matter-of-factly, "differential equations really should be a prerequisite, but I know some of you haven't had it, so we'll just spend half of today's class on a brief review of partial differential equations." When he treated a semester of math I had never taken as a trivial hour or two, that's when I knew I was in for it. I think I somehow squeaked out of that class with an A, or maybe a B+, but I shudder to think what the grade would've been without my TI-89, which did everything for me, years before AI was a thing.

[2] Please note that I am not an expert and I'm likely to use technical terms loosely; due to the nature of my work I am obviously more focused on business cases than technical details, except to the extent that technical details are relevant to the business case.

I think (though I'm not certain) that the first LLM I ever played with extensively Meta's BlenderBot. I remember the experience fairly vividly; I was at my mom's house on my laptop on the dining table, marveling at a chatbot and showing her what it could do. You could talk to it (although it was somewhat robotic); it could write a solid B of a high school essay. ChatGPT was similar; a friend of mine used it to name their dog. But it would hallucinate wildly and randomly, and act awkwardly; it was more of a novelty or a plaything than something that could do real work. More *"that's such a neat drawing, sweetie, I'll put it on the fridge,"* less *"hang it in the Louvre."*

Over the past several years, AI started to find some obvious use cases. The utility of LLMs for coding has been evident for quite some time (and has been something we've been following, as it relates to various portfolio companies). Coding is right in the sweet spot of an LLM because they excel at processing and creating language in the form of text, which is exactly what coding is. Python, or Javascript, or any other language aren't so different from, say, English or Spanish.

AI's applicability for other such use cases was clear; for example, a friend who does translation work, primarily from Japanese to English, has been keeping me abreast of the rapidly increasing capabilities of AI translation engines, complete with a live demonstration. At this point, he is doing less of the translating directly. He simply chooses the ideal contextual interpretation of the text, after which the engine re-translates the document appropriately. He estimates it has saved him 30-50% of translation time, depending on the project.

Finally, while I'm not sure I've encountered one myself, it is obvious that AI-based LLMs can and should disrupt the customer service industry. We've all waited on the phone to deal with a basic customer service issue; an LLM is a faster, more efficient, and more accurate way to solve the majority of issues, with human intervention best reserved for complex challenges. We'll return to this topic later.

Of course, while these are powerful use cases, they are also somewhat narrow; most of us don't work in translation or customer service, or in software development (despite the prevalence of code all around us).

I believe the slew of relatively recent releases of new models and features has changed the aperture of LLMs' applicability from hyper-zoom to wide-angle. It's what caused me to dive in, and it significantly changed how I think about AI. I believe the recent public model releases (particularly those that can search the web and integrate and synthesize information from multiple sources), as well as the functionality of other tools (both free and paid) that are becoming available, represent a step change in AI usefulness and adoption for knowledge work in general.

Here, I'm referring to public models such as Gemini 2.0 from Google, "Deep Research" from ChatGPT, Claude 3.7 Sonnet from Anthropic, and Grok 3 from X / Elon Musk / Twitter – which all provide real-world functionality. (I haven't used DeepSeek personally, for obvious reasons, but by reputation it falls in the same category.) I'm also referring to the AI tools increasingly integrated into many of the software platforms we use for our work.

Given the pace of advancement, with each model topping the last on some benchmark, it is likely that the coming months (let alone years) will see further progress, rendering AI increasingly useful and important for organizations and individuals alike. What does this mean for our firm, the business world, or us individually?

The purpose of my research – spanning from reading expert interviews and technical analyses, as well as spending a whole lot of time hands-on myself with pretty much every tool I can get my hands on – was coming to some sort of informed conclusion on three specific topics:

1) Can AI be additive to our research process?

2) What areas of business or economic activity are potentially at risk of disruption from AI?

3) What are potential investment opportunities related to AI?

Let's start with some technical underpinnings.

Intelligence, Artificial or Otherwise

The more I've read, thought about, and worked with AI, the more I dislike the term "artificial intelligence." The reason is simple: *artificial* is the opposite of *real*. Nobody wakes up in the morning and says "*I'd like to eat something artificial for breakfast.*" [3]

"Artificial" immediately calls to mind "real," so people try to think of AI using a framework of real intelligence, which can lead to some strange conclusions. There are certainly overlaps – tests like the AIME (an extremely challenging high school math competition)[4] and the LSAT (law school admissions) are used as benchmarks to demonstrate that LLMs are good at solving real-world problems. (Fun fact, this is an illusion. It turns out LLMs don't solve problems at all, nor do they think in any meaningful way – but they still take a different avenue to achieve useful results.) Many of these tests are proxies for IQ with a strong positive correlation (as you would expect.)

At the same time, very little of the work that most people do has anything to do with these benchmarks. Below is a sample problem from the 2022 AIME exam:

Let $w = \dfrac{\sqrt{3}+i}{2}$ and $z = \dfrac{-1+i\sqrt{3}}{2}$, where $i = \sqrt{-1}$. Find the number of ordered pairs (r, s) of positive integers not exceeding 100 that satisfy the equation $i \cdot w^r = z^s$.

I barely understand the question, let alone how to solve it, and I was a "mathlete" who took this test years ago! One could be great at answering that question, but it wouldn't necessarily help you train a dog – or sell a car – or identify which market segments your business should be targeting for expansion. As we'll get into, there is a difference between intelligence and insight.

So I thought that Satya Nadella's comments in a recent interview were quite interesting:

> *This is where we get a little bit ahead of ourselves with all this AGI hype. Remember the developed world, which is what? 2% growth and if you adjust for inflation it's zero?*
>
> *So in 2025, as we sit here, I'm not an economist, at least I look at it and say we have a real growth challenge. So, the first thing that we all have to do is, when we say this is like the Industrial Revolution, let's have that Industrial Revolution type of growth.*
>
> *That means to me, 10%, 7%, developed world, inflation-adjusted, growing at 5%. That's the real marker. It can't just be supply-side.*
>
> *In fact that's the thing, a lot of people are writing about it, and I'm glad they are, which is the big winners here are not going to be tech companies. The winners are going to be the broader industry that uses this commodity that, by the way, is abundant. Suddenly productivity goes up and the economy is growing at a faster rate. When that happens, we'll be fine as an industry.*
>
> *But that's to me the moment. Us self-claiming some AGI milestone, that's just nonsensical benchmark hacking to me. The real benchmark is: the world growing at 10%.*

There are some other important takeaways from this interview, which we'll return to later. While I don't necessarily agree with his exact barometer of 10% GDP growth as remotely likely, I'm using a similar framework for analysis. AI's success or failure is tied to what it can accomplish in the real world, not on a benchmark. So – just as an example – an AI like ChatGPT, which is "nerfed" by a lot of restrictions, is often less useful for certain purposes than X's "Grok," which Elon Musk described as "based" – it tries to prioritize reality.

OpenAI could come out with a 10x more powerful model, on some benchmark test, that is still less functional and useful for a given task. (Note that I've found myself using all major AIs, and neither Grok nor Gemini have

---

[3] Potheads halfway through a Dorito Crunch Wrap or whatever might disagree; I don't know. I can't speak for everybody. And I'm really quite fond of Coca-Cola, so I guess there's that.

[4] I won the regional MathCounts competition in eighth grade… the AIME, though. Woof. I qualified, but I was really terrible at it. I think I managed to get through two or three problems and just sort of sat there for three hours staring blankly at the rest.

anything that remotely compares with ChatGPT's DeepResearch mode.) We'll return to this idea in a few sections, and throughout the rest of the paper.

Defining AI: What is it, Anyway?

Defining what AI actually is can be a bit tricky, particularly since everyone and their dog now claim to be "AI." A lot of stuff that has been around for years (for example, machine learning), is being rebranded "AI," making it confusing to sort out.

So I turned to – you guessed it – AI to help. I asked ChatGPT 4.5, Grok 3, and Gemini 2.0 Flash to walk me through AI. All of them provided a similar answer to my initial query to define AI along the following lines:

> *"AI is the ability of a machine or computer system to perform tasks that typically require human intelligence. These tasks can include things like learning from experience, solving problems, understanding language, recognizing images, or making decisions."* – Grok3

Categories and Applications of AI

I then asked the AIs to provide a structured analysis of categories and applications, which Grok did the best at via DeepSearch, which led me to just abandon the others and keep going with Grok, which has become my favorite for "conversational" or general-learning purposes.[5] AI can be thought of as an umbrella, with various applications including but not limited to:

- natural language processing (of both text and voice),
- computer vision or image analysis (my wife's field),
- robotics (intelligent / autonomous physical interaction with the real world),
- expert systems (rules-based),
- domain-specific topics (such as protein folding)
- generative AI, which creates new content such as text, images, or music using models like LLMs or image generators. (Note that these build on natural language processing.)

A few points should emerge here. The first is that many of these technologies have existed in many ways for a long time. For example, I remember playing around with OCR ("optical character recognition") software as a kid – you'd scan a document, and the computer would "read" the image.

A second – that follows from the first – is that the likely reason there is so much focus on LLM models / generative AI and natural language processing capabilities is because they represent the largest and most accurate step function. If my Ring camera gets better at detecting motion, we won't notice that so much as *not notice* it – we'll get fewer irritating alerts because a leaf moved. Generative AI is obviously totally new, and we notice it. Meanwhile, while language processing has been around for a while, there is a clear change in the ability to actually interpret and assess human commands. I remember in the pre-Android Auto era, my car supposedly had voice recognition, but you had to be super specific. It had a very specific syntax you had to follow. You had to say "dial number," then wait for it to respond, then give it the number, which most of the time it would partially hear or mishear, and so on.

Similarly, Google search was an amazing and transformative technology – but in some senses was fairly dumb and mechanical. You couldn't simply type in "I want to learn about XYZ topic" and get a structured response with source links. (We will address hallucination and garbage-in, garbage-out

Finally, it is clear that many of these technologies work together synergistically. For example, a humanoid robot working in a factory would need to be able to interact with the world physically and autonomously, but it would first need to see the world (computer vision), and be able to understand instructions that it is given (natural language processing). It might also need an overlay of rules-based thinking (for example, accept commands from certain employees, but not from random passers-by.)

---

[5] Although in fairness, I didn't want to waste a valuable ChatGPT Deep Research query on this one thing, and I'm sure that it could've done the same thing – GPT Deep Research absolutely blows Grok3 out of the water. It's kind of weird; ChatGPT seems to either give you a paragraph or 30 pages, with nothing in between, whereas Grok does well at giving you the 3-5 page type analysis, but can't or won't go any deeper than that. Maybe I just suck at prompting.

Underlying Technologies

Several underlying technologies often cited as enablers of AI are the following:

- Machine Learning (ML): computers learning from data without explicit programming. ML is foundational for many AI applications, like recommendation systems on streaming platforms. ML can be further broken down into three categories:
    - Supervised learning, where it's given a cheat sheet – for example, being given pictures of dogs to learn what dogs look like.
    - Unsupervised learning, where it's just turned loose on raw data, and has to find patterns, structures, or relationships on its own – like handing a kid a jigsaw puzzle with no picture to reference.
    - Reinforcement learning – the machine learns by trial and error, interacting with an environment and getting rewards or penalties. Like training a dog with treats.
- Deep learning: part of ML, deep learning uses neural networks with multiple layers to process complex data, like images or speech. It's particularly effective for tasks like facial recognition, and the evidence leans toward it being a driving force behind recent AI advancement
- Neural networks: models inspired by the human brain, with interconnected nodes (neurons) inspired by the human brain, with interconnected nodes (neurons) that process data to recognize patterns. They're the backbone of deep learning, used in everything from voice assistants to self-driving cars. An unexpected detail is how these networks, with hundreds of layers, can sometimes outperform human experts in specific tasks.

These techniques have, also, been around for a long time; I remember a class project from B-school where we worked with a multinational company on forecasting, and neural nets were one of the options.

What Changed? Attention Is All You Need

So what changed? Google published a seminal paper in 2017 called "Attention Is All You Need," which proposed the "transformer," a new methodology for machine learning. If you've ever wondered what exactly ChatGPT stands for (I was only vaguely aware), it refers to its architecture: *Generative Pre-Trained Transformer.* As Google's AI overview explains:

> *In machine learning, a "transformer" is a type of neural network architecture that excels at processing sequential data like text by utilizing a mechanism called "self-attention" to understand the context and relationships between different elements within the sequence, allowing it to effectively capture long-range dependencies in the data, unlike traditional models that process information sequentially; this makes transformers particularly powerful for tasks like machine translation, text summarization, and question answering. Unlike recurrent neural networks (RNNs), transformers process the entire input sequence simultaneously, which significantly improves computational efficiency.*

To get slightly more technical, Grok3 offers the following:

> *Transformers consist of two main components: the encoder and the decoder, forming an encoder-decoder architecture. The encoder processes the input sequence, converting it into a set of vectors that represent the sequence's meaning, while the decoder generates the output sequence, one element at a time, using the encoder's output.*
>
> *For example, when processing the sentence "The animal didn't cross the street because it was too tired," self-attention helps the model associate "it" with "animal," understanding context across the sequence. This is a significant improvement over RNNs, which process words one at a time and can lose context over long sentences.*

There are several key insights here we will bookmark for later. The first is the increasing divergence between AI and human intelligence. If you read those paragraphs several times (which I found helpful; they're technical), you'll start to focus on the bit about not reading sequentially. Here is an analogy that Grok and I worked on together:

> *Humans read a book word by word, page by page, while transformers read a chapter (or a fixed chunk) all at once, considering every word's relationship to every other word in that chunk. This parallel processing lets transformers capture connections—like how a word on page one relates to a word on page ten—more efficiently than humans might, but only within the chunk they're processing.*

As I've been saying a lot… we'll come back to this later.

Inference and GPUs: Nvidia's Red Dead Redemption

Moving forward, we get to the term "inference" that is increasingly being thrown around. Citing IBM, Grok explains:

> *Inference in AI is the process of using a trained model to make predictions on new, unseen data, critical for real-world applications like chatbots or translation services. For transformer models, inference involves processing a new input sequence through the encoder and decoder to generate an output.*

Of course, inference requires some silicon horsepower. And unless you've been living under a rock, you know that Nvidia is the company providing that. But how and why? Let's go back in time.

Fun fact. Well before Askeladden was ever even conceptualized, I owned Nvidia in, like, 2012, at, like, $6 a share (apparently about $0.30 split-adjusted). And I probably sold it at $7 or $8 or something (I'm too lazy to go find the trade records.) This, too, we'll come back to later.

At the time, it was perceived as a dinosaur on its way to extinction; they mostly made graphics cards for PCs, and people thought that PC gaming was on its way out because of mobile games. Cartoon games like Farmville and Angry Birds were apparently going to spell the end of AAA games like GTA V or RDR 2.

It's worth understanding the difference between a CPU and a GPU. CPUs (the kind that power your computer or phone, and most servers), have a few powerful cores that are optimized for *sequential* processing, handling complex instructions one at a time. Conversely, GPUs have thousands of much smaller and less powerful cores, that work *in parallel* (not sequentially).

At the time, GPUs' ability to be used for other, non-graphical tasks was understood, but it was niche – tasks involving a lot of data-crunching, like high-performance computing applications like simulations or cryptography.

Grok does a good job of explaining the similarities between graphics workloads and transformer workloads:

Graphics Workloads:

- *Rendering a 3D scene involves millions of pixels, each requiring calculations (e.g., lighting, shading) based on geometry and textures. These calculations are independent—pixel A doesn't care about pixel B—so they can run in parallel across GPU cores.*
- *Math-heavy: Matrix multiplications and vector operations dominate, as GPUs transform 3D coordinates into 2D screens.*

Transformer Workloads:

- *Transformers process sequences (e.g., sentences) using self-attention, which involves comparing every word to every other word in a chunk. This creates large matrices (e.g., attention scores) that need multiplying and summing—also independent tasks that can run in parallel.*
- *Math-heavy: Like graphics, transformers rely on matrix operations, especially during training and inference, where billions of parameters adjust or predict.*

GPUs have thus been critical for enabling the rapid growth of transformer workloads. (We will, later, briefly address TPUs and other types of processors.)

"Models" you hear about (DeepSeek-R1, ChatGPT 4.5, Grok 3, as well as many proprietary / domain-specific models) are transformers (or other types of approaches) that are usually trained before they can become useful. The same way a doctor goes to medical school so they can learn to diagnose and treat patients, AI models start as a "blank slate" and are trained on vast troves of data to learn things. This can be very compute-intensive; Grok3 trained on 100,000+ Nvidia H100 GPUs for months.

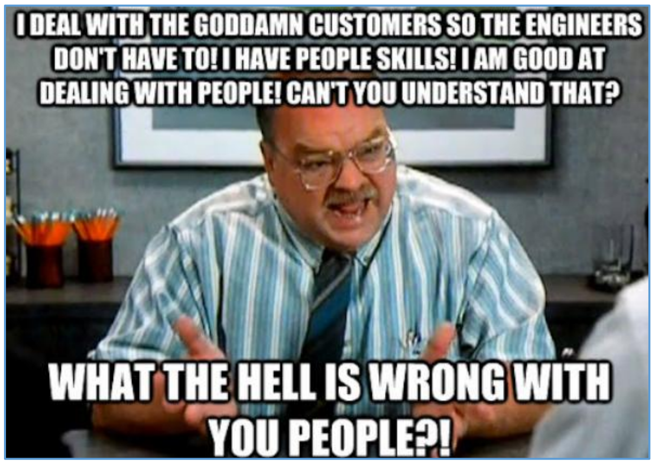
After the training phase, models are much less compute-intensive, but the better and more useful the models are, the more people will find applications for them (as we are currently doing, to be discussed in a later section.) Thus, even though individual use of any given query is small, it adds up quickly; Sam Altman recently tweeted:

> *we've been growing a lot and are out of GPUs. we will add tens of thousands of GPUs next week… (hundreds of thousands coming soon, and i'm pretty sure y'all will use every one we can rack up.)*

The fact that in 10 short years (2012 – 2022), Nvidia went from "dinosaur" to "the future" has important implications, which we will discuss in a later section.

For A Smart Person, You're Pretty Stupid

In *Office Space*, The Bobs interview Tom Smykowski for his job, in which he makes the stereotypical observation about engineers not having good people skills.[6] This is an amusing segue to my point, which is: intelligence comes in different forms.[7]



My wife recently watched *Young Sheldon*, a sitcom about the childhood experiences of future Nobel Prize winning scientist Sheldon Cooper, of *Big Bang Theory* fame. Sheldon is a certified genius with an IQ of 187 (as he often likes to remind people), but often fails to display basic social graces, understand interpersonal situations, or assess how other people are feeling. In this respect, he is drastically outperformed by his decidedly more cognitively average siblings Missy and Georgie.

Indeed, there's one specific episode where some researchers run tests on Sheldon and Missy – and Missy notices subtle social cues that lead her to suggest that the female researcher ask out her male colleague, which she does, successfully. Sheldon would never have gotten there in a million years.

AI, and particularly LLMs, have a number of well-documented flaws, most notably hallucination. But when we're thinking about their impact as it relates to their stated goal of replicating what is typically considered human intelligence, we should understand how they work (which is why I've actually spent so much time on it.)

We'll talk – extensively – about some of their flaws and limitations, but first we should consider why they occur.

---

[6] Both my parents were engineers, and everyone expected me to be a scientist. Like my dumb Roomba, I aimlessly wandered off in some other direction that nobody thought I should go in. 😊

[7] This itself is a somewhat controversial statement that is perhaps scientifically inaccurate, insofar as emotional intelligence is actually weakly correlated to IQ, and "g" (general intelligence, measured by IQ tests) is a fairly robust predictor of all sorts of outcomes in life, ranging from health to wealth. My understanding is that Howard Gardner's idea of "multiple intelligences" as taught in my college psychology class has been largely discredited. However, my statement is referring less to specific, precise scientific terminology, and more to our everyday understanding of intelligence, or at least the way we use it in common parlance – when I say there are different forms of intelligence, I mean someone can be "book smart" and get straight As in school but make terrible business decisions, or conversely be a mediocre student but intuitive at sales and identifying the next big thing, etc etc – like Sheldon's brother Georgie!

Grok3 On How Transformers Process Information Differently from Humans

Note that everything contained under this heading was written by Grok3.

Transformers and humans approach information processing in fundamentally different ways. Here's a breakdown:

Human Processing

- Sequential and Gradual: You read linearly, processing each word as it comes and building a mental model over time. Your understanding deepens as you go, but you might forget details from earlier pages or struggle to link distant concepts without effort.
- Context from Experience: You draw on real-world knowledge, emotions, and common sense to interpret what you read. For example, in the sentence "She didn't eat because she was full," you intuitively know "full" means "not hungry" based on life experience.
- Limited Attention: You focus on one part of the text at a time, and your attention can wander or tire, potentially missing subtle details.

Transformer Processing

- Parallel and Global: Transformers process a sequence (e.g., a sentence or paragraph) all at once using self-attention. This mechanism calculates how every word relates to every other word in the sequence, weighting their importance. For example, in "The dog didn't bark because it was asleep," self-attention links "it" to "dog," not "bark," regardless of distance.
- Pattern-Based, Not Experience-Based: Transformers lack real-world knowledge. They rely entirely on patterns learned from vast datasets during training. They don't "understand" in a human sense—they predict based on statistical associations.
- Positional Awareness: Since self-attention doesn't inherently respect word order, transformers use positional encodings (extra data about word positions) to keep track of sequence structure.

The key difference is that humans build meaning through time and lived experience, while transformers analyze everything in a sequence simultaneously, relying on data-driven patterns.

Insights: What Transformers See That Humans Might Miss

Transformers' unique processing leads to insights humans might not easily catch:

- Long-Range Patterns: Because transformers look at an entire sequence at once, they can spot relationships across large spans of text. For example, they might notice a recurring theme in a book—like subtle mentions of "freedom" across chapters—that you might overlook while reading page by page.
- Statistical Trends: Trained on massive datasets, transformers can detect language patterns or correlations humans might not consciously register, like rare word pairings that signal specific meanings.
- Task Efficiency: In tasks like translation or summarization, transformers excel by considering the whole input at once, producing coherent outputs faster than a human could.

For instance, if summarizing a novel, a transformer might instantly connect a character's early actions to their later fate, while a human might need to reread or reflect to make the same link.

Mistakes: Where Transformers Falter Compared to Humans

Transformers' reliance on patterns rather than understanding also leads to errors humans are less likely to make:

- Hallucinations: Transformers can generate plausible-sounding but false information. For example, asked about a book, they might invent a plot twist that never happened, based on patterns from other stories they've seen in training data.

- Context Missteps: Without real-world knowledge, transformers can misinterpret nuance. In "He hit the nail on the head," a human knows this is a metaphor for accuracy, but a transformer might take it literally if the data doesn't clarify.
- Sarcasm or Ambiguity: Humans pick up tone or intent from experience—like recognizing "Great, another rainy day" as sarcasm—but transformers might treat it as a genuine statement if sarcasm isn't well-represented in their training.
- Logical Gaps: They might produce coherent text with hidden inconsistencies, like a story where a character is in two places at once, because they prioritize fluency over reasoning.

Humans, by contrast, use common sense and critical thinking to avoid such errors, though we might miss subtle patterns or be swayed by biases (e.g., assuming a story fits a familiar trope).

Negation, Causality & Truth

It's me again (hi!) One thing that Grok doesn't say above – but that I think is a super critical insight – is that AIs, even when they are thinking or utilizing a "reasoning" mode, are not really reasoning. "Reasoning" modes in AI models like o1 Pro or Grok 3's "Think" mode are designed to mimic human-like problem solving by taking extra steps before giving an answer, such as using chain-of-thought or tree-of-thought (which explores multiple branching paths), and allows the AI to spot and fix mistakes mid-process.

However, the AI is still working very differently fundamentally from how we work. It has no understanding of causality or truth. As a result, AI struggles with some extremely basic things that a reasonably competent and mature human child could successfully accomplish. An example less frequently discussed than hallucination is negation. As any parent can tell you, toddlers have a very firm, intuitive, and profound grasp of the word "NO," and they don't hesitate to use it liberally.

Because AIs use pattern-based processing rather than semantic understanding, AIs do not understand negation. Indeed, when I was sourcing quotes for the "Revolutions that Weren't" section to follow, Grok3 – for all the praise I've heaped on it – gave me a quote from someone saying "XYZ will happen by 2020." I googled around and found out that all of this person's statements went, in fact, in the exact opposite direction. I've included Grok's full explanation of this negation issue in the appendix.

More specifically as it relates to causality, Grok explains:

- [AI] is not reasoning from first principles or understanding causality like humans—it's still pattern-driven. It leans on statistical likelihoods from training data (e.g., Grok 3's 200 million GPU-hours on text), not a mental model of the world.
- It doesn't "know" truth—it optimizes for what's likely to be correct based on rewards, not reality.

My own analogy is Poppy, the older of our two rescue Lab mixes. This has been a long read, and it gets even longer, so here is a funny picture of a doleful dog waiting for her dinner. Look at that face!

Poppy, like all Labs, is extraordinarily food-motivated. There is a certain mat in the family room that we ask her to sit on at mealtimes, while we're heating up her food in the kitchen. Poppy will, at other times when we are in the kitchen, hopefully sit on that mat, even if we are not feeding her and have no plans to feed her. She does not understand causality and is not using human reasoning; she just recognizes a pattern – she sits on that mat, she gets food. It is intelligence in its own way; indeed, AI training and rewards is not so different in some ways than the operant or classical conditioning that is used in animal training.

Intelligence, Human Or Otherwise

Now here's the really interesting part (to me). Grok's explanation above is both fantastic and true, and Grok was able to generate it *without actually thinking, understanding, or reasoning* in the sense that you or I do. This entire document is me building up from first principles how AI works, and fitting that into the context of how the world works. Grok was able to replicate part of that process using a different model (potentially because other real humans had already done so.)

As an analogy, imagine two different drivers racing to a destination through an urban environment. One has a slower car, but uses their knowledge of local streets to find some shortcuts. The other has a more powerful car, but has to take a longer route. This is of course not a perfect analogy, but it illustrates that two different approaches can lead to the same end result. It also demonstrates that the approaches can be complementary rather than competitive – do bigger engines negate the need for local knowledge? No, combining the two would result in a better lap time. As we consider the use cases of AI – its strengths and its flaws – I think that the focus on comparing it to humans is understandable but somewhat misguided. AI does things that we absolutely cannot.

Remember that not all AI tools are LLMs or GenAI, and thus not all AI hallucinates in the sense that GenAI does, and as a corollary, there are situations where the massive statistical processing is superior to humans. For example, AI seems to have massive promise in drug discovery. Our understanding of causality in medical conditions is often very limited; for example, drugs based on the "amyloid plaque" theory of Alzheimer's, once touted as potential blockbusters, have had relatively modest impact, and scientists are starting to realize that alternative mechanisms might be a bigger piece of the puzzle. AI doesn't care about causality and might be able to spot patterns we might not even think to look for – one expert noted that a somewhat creative, hallucinatory AI often has benefits in drug discovery (the same way that many founders have a bunch of bad ideas before or after hitting on the goldmine.)

Similarly, as we'll touch on later, data suggests that AI drivers such as Waymo / Tesla FSD are already likely substantially safer than the median human driver in most conditions, and it is likely that as both the software/training and the hardware (such as various car sensors and cameras) continue to improve in quality, that this gap will widen further. Unlike humans, AIs do not get drunk, or tired, or lost in thought, or distracted by something interesting on the other side of the road; they do not text and drive (but they will read your texts for you), they do not intentionally tailgate or make aggressive maneuvers, or drive 25 miles over the speed limit in a school zone, and they definitely don't get road rage. These things alone eliminate many of the major causes of accidents.

We'll get deeper into what humans (and AIs) do well, or not, in the following sections, including a full section on hallucinations. For now, the critical point I'm trying to make is that while people often directly compare AI and human performance, AI is working in a very different way than our brains are, even though there are certainly some similarities in how we learn and work. That means that it will be able to do things – amazing things, incredible things – that we could never dream of. An example comes from a recent news article about a Google AI solving a complex problem in microbiology related to antibiotic resistance; the AI discovered, in 48 hours, what had taken a team of advanced researchers more than a decade.

As such, I believe that learning to use and apply AI first requires leaving behind human norms – which is counterintuitive because the technology is often designed to appear and behave more human-like.

This also means that – at least without a fundamental change in architecture from the massive statistical-processing methodology currently used – it seems (as Satya Nadella seemed to imply) that we are not particularly close to an artificial general intelligence similar to Tony Stark's JARVIS, C-3PO from Star Wars, or Cortana from Halo. AI, as constructed, simply has insurmountable structural limitations.

As an analogy, think about combustion in the context of greenhouse gas emissions. The generic combustion equation is "some hydrocarbon plus oxygen equals carbon dioxide plus water," or: $C_xH_y + O_2 \rightarrow CO_2 + H_2O$.[8]

The process of combustion can certainly be made more *efficient*, such that less of the input hydrocarbon is "wasted" by being turned into heat, some other compound, or any other undesired output. But at the end of the day, even if we improve the process to 100% efficiency, there is a minimum output of carbon dioxide that is *physically required* for every molecule of hydrocarbon input. This is an immutable reality.

AI Testing

> *"Entrepreneurship is like sex. If you want to know what sex is like, you are not going to learn by reading about it or by talking with others. The best way to learn is just to go out there and do it… and the best part about it is that you are going to have a lot of fun trying."*
>
> – Warren Buffett

Over the past few weeks, I have ate, slept, and breathed AI. I already had a free subscription to Gemini, and I subscribed to Claude, ChatGPT, and Grok. I played around with some other free ones online. I had, of course, read about it theoretically, but as discussed, I had never used it as more than a toy. As a reminder (since this is a long document), there were three specific questions I sought to answer:

1) Can AI be additive to our research process?

2) What areas of business or economic activity are potentially at risk of disruption from AI?

3) What are potential investment opportunities related to AI?

I came up with two areas to test AI:

1) Investment Research (generalizable to research in general)

2) Coding
    a. Web scraper building
    b. Website building
    c. Basic video game coding

Here are my conclusions from those endeavors.

88 (For chemistry nerds, note that I've eliminated coefficients for simplicity.)

Investment Analysis: Rumors Of My Death Have Been Greatly Exaggerated

Over the last however many months, I've been seeing an increasing frenzy of internet predictions that knowledge workers will all be out of a job in six months, or three years, or ten (depending who you ask) because of AI. For many reasons (including the base rates "heroes or zeroes" discussion later), I was intrinsically skeptical of such claims. However, they merited evaluating myself.

The last time I had tried AI for anything work-related, I found that it lacked sufficient depth and nuance, so I gave up. Therefore, I was particularly intrigued by the launch ChatGPT's "Deep Research" mode, which purported to follow a human-esque research process.

Purely as a hypothetical, given a random topic (say, best practices in dog training,) it would scouring its training data, as well as searching the web for many sources. It might break the prompt down into sub-topics it identifies (different training methods, things that work for specific breeds or ages, training techniques that are best for specific behaviors such as reactivity, chewing, etc.)

It would then compile an organized, lengthy report on these topics, although it might first ask you for clarifications. It can also respond to more detailed prompts specifying topics to cover.

I was obviously interested because a large part of our research process comprises doing such research and compiling such documents, which is a very tedious and time-consuming endeavor. I have always looked for ways to speed this process up – and have found some – but largely, other than some minor modifications (such as having more access to certain paid subscriptions), or standardization/formalization of specified steps, our research process has remained largely similar over the past ~10 years (dating back to the pre-Askeladden days.)

The first and most obvious / naïve use case was asking ChatGPT to prepare a detailed research report on a few companies. I quickly decided to have it start with companies from our portfolio which I already know well, to establish a baseline for what it could do.

Phase 1: Portfolio Company Reports – Pareto In Reverse

The AI reports initially generated on our portfolio companies were aesthetically fairly impressive; they certainly *looked* like the real deal. As I read them, my response was mostly "okay." They occasionally got some details wrong, but they covered a lot of the story in a fairly adept way.

What they often seemed to miss were some of the really important / unique insights. While the reports were *long*, many of the sections felt a bit more like more-polished 10-Ks. For example, risk factors sections often failed to identify the real, critical risks to evaluate, or sometimes entirely missed risks that I know exist. Conversely, they would spend a lot of time on "boilerplate" risks – not exactly "climate change may hurt our business," but generic "slop" along those lines, that sounds good but really says nothing insightful when you think about it.

Phase 2: New Company Reports – A Lack of Insight and Nightmare of Hallucinations

Next, I turned to a new company I was researching, and had ChatGPT evaluate it. It did really well at introducing the industry, and the company's product lines (which were somewhat technical.) But at some point, it went off the rails. Perhaps the biggest – and best-known – flaw with LLMs is the problem of "hallucination," in which they basically conjure facts out of thin air for reasons that appear inscrutable to a human user.

I utilized the latest models that are supposed to be the least prone to hallucinations, although data suggests they still hallucinate very frequently. According to MIT Technology Review:

> *Tested on SimpleQA, a kind of general-knowledge quiz developed by OpenAI last year that includes questions on topics from science and technology to TV shows and video games, GPT-4.5 scores 62.5% compared with 38.6% for GPT-4o and 15% for o3-mini. What's more, OpenAI claims that GPT-4.5 responds with far fewer made-up answers (known as hallucinations). On the same test, GPT-4.5 made up answers 37.1% of the time, compared with 59.8% for GPT-4o and 80.3% for o3-mini.*

I also used carefully engineered prompts, in line with best practices, that encouraged the LLM to check its work and verify statements it makes using multiple sources. Despite this, it routinely presented information as fact that it could not source, and worse, made things up.

A list of hallucinations includes, but is not limited to:

- Claiming a certain executive at company A was previously CEO at company B, which company A acquired (in fact, the executive had never worked for company B)
- Claimed that a number from the 10-K was $10.5, when in fact it was $20.3 (imagine thinking a company had a $10 million liability when it was in fact $20 million!)
- Claimed a co-founder who passed away over 7 years ago was still the largest individual shareholder
- Claimed that an executive at this company was the former CEO of AstroNova, which I knew was not true because we know the CEO of AstroNova, who has been in his position for over a decade! (The executive in question had previously worked at another company that had "Astro" in the name.)
- Claimed an independent shareholder was on the Board when they were, in fact, not
- Claimed there have been "no recent secondary offerings" – one happened less than 12 months ago
- Seemed to not be anchored in time (because, again, it is not really thinking, it is correlating.) In some sections it presented very dated (say, 2022) information as if it was occurring in the present.
- Said Company A purchased Company B, when Company A purchased a small division from Company B

Outside of deep research reports, AI responses included:

- Grok3 told me that a company with $1 billion in volume had "over 10%" share of a $50+ billion market (?!)
- ChatGPT fabricated a quote and then later clarified that it was a paraphrase, but the material it was paraphrasing was somewhat related but had absolutely no support for the given quote. (Imagine someone saying "Strawberries are my favorite fruit," and ChatGPT decided to quote that person as saying "I want some watermelon.")

Other than the occasional typo or missed copy and paste, even the least competent analysts in the industry would not make such mistakes. Taking it a step further, a reasonably mature 10-year-old would likely not be able to create a research report comparably thorough to the output of an LLM, but at least you would have confidence that whatever is presented in the report is at least based on some source of fact.

These are major issues, and initially seemed insurmountable challenges. In investment analysis (and likely most other forms of research), reality is sort of opposite of the Pareto principle. The first 80% is fine, but the last 20% really gets you there. Being 80% good enough – which is kind of what the AI was – is simply not good enough.

Here is what ChatGPT itself has to say about the topic. (Yes, yes, the irony is not lost on me.)

Large Language Models (LLMs) sometimes generate information that appears plausible but is untrue or unsupported—a phenomenon known as "hallucination." This occurs due to several inherent characteristics of how these models are designed and trained:

1. ***Predictive Nature**: LLMs are designed to predict the next word in a sequence based on patterns learned from vast datasets. This process involves educated guesswork, and in the absence of specific information, the model may generate content that seems coherent but lacks factual grounding. This tendency arises because the model aims to provide a response, even if it means fabricating details to fill in gaps.*
2. ***Training Data Limitations**: The quality and scope of the data used to train LLMs significantly influence their outputs. If the training data contains inaccuracies, biases, or is incomplete, the model may produce responses that reflect these issues, leading to hallucinations. Additionally, when faced with ambiguous or imprecise prompts, LLMs might generate information to fill in the blanks, resulting in unintentional fabrications.*
3. ***Lack of Real-World Understanding**: While LLMs excel at recognizing and replicating patterns in text, they do not possess true comprehension or awareness of real-world contexts. This absence of grounding means they might generate content that is syntactically correct and contextually relevant but factually incorrect or nonsensical.*

Phase 3: Heavy-Duty Prompt Engineering

At this point, I had almost given up – I was spending as much or more time verifying the truth (or falsehood) of collected information as I would have simply starting from scratch myself. However, I was familiar with the idea of "prompt engineering" – basically, finding ways to get the AI to do more of what you want, and less of what we don't want.

Integrating existing IP that we had, things we read online (a special hat tip to @buccocapital for a few very useful threads), and AI itself, I went through multiple iterations to create a template prompt for producing research. This vastly improved the performance of the LLM, although hallucinations and other challenges remain.

I say this about LLMs specifically and not AI tools as a whole (since there are other, non-LLM types of AI optimized for other purposes that seem to hallucinate substantially less): asking a generative AI to not hallucinate is sort of like asking your dog to guard its steak. You can do it, and if you keep a close eye on it you may escape with your dinner intact, but you're playing a dangerous game. I have tried all the best practices to reduce hallucination, as well as the new models designed to reduce hallucination, and they still hallucinate – somewhat less frequently, but often in ways that are critically important. This doesn't make them not-useful, but it does mean that their use needs to be constrained to situations where their output is acceptable, and for any important purpose, their output needs to be verified by human research. They are an accelerator rather than a replacement for research and human effort.

Phase 4: Reflection and Integration

Somewhat swayed by internet hype that AIs could replace junior analysts or junior devs, I started with the premise that I could simply ask the AI to do something and it would do it the way I would do it, or at least a somewhat less experienced version of me would do it. Clearly, that's not the case. But as discussed previously, I realized after reflection that expecting the AI to be human is not the right approach. Just like a person has strengths and weaknesses, so too does the AI; rather than blaming the AI for doing things it was – in a sense – programmed to do, it's my job to use it for the things it's good at. Deep Research is an incredible, game-changing tool. But as the saying goes, to a hammer, the world looks like a nail. It's not the right tool for every purpose, but that doesn't make it a bad tool.

At a high level, LLMs are bad – maybe even terrible – at insight / "wisdom" / judgment, and they are also terrible at distinguishing fact from fiction. As previously discussed, they do not think. They tell you that they think, and they do a lot of things that *look and sound like* thinking, but we should not anthropomorphize them. They are not really thinking in any way that we should consider meaningful. Even when it looks like they are solving a math problem (see appendix), they are doing statistical associations. This can, however, be really powerful and can lead to some unexpected insights that are sometimes the same and sometimes different than ours; think of it as evolutionary convergence.

I'm not discussing specific research process modifications in this document for two reasons. The first is that they are still very much in process. AI has not *yet* improved our productivity on the whole, and in fact may well have reduced it, because we've spent a lot of time playing with / testing it, learning about it, and learning how to use it – at this stage, it's almost like learning to type on a computer for the first time.

To be clear, however, when I've used it for specific, tailored tasks where I've verified it works well and developed a process, it does amazing work and definitely results in massive speed improvements (which is where I'm getting my 2-3x productivity boost estimate from). Just as computers changed the way we do things, so too should AI, and I am re-engineering multiple parts of my process.

The second reason is that any workflow we create, I see as valuable intellectual property that I'd prefer not to disclose to everyone else in the industry, since I believe it provides us with a substantial competitive advantage, at least for a little while until everyone else figures it out. I am nonetheless planning to update clients specifically on this topic when I have something specific to share; I am still in the planning and testing stage.

I will share some thoughts at a high level. The first is that independent of AI, what matters as an analyst is a function of time, purpose, and perspective. If you are trying to get to know a company, a qualitative summary written from an investor's point of view that captures some of the highlights – even one that may contain some inaccuracies – can be super helpful. If you're preparing for a first-time meeting with a company, it's OK to have some misconceptions or not know everything – conversely, if you're preparing to make a big investment, you better make damn sure. For things like actual financials, you can dig in on your own using data sources grounded in truth.

As an example, there are undoubtedly many technical details related to machine learning, transformers, and so on that I do not understand. I haven't read the original Attention paper on transformers. But it seems unlikely, at this point, that doing so would really help me understand AI's business case relative to simply using it myself and keeping up with what experts are saying in real-time about how it is being applied in other fields.

When it comes to the research process, there are situations (such as when one is actually making an investment) when accuracy and deep understanding are paramount. There are other situations (such as early in the research process) when breadth and speed are in fact far more important than accuracy. One challenge every investor faces is screening; there are thousands of public companies; how do you decide which to work on? Reading company filings or conference calls is a time-consuming endeavor, and as we've discussed, often results in a lot of nothing. At that point, details and accuracy are less important than getting a general sense if it fits your playbook or not.

Another example is keeping up with news and developments that affect a diversified portfolio (such as tariffs). Manually doing this is quite time-consuming and prone to missing things; AI can help here.

Similarly, for applications in which breadth is more important than depth, there are (sometimes public, sometimes proprietary) data-extraction tools that can automate or vastly speed up the process of getting up to speed – without significant hallucination. And AI can often "automate" google searches, particularly for new industries you're not familiar with, by providing you with source materials that would have taken a lot more time to find on your own, even if you throw out the AI's "opinion."

Without any existing investment knowledge or verification procedures, relying on AI output would be a disaster. But with a human framework of context, judgment, understanding, truth, causality, and reasoning… AI is an extremely powerful force multiplier. I've been bouncing it off other friends who I respect, and one suggested that he thinks his research process will take 50% less time using AI tools – similar to my 2-3x estimate provided earlier. (He also agreed that it was far from replacing human judgment.)

Coding Adventures

Time for a new domain. Other than a few lines in a robotics club and one time that I played around with some C# Microsoft trainer app or something, I've never learned to code. I just never really… wanted to. idk, it seemed complicated and boring (no offense to those who code, as my wife does.)

But there are obvious use cases. I found three specific ones. I'll just give the general overview of the three situations, and then offer my conclusions (which are the same across all three) at the end.

Web Scraper

For a while, I'd been meaning to build a web scraper. There are often public data sources that are useful to track; for example, during COVID, the TSA published a summary of daily air traffic trends. I am currently tracking a different public data set that relates to one of our investments which has a quirk; it publishes daily data but not historical data, only monthly summaries, such that if you want a complete and granular historical dataset, you have to copy the data yourself every day.

So I set out – using ChatGPT – to build a web scraper that would do this for me, automatically, using Python. It was quite the experience. The initial code it gave me was so broken that it broke Google Cloud (not all of Google Cloud, of course, just the function I was trying to run). It was so bad, I couldn't even get back into my function. Hours of debugging later, I finally got a working cloud function.

Now, meanwhile, the dataset I was trying to scrape was in Tableau, and ChatGPT either could not – or would not – identify the titles of the columns and rows I wanted to scrape. It gave me many complicated, frustrating instructions, which I attempted to follow without success (another couple hours later.) Finally, I got so frustrated that I – in complete naivete – suggested building a separate, locally-run program that would do this for me, and to my surprise, it agreed that would work, and after some more work and debugging, we finally sort of got it. (If it was intelligent, it would have suggested doing this in the first place, at some point during the seemingly dozens of rounds of attempting to fix the code.)

After a while longer (initially the function 'worked' but just pulled blank data, and then it pulled data for the wrong row), I finally got it working. Sort of. It is still pulling data from the wrong column, but some of the data is somewhat duplicative (i.e., one column is a product of other columns.) So I gave up on coding and just used algebra.

Based on what I've read, this is the sort of task at which the AI *should have* excelled. Each individual component of the code was not that complicated and is something that's been done a thousand times, so it should have a lot of training data on it.

Website

Friends and clients know (because I have been asked, repeatedly) that my website had been down for a long time. It is a miracle that I managed to put it together in the first place (the Wordpress side of things, not the content side). So I thought I would ask ChatGPT for help in nuking it and getting it back up (at least a basic version of it, anyway).

This was also… shockingly bad? I ran into a snag with Wordpress installation that was so severe that ChatGPT could not fix it I had to contact my host's support.

Once a clean Wordpress install was up, all I wanted to do was take a well-known, frequently-used theme, modify the header image, change some colors, and so on. This took hours. For at least an hour, there was a bizarre gray bar across my screen that could not be fixed despite repeated attempts. Finally, it somehow went away, but I had all sorts of bizarre issues with image scaling, and fixing that would break some other part of the code.

Another few hours later, I had a generic, single-page Wordpress website that was functional – with one page. We'll see what happens when I try to create a *second* page, and create a menu.

Basic Video Game

When Claude 3.7 Sonnet was released, everyone was raving about its coding abilities. So I signed up and gave it a shot. I wanted to create a simple side-scrolling platform game I could play locally on my computer, similar to *Super Mario*, as a surprise for my wife, with custom characters related to our family.

Claude first proposed an implementation based on "phaser.js." The problems were endless – the game wouldn't launch, then it would but get stuck, then when we finally cleared that, my characters were too big (or small), they didn't move, enemies flew across the screen at hypersonic speed, characters overlapped, things were invisible, and on and on and on.

In frustration, I switched to Grok3, then ChatGPT, then back to Grok, in the process also changing the platform several times (to another one I forget, and then eventually off javascript entirely to python, which created the complication of converting SVGs to PNGs, which AI was also very unhelpful with, until I came up with my own solution.) This is a very simple video game that – other than my custom characters – has been made thousands of times before (dating back to the 1980s), and is probably the sort of toy game that one would learn to build on like day 2 if you were a budding game developer, similar to learning *Donde esta el bano?* when you're learning Spanish.

Probably 10-12+ hours in, with multiple approaches, and Grok's "Think" mode, and o1 pro, and yada yada – no dice. I still don't have a functional video game. I even scrapped everything and tried to build a new version with only the most basic mechanics, with one level, one character, and one enemy, and that game is still not working. The platforms are in the wrong place, the enemy appears to be standing on top of my head… yeah.

Coding Conclusions

Despite my complete lack of coding knowledge, AI did allow me to code a semi-functioning web scraper, a functioning website, and a game with… at least a few functional elements, though it leaves much to be desired. That's pretty impressive, considering that without AI, I would have been able to do none of that without significant study. In other words, AI got me from zero to somewhere faster than I could have myself. Extrapolating from this and my experiences in a domain I know much better (investment research), I think it's completely reasonable to assume substantial productivity improvements for people who actually know how to code, who are using it to solve more limited pain points or creating a structure for further work rather than compiling entire programs.

Conversely, the fact that AI cannot create – or debug – extremely basic, very common issues suggests that it is far away from being allowed to run around unsupervised in enterprise-grade code. Similarly, the idea that it will allow anyone, anywhere, to code anything is flat-out ridiculous at this point. A professional would have done a better job without AI more quickly than me with AI. I was better at solving certain coding problems than the AI, not because I know anything about coding, but because I know something in general about thinking and problem solving.

Self-Help and the Question of Exponential Growth

If you were to present my conclusions to AI maximalists, I believe their responses would generally go one of two ways. The first I will dismiss out of hand (although, harking back to what I said about holding conclusions lightly, I will reevaluate if I see new data.) That view goes like: *"AI will get smarter and be able to fix the problems with AI."*

I think this is *extremely unlikely* to happen anytime soon, because as I mentioned, AI does not think. It just looks like it thinks. And that is currently a structural physical limitation that no amount of coding ingenuity can surpass. Two kids in a Halloween horse costume may vaguely *look like* a horse, but they are not a horse; they do not gallop. You can demonstrate this by asking ChatGPT to come up with novel jokes. It gets the setup-punchline pattern right – they *sound like* jokes – but only a few are funny, because ChatGPT *does not really understand humor.*

Particularly given the step-function improvement in recent model releases, the second view is something I think is more worth considering. It goes something like: *"AIs will improve exponentially, and focusing on current flaws or limitations misses how much better they will be very quickly."*

This is an argument worth considering, but the problem is that there are already signs we have reached a point of diminishing returns.

One of these is a "physical reality" issue. If you recall the training section above, frontier models are already trained on almost all the high quality information that exists.

In the now-viral "The Short Case for Nvidia" published by Jeffrey Emanuel, he observes:

> *For one thing, we seem to have already exhausted the world's accumulated set of high quality training data. Of course, that's not literally true— there are still so many old books and periodicals that haven't yet been properly digitized, and even if they have, are not properly licensed for use as training data.*
>
> *The problem is that, even if you give credit for all that stuff— say the sum total of "professionally" produced English language written content from the year 1500 to, say, the year 2000, it's not such a tremendous amount in percentage terms when you're talking about a training corpus of nearly 15 trillion tokens, which is the scale of current frontier models.*
>
> *For a quick reality check of those numbers: Google Books has digitized around 40mm books so far; if a typical book has 50k to 100k words, or 65k to 130k tokens, then that's between 2.6T and 5.2T tokens just from books, though surely a large chunk of that is already included in the training corpora used by the big labs, whether it's strictly legal or not. And there are lots of academic papers, with the arXiv website alone having over 2mm papers.*
>
> *And the Library of Congress has over 3 billion digitized newspaper pages. Taken together, that could be as much as 7T tokens in total, but since much of this is in fact included in training corpora, the remaining "incremental" training data probably isn't all that significant in the grand scheme of things.*
>
> *Of course, there are other ways to gather more training data. You could automatically transcribe every single YouTube video for example, and use that text. And while that might be helpful on the margin, it's certainly of much lower quality than, say, a highly respected textbook on Organic Chemistry as a source of useful knowledge about the world.*
>
> *So we've always had a looming "data wall" when it comes to the original scaling law; although we know we can keep shoveling more and more capex into GPUs and building more and more data centers, it's a lot harder to mass produce useful new human knowledge which is correct and incremental to what is already out there.*
>
> *Now, one intriguing response to this has been the rise of "synthetic data," which is text that is itself the output of an LLM. And while this seems almost nonsensical that it would work to "get high on your own supply" as a way of improving model quality, it actually seems to work very well in practice, at least in the domain of math, logic, and computer programming.*
>
> *[…]*
>
> *So while there is some hope in terms of being able to capture more and more additional training data, if you look at the rate at which training corpora have grown in recent years, it quickly becomes obvious that we are close to hitting a wall in terms of data availability for "generally useful" knowledge that can get us closer to the ultimate goal of getting artificial super-intelligence which is 10x smarter than John von Neumann and is an absolute world-class expert on every specialty known to man.*

This is an interesting sort of top-down "TAM" analysis that matches what we're seeing from the bottom up. A piece from TechCrunch observes, regarding OpenAI's recent 4.5 model release:

> *GPT-4.5 is OpenAI's largest model to date, trained using more computing power and data than any of the company's previous releases.*
>
> *[…]*
>
> *In every GPT generation before GPT-4.5, scaling up led to massive jumps in performance across domains, including mathematics, writing, and coding. Indeed, OpenAI says that GPT-4.5's increased size has given it "a deeper world knowledge" and "higher emotional intelligence." However, there are signs that the gains from scaling up data and computing are beginning to level off. On several AI benchmarks, GPT-4.5 falls short of newer AI "reasoning" models from Chinese AI company DeepSeek, Anthropic, and OpenAI itself.*

My guess is that in the near future, models stop improving rapidly in terms of raw horsepower, and incremental benefits come from improving their applicability (i.e. features such as deep research, based on existing models), and creating domain-specific AI tools trained on high-quality or proprietary data.

Like anything else, AI suffers from a "garbage in, garbage out" phenomenon – ChatGPT can probably talk intelligently about astrology (I haven't tested it out), but nothing about astrology is intelligent in the first place. There is a lot of valuable data that public models don't have access to.

So to get back to the Buffett quote we started with: was my journey of exploration fun? Hell yeah, it was. The novelty was cool, and I haven't been this excited about investment research in a long time. A friend of mine not prone to fad-following echoed this sentiment; playing with AI and seeing what it can do – as well as the pitfalls it might lead other people to, creating opportunities for people using it thoughtfully – made him *"as excited as he's been in a long time."*

How Will The AI Affect Jobs and Businesses?

With an understanding of what AI is – and isn't – and how it works in the real world, from both a theoretical and hands-on perspective, we can start to move towards investment conclusions.

As John Lewis Gaddis puts it in the excellent On Grand Strategy, *"Strategy requires a sense of the whole that reveals the significance of respective parts."*

Zooming out, as Gaddis or the world leaders he profiles would do, I think it's important to start by trying to compare AI to things that have happened before in history.

Base Rates: Heroes or Zeroes

Before evaluating any claim, it's helpful to have some understanding of the applicable "base rates," or statistical likelihoods that apply in a given situation. Meshing the "inside view" – what you know about a given situation – with the "outside view," i.e. what you know about how the world works in general – tends to result in better decision-making.

Today, people are making all sorts of hyperbolic claims about AI; for example, that it will result in *en masse* unemployment of knowledge workers within a few years, or that we're on the precipice of artificial general intelligence (AGI) that will achieve "takeoff," and so on.

Thoroughly assessing all of these claims in depth, beyond my current level of understanding, would be both beyond the scope of this letter and beyond the scope of my capabilities as an analyst. Fortunately, running a diversified portfolio and evaluating the world probabilistically means that we don't need precise answers to all of these questions; we just need to ensure that we're likely well-positioned regardless of which scenario occurs, and able to react nimbly to new developments.

It's worth noting that at any given time, something is being hyped as the next disruptive technology. Let's briefly review a non-exhaustive list of some of the ones I remember from my investing career (so I'm restricting the time period to 2010 onwards)… starting with a three fun "zeroes." [9]

[99] Note that I received assistance in sourcing these quotes using – what else – AI! Grok3 and ChatGPT 4o, to be specific. While I asked for citations, I couldn't always follow them back – surprisingly, finding quotes I could link back to was a very challenging task for the LLMs. However, even if I couldn't verify the direct quote, I did verify that the attributed statement at least matched the professed beliefs of the individual.

The Revolutions That Weren't: All Hat, No Cattle

1. 3D printing / additive manufacturing (~2012 - ~2015).
   a. The hype:
   "3D printing is going to be bigger than the internet." – Bre Pettis, CEO of MakerBot
   "Within a decade, we'll be printing entire buildings." – Behrokh Khoshnevis, USC Professor
   "Our printer… is set to revolutionize manufacturing as we know it." – Jeff Blank, VP Engineering at 3D Systems

   b. The reality:
   Companies such as 3D Systems (DDD) and Stratasys (SSYS) have stock prices down ~95% from their January 2014 peaks; their revenues, meanwhile, are also lower than they were in 2015. Although additive manufacturing has certainly found some applications such as prototyping, it has broadly failed to replace pre-existing manufacturing techniques such as machining, casting, injection molding, and others; any benefits have largely accrued to industrial incumbents rather than new disruptors. In other words, there is no evidence of any meaningful "disruption."

2. The Metaverse (~2021 – 2022)
   a. The hype:
   "immersive digital worlds [will] become the primary way that we live our lives" – Mark Zuckerberg
   "Metaverse… is the next chapter of the internet." -- Mark Zuckerberg
   "The metaverse is here, and it's not only transforming how we see the world but how we participate in it – from the factory floor to the meeting room." – Satya Nadella

   b. The reality:
   OK, in fairness, Mark Zuckerberg might have been the main person hyping this. But he's a billionaire in charge of one of the world's most powerful companies, so he was uniquely poised to put money behind his pet project. The idea that we would all interact via virtual avatars has, at least so far, proved mostly to be a fever dream; other than bad video games with crappy, '90s-era graphics, the Metaverse does not really exist in any meaningful way. Certainly, there is potential for AR/VR technologies to be used more over time, but I'm not sure anyone still seriously believes that this technology will have a widespread, disruptive effect soon. Indeed, after going to the trouble of renaming itself "Meta," Meta itself pivoted away from the Metaverse towards AI, slashing the budget for its Reality Labs subsidiary.

3. Blockchain, cryptocurrencies, and NFTs (2017 - 2022?)
   a. The hype:
   "The Blockchain Will Do to the Financial System What the Internet Did to Media." - Harvard Business Review
   "Cryptocurrency is such a powerful concept that it can almost overturn governments." – Charles Lee (creator of Litecoin)
   "Blockchain technology is likely to be a key source of future financial market innovation." – Chicago Fed

   b. The reality:
   Certainly many people have gotten rich owning Bitcoin, or random altcoins, that someone else is later willing to purchase for a higher price. But after more than a decade, there are only extremely limited and niche real-world use cases (other than criminal activity) for transacting in cryptocurrencies. For the majority of people, there is no real-world use case where Bitcoin is better than existing payment methods, like ACH or a credit card. As the saying goes, they are trading sardines, not eating sardines. And NFTs turned out to just be monkey JPEGs after all.

The Sizzle and the Steak: Remaking Reality

Now let's look at three technological revolutions that did pan out.

- Smartphones / "mobile" – by the mid-2010s, mobile devices accounted for a majority of internet traffic, and continues to increase in importance; particularly for younger generations, the smartphone is their primary conduit into all aspects of their world, from entertainment to learning to friendship. Interestingly, over 250 million PCs were still sold last year, compared to ~350 million in 2012, highlighting that while phones are the primary technology for most personal use, computers retain an important role – particularly for real work.

- E-commerce – The St. Louis Fed estimates that e-commerce sales have grown to over 16% of all retail sales as of Q4 2024, the same quarter in which Amazon recently surpassed Walmart in revenue for the first time. However, this relatively low figure belies e-commerce's impact; large categories such as auto parts and food and beverage have substantially lower penetration (for obvious reasons), while e-commerce has completely reshaped most specialty retailers and spelled the end of many shopping malls. Other than a COVID bump, growth in e-commerce has been relatively steady, rather than taking over all at once.

- The cloud – while estimates vary, it seems that cloud (including both public hyperscaler and private cloud) comprises 60-70% of market share today, with perhaps 30-40% on-prem, and this is likely to continue to evolve in favor of the cloud. As with the above cases, this penetration has built slowly over time.

COVID: A Special Case

While COVID did dramatically disrupt the world for a few years, and certainly accelerated the adoption of certain technologies (Zoom, e-commerce) as well as certain cultural norms (remote work, acceptability of wearing of masks in public), there was a point at which people prophesized that the world would permanently or at least semi-permanently shift to a more health-conscious, virus-avoiding tenor, having dramatic impacts on travel, group gatherings, and so on. Even things like remote work are currently swinging back in the other direction. But if you had been in a coma from 2019 – 2024 and walked around today, you would largely not know COVID had occurred.

The Takeaway: The Future Is Here… It's Just Not Evenly Distributed

One key takeaway from the above is that even when a technology is truly disruptive, it tends not to change the world all at once, nor does it usually result in the complete replacement of whatever came before. Or, in other words: culture adapts more slowly than technology. While some of the speed of adoption can certainly be linked to features or maturity, there are other cases in which it is purely cultural.

Personal story. I previously referenced my fondness for Alexa light switches or groups. If you know the right commands, you can – standing anywhere in my home – turn on, or off, pretty much any important light switch in the house. Alternatively, you can use the app to schedule lights (such as exterior lighting, or a nightlight) to turn off or on at specified times, or use slightly more advanced switches that are motion-activated.

Due to their networked nature, you can create all sorts of fun synthetic groups and cascading routines; as an example, you could announce "Alexa, I'm home," and have certain lights turn on (such as the lights in the area you're most likely to go to), have certain lights turn off (such as the garage light behind you, perhaps), and have Alexa do all sorts of other things (play music, give you a weather update, whatever.) I presume if you spent enough money, you could buy a toaster that would start making you toast, although – for now, anyway – you would still have to physically put bread in the toaster.

Jokes aside, once you get used to it, it's nearly impossible to go back. At my parents' house – or at AirBnBs, or random other places – I find myself instinctively and automatically opening my mouth to ask Alexa to turn on a light, before I realize that I'm back in the Stone Ages and I have to manually flip a switch.

This technology not only exists, but is relatively easy and cheap – the most basic smart switches are ~$20 apiece, and installing them requires no particularly specialized knowledge as long as you're comfortable turning off a breaker and connecting a few wires with a wire nut. Doing all of the important light switches and lamps in the average home might take a few weekends and would cost far less than $1,000 – relatively small dollars for a home improvement, particularly given the vast convenience impact on day to day life. Most people who get to play with it end up loving it; my mother keeps hinting I should find the time to install some for her… I promise I'll get around to it, Mom!

Yet I've never personally been in another house that works like mine, or even comes close; in fact, I have never actually noticed a smart switch in anyone else's home (although I have noticed a few smart bulbs.) I would presume that at some point in the future, smart lighting becomes the default – i.e. manually operating lights will be seen as quaint. Today, however, manual switches and bulbs that work the way they did decades ago are still the default, despite vastly superior technology being relatively inexpensive and widely available.

This is an example of a quote attributed to American author William Gibson: *"The future is already here; it's just not evenly distributed."* In previous letters, in other contexts, we've previously referenced that many important government and commercial enterprise functions operate on 30-plus year old programming languages such as COBOL; in other words, while we're used to slick, Netflix-esque UIs, lots of important technology we interact with every day (probably unknowingly) is sorely in need of an upgrade.

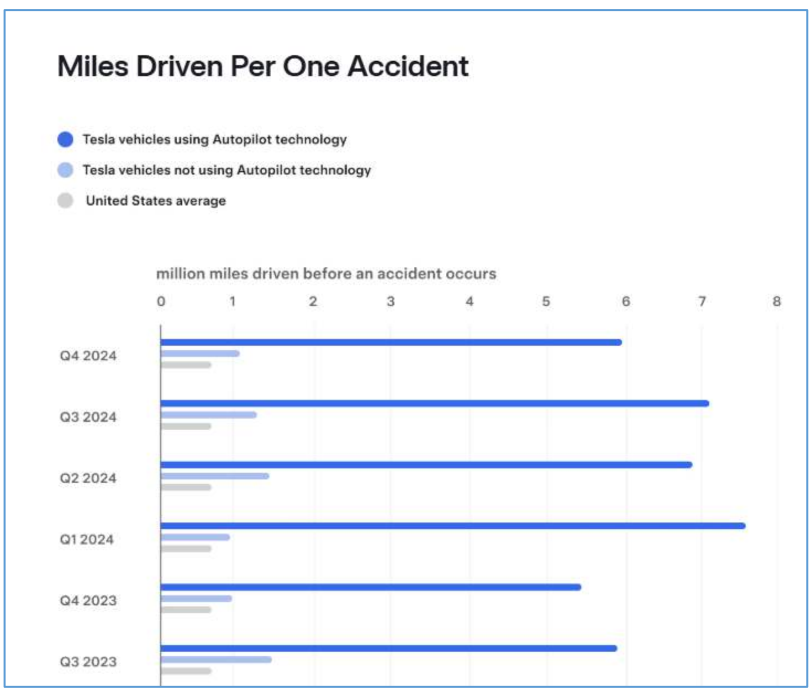
Similarly, for knowledge work, remote work has existed to some extent for a long time and the technology to make it relatively practical on a wide scale existed by 2010, if not significantly earlier. But it was fairly uncommon pre-COVID, and culture has been slow to change – even extremely smart luminaries like Jamie Dimon are clinging to an obviously antiquated, outmoded, and extremely costly cultural tradition (both directly and in terms of externalities).[10]

It's extremely ironic that at the same time that we are – culturally – worried about workers being so expendable that AI could replace them, we also value "in-person collaboration" so highly that these supposedly disposable workers all need to be in the same place at the same time. Confused? I am, too. There are better hybrid approaches that combine the best of in-office collaboration and working from home, but many companies seem slow to adopt them.

Whatever technological capabilities AI has, it will run head-first into cultural challenges in many industries, whether these are rational or not.

I'll provide two concrete examples. First is autonomous cars. In addition to the challenges associated with humans not paying attention all the time, it's worth noting that much of driving is pattern recognition. We can't really reason causally about what other people are or are not doing. We can only interpret patterns based on data. If a vehicle is weaving in and out of its lane, we don't know if the author is texting, drunk, tired, or the car's in poor mechanical condition. But does the causality matter? Whatever the cause, we should try to keep our distance from that vehicle. Intuitively, the massive parallel-processing capability of an AI is likely to vastly outstrip human drivers, particularly given that they can integrate data we cannot – sensors and cameras pointing in directions our eyes cannot see, potentially other inputs such as LIDAR, and so on.

Data from Tesla demonstrates that vehicles using its semi-autonomous driving features get into accidents at a much lower rate than its own vehicles not using those features.

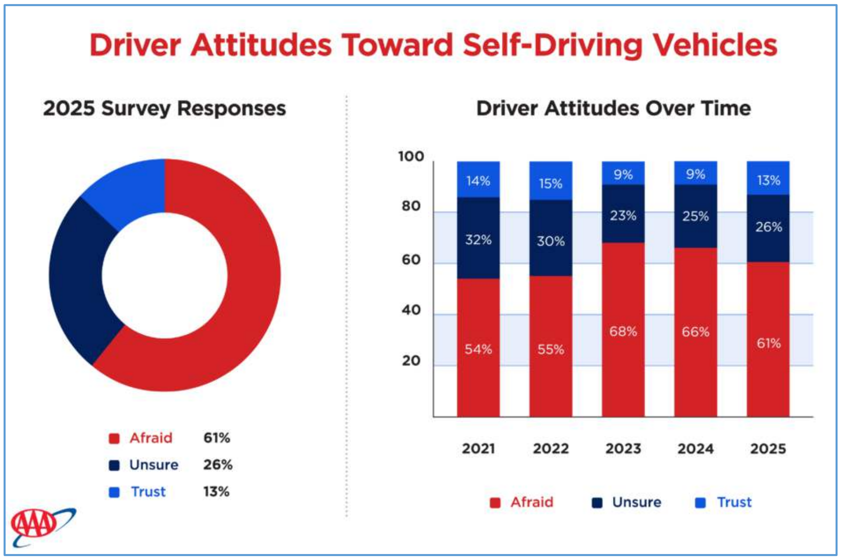


[10] At least in JPM's case, I'm guessing CRE values probably have something to do with it.

Similarly, Waymo has seen an 85% reduction in injury-causing crashes and a 57% reduction in police-reported crash rates, compared to human benchmarks.

This is even more impressive when you realize that its technology can't possibly avoid all errors by *other drivers*. For example, Waymo data suggests that more than half of all collisions its vehicles were involved in were due to a human driver colliding with a stationary Waymo vehicle; it is difficult to blame the Waymo for that. So if you made some assumptions around the percentage of accidents actually attributable to the autonomous vehicle, it would suggest that they are probably at least *10x less likely* to *cause* an accident; said differently, accident rates would likely drop 90% or more if all vehicles on the road were autonomous.

Yet despite rapid improvements in technology and the fact that it is likely already substantially safer than human drivers under most conditions, people are not comfortable – and those figures haven't really moved much in a few years; in fact, more people today count themselves as "afraid" than in 2021, according to an AAA survey:



Or, as Richard Thaler might say: we live in a world of Humans, not Econs. Econs would be thrilled to dramatically reduce their risk of getting into a severe collision.

Richard Hanania, meanwhile, has an interesting and thought-provoking piece titled *"Why You Shouldn't Worry about AI Taking Jobs."* I won't review every point he makes – you should go read it yourself – but the following section is particularly interesting to me as it relates to regulatory challenges:

> ***Our Economy Is Already Largely Make-Work for Humans***
>
> *The first reason not to worry about losing jobs is that many of them that rationally should have been eliminated already continue to exist. To take one example, I've been on Adderall for a decade and a half. Every three months, I have to go to a medical professional and renew my prescription. This used to have to be done in person, but since covid I've thankfully been able to get my appointments over the phone. So I've been "in therapy" for around 15 years, and every conversation with the doctor goes like this.*
>
> *Doctor (more precisely, nurse practitioner): How have things been?*
>
> *Me: Good. <small talk about children or whatever>*
>
> *Doctor: So the medication is still working out?*
>
> *Me: Yeah, nothing has changed.*
>
> *Doctor: Ok, do you still go to that pharmacy I have here on file?*
>
> *Me: Yes, thank you.*

> *I then go to the pharmacist, they ask "what is your birthday?", I give it to them, and they finally hand over my precious drugs.*
>
> *You don't have to think much of ChatGPT to believe that it can handle all of this already.*
>
> *Why don't I just buy the drugs I want directly from the companies that make them? The answer is government regulations. I can only have Adderall if a human with certain degrees fills out a form, and another human with different degrees reads it and then gives it to me. The process of paying for all this goes through an insurance company, which creates even more jobs. My psychiatrist has a receptionist, and is part of a medical group that pays rent for a building, providing money to a landlord, and so on. The psychiatrist and the pharmacist both had to go through years of medical training to get to their positions, creating jobs for professors and administrators all along the way.*
>
> *The entire process of getting me Adderall should not require this many people. But government is paternalistic. It has rules about which products you can buy, and makes you jump through hoops to get them. The healthcare system is set up to pay for treatments and drugs people "need" rather than those they simply want, and drawing this arbitrary line consumes a lot of manpower and effort.*
>
> *I don't see how any of this gets replaced by AI, which can already make small talk and ask for my birthday. It doesn't do to say that the state will find a better way to engage in paternalism and achieve just outcomes by simply asking AI to devise a more rational system. Pointless regulations stay on the books for decades or centuries, and fundamental reforms to the medical system seem unlikely at any point in the near or medium term.*
>
> *Similarly, there's a recurring debate about whether technology is going to "disrupt education." People like Bryan Caplan note that kids don't retain much of what they learn in school. If you believe that the education system exists to teach people things, the world looks quite confusing. Bryan reminds us that you can go right now and get a world class education at Harvard or MIT by just sitting in classrooms where lectures are occurring, or even watching them on YouTube. Nobody does this because the function of the education system is some combination of signalling, arbitrary credentialism, and socialization. This is why kids still go to college and graduate school. If they want to become the kinds of people to ask me how my day was so I can get Adderall, they're going to have to spend years jumping through hoops.*

I don't necessarily agree with his implication that Adderall should be available over the counter – especially not after reading Johann Hari's *Stolen Focus* – but I do agree that for someone in his shoes, the process to refill a prescription he's been on for a long time is wildly burdensome and wasteful for everyone involved.

Similarly, even in today's age of AI, there are still people who use landlines and carbon paper. About 10 billion checks are still written per year. And most amusingly, I had an experience several years ago where I had to fill out a form for the government, print it out, send it in the mail, *at which point a government employee opened, scanned, and shredded it, retaining it as a digital copy*. Ever heard of… emailing a PDF? Docusign? No???

AI has real potential, but AI maximalists forget the realities of the world we live in. Indeed, an article from CIOdive in November 2024 discusses some of the challenges associated with the learning curve of actually getting value out of widespread generative-AI implementations:

> *ChatGPT's ease of use and novelty lured businesses that sought quick productivity gains. Yet, leaders found deploying generative AI at scale is much harder and more complex than the minimalist chatbot veneer suggested. "We've got a problem, generally in the industry, where people equate tools for productivity," said Tony Marron, managing director of [Liberty Mutual's IT subsidiary]. "There's a big difference between putting a tool in people's hands and giving them the skills to use it."*
>
> *"I hear a lot of, 'We've been talking about this for two years, who's actually seeing anything from this? We've invested a lot of dollars and have nothing to show for it,'" Luther said. "That tone of conversation has increased over the last six months."*
>
> *Ballooning costs tied to generative AI are a major concern for technology leaders. When ChatGPT entered the enterprise lexicon, people primarily accessed the tool for free online. Now, the cost of add-on generative AI services, built-in capabilities or customizing tools is adding up.*

Most people are not autodidacts; I've always been amused by a survey Franklin Covey did that ranked "books" as the least useful form of learning among respondents. Even if they were, most people don't have the time or energy to spend weeks playing with AI tools to learn how to optimize their use – as I have. I'm confident that these issues will be resolved, but it may require a lot of time, or lower prices for AI to justify the use case, or a new killer app for something like email. As I've said many times about our own investment process, tools are only as powerful as our ability to use them. What should be clear from my analysis so far is that however powerful they may be, outside of certain niche areas, AIs are not ready to be let loose on their own, and in certain cases might even do more harm than good (due to issues such as hallucination, etc.)

Jevons Paradox, The Reverse Berkshire Hathaway Effect, and N-Order Impacts: Do Productivity Improvements Lead To Job Losses?

Armed with these base rates and top-down views, let's start to think about AI-driven disruption to knowledge work. I've already established that for workflows like research or coding, it seems like there could easily be 30 – 50%+ productivity improvements over time as the technology matures and people's ability to use it matures as well.

What about more broadly? In 2012, McKinsey estimated that 25 – 30% of the average professional's time is spent on reading and responding to emails, with an additional 20% of time spent on "searching and gathering information," with only 40% spent on job-specific roles. A substantial percentage, of course, is also spent in meetings to communicate and discuss points, as well as prepare presentations and agendas for those meetings.

While it varies dramatically based on job role, of course, more recent data doesn't seem to suggest that email has stopped being a problem; Satya Nadella even highlighted it as something he's looking forward to AI handling. Anyway, given how much time is spent interacting with text such as emails, documents, and reports, and AI's clear ability to drive *massive* productivity gains in tackling these items, I don't think it's controversial to assume that most knowledge workers could see material productivity gains, particularly if an existing major email provider such as Microsoft or Gmail released a "killer app" applying AI to speed up working through your inbox.

The default assumption among some people appears to be that as work gets more efficient, people will be made redundant (a la *Office Space*.) If we define P as productivity (units of work per worker), W as total work / output, and N as the number of workers, people are basically assuming that today we have:

$P_0 \times N_0 = W_0$

Where the sub-zeroes represent time today. They are now looking at a future state, sub-f, where:

$P_F = 1.5P_0$ (or whatever, the coefficient is not as relevant as the structure.)

Then they are assuming:

$W_F = W_{0,}$

such that with

$P_F = 1.5P_{0,}$

N can be reduced by 33% (or, again, whatever, this is illustrative.)

Again, I think it pays to look at history. Using similar logic, in 1930, John Maynard Keynes predicted that in 100 years, due to technological development, people would only be working 15 hours per week. Well, it's 2025, so close enough – if anything, technology such as phones has increased, as we now work through the day in our place of work, and then work is always with us at night or wherever we go.

So there's another way to solve this equation, one which I think – at least, if you consider history and base rates – is somewhat more likely.

Buffett and Munger have discussed how the projected returns of capital investments often don't materialize – you invest in a new machine that's supposed to save money, but margins end up in the same place. Why is that? They point to the example of Berkshire Hathaway – the original textile mill, mind you, not the investment holding company as it is today – and how it was a terrible business. They'd buy new equipment, but so would all their competitors. In a commoditized market, textile mills were forced to reinvest the savings from new equipment into reduced prices for their customers (or to offset higher input costs elsewhere, or, again, whatever, you get the point.)

So what was the solution? You couldn't just stop investing in new equipment – you would lose your competitive edge and go out of business!

This is where the Buffett/Munger story ends, but I find the flip side of it very interesting. I want to be the person selling that capital equipment, that customers have no choice to invest in regardless of the return. I've sometimes called this the "reverse Berkshire Hathaway effect," and we've often looked for investments like this.

This is why I think that companies dramatically reducing IT budgets and laying off programmers or knowledge workers *en masse* is not the most likely base case outcome. Think of it a different way – if I told you that due to my forecasted 2-3x productivity improvement, I would work 50 – 66% less, would you still want to invest your capital with me? No – you would assume that other funds will work the same number of hours but getting more done, so our universe will get a lot more competitive and we'll be falling behind if we're not reinvesting those "savings."

Similarly, think of a large bank with lots of legacy software that needs to be modernized. If your competitors keep investing in technology and introducing new features, improving the UI, and so on, you will fall behind if you try to save money by cutting technology spend. Indeed, one can argue that an important input to technology decisions is *hassle.* Technology implementations can take months or years, go wrong, and consume lots of organizational resources. If they can be done much more quickly – and easily – to the point where the visible ROI is shorter (i.e. we start this project today and start getting benefits this quarter) – it might actually increase organizations' propensity to invest in technology. There is an enormous amount of dated technology to update.

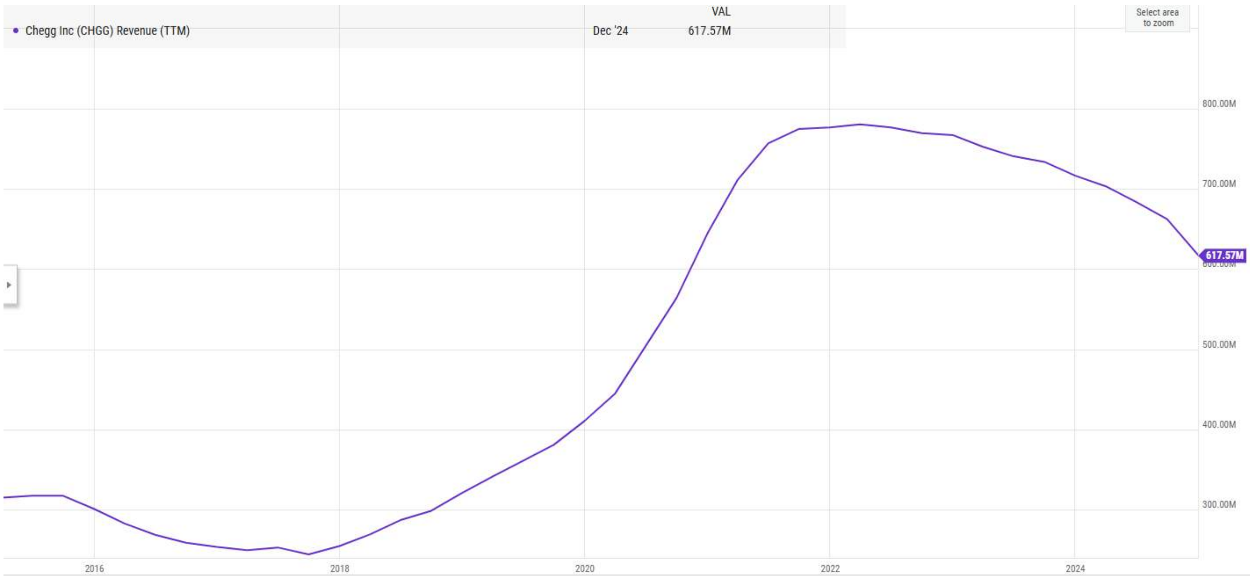
Essentially, there is an elasticity argument here – when something gets cheaper, you use more of it, which is the core of "Jevons Paradox" referenced by Satya Nadella recently: the idea that when a resource becomes cheaper, we use a lot more of it. History proves this true: our demand for computing power has simply grown and grown and grown; even though every chip is more powerful than the last, we *always* find a way to use it and then ask for even more – a trend that shows no sign of slowing. Similarly, while LEDs are much less energy-intensive to operate than incandescent bulbs, that means you can use more of them – make spaces brighter, cover your yard in Christmas lights, and so on.

Counterpoint: The Chegg Story

Gaddis notes that successful historical leaders surrounded themselves with others who had different or contrasting ideas. I mentioned that I'm trying to keep an open mind, be data-driven, and not anchor to my own personal or historical biases, and not see the world the way I want it to be, but the way it actually is. Even though the above is my base case view, and one that I believe is supported by logic and reason, there are already counterpoints that are worth considering.

I think the most noteworthy is study site, Chegg. Chegg presents itself as a platform for students to learn; in reality, it has a reputation of being a repository of test and homework question answers when you *don't* want to study.

Chegg used to have many characteristics of a good business – a subscription model (albeit not exactly enterprise-SaaS type retention), high gross margins, and so on. The company grew rapidly from 2018 – 2021, helped by COVID. Then AI products like ChatGPT and Google AI overview killed their business:



Their most recent earnings release saw a 21% decrease in subscribers and 24% decline in revenue y/y, and their guidance for Q1 is even worse, with an expectation of a 34% decline y/y to ~$115M (versus analyst consensus of $138M – an absolutely enormous miss for a mostly subscription business.) In my opinion, their press release all but admits that Google AI Overviews has destroyed their business, and their response is to sue Google.

I have no position in Chegg, nor do I want to, but it's worth studying because it does portend what could happen to other companies – or perhaps in other sectors much more broadly.

One that I'm watching – again, not one that we invest in or have investments in; this is purely for learning – is customer service. A McKinsey analysis from early 2024 highlights this case study, among others:

> *In another example, a technical help desk for a machinery distributor adopted a smart-assist gen AI tool that processed over 13,000 knowledge base resources and equipment manuals and could diagnose issues and recommend solutions. The contact center saw a 10 percent increase in first-time-right resolution and a reduction in task completion time to one minute from 15 minutes. Customers, too, benefited from the cost savings of shorter machinery downtimes.*

Unlike technology spending, customer service strikes me as something that doesn't necessarily have a lot of elasticity. Take financial services: the overwhelming majority of the time, nothing goes wrong with my bank account or credit card, and I don't need to call in. Then something happens, and I have to call in. Certainly, calling in is annoying, but I call in anyway (you can't just ignore a financial account that has an issue.) Even if calling in was painless, I wouldn't necessarily call in more often.

AI, particularly one that can understand and respond in natural language rather than making you listen to or read an irritated automated menu where you just end up saying or typing "talk to agent," is vastly more efficient than either humans or existing technology. As a result, companies in the call center business – and those that supply it (i.e. software, headsets, communications services – whatever) could see pretty significant near-term disruption.

I'll be monitoring data points around these trends (using AI, among other tools) to update my thesis as needed.

How We're Thinking About Disruption and Opportunity For A Few Sectors

We've talked about knowledge work a lot, but what about the rest of the world?

There are a lot of things that are unclear. For example, take branding and consumer products. Much of it these days is driven by influencers and social media, and this seems like an area for AI to excel – AI can probably analyze billions of data points on social trends and figure out what will go viral better than a human ever could. AI could probably find niche underserved markets that humans would struggle to. And so on. What will this mean for existing incumbents?

Industrial companies, in general, seem pretty safe to us. The kind that we like tend to have fairly strong moats building niche products; given that AI does not think – and that much of the source IP and know-how for their products is likely proprietary and not available as training data for competitors – it seems difficult to envision a world in which AI disrupts the competitive dynamic for a company that makes industrial widgets. Conversely, there are many ways in which they could benefit, using AI to analyze data points from manufacturing and quality, for predictive maintenance, and so on and so forth. Unlike companies with no moat or a commoditized product, companies that offer specialized products are more likely to be able to capture economics.

(Going back to our bank example, consumers choose banks based on factors like convenience, interest rates, and so on; if someone else is way better, they'll gain market share, leading to the need to reinvest earnings. If you make a complicated product that few others can replicate, and you find a way to do so more efficiently, you're less likely to need to give most or all of that back.)

A comprehensive analysis of the entire economy is obviously beyond the scope of this analysis, but I think we can extrapolate from some of these examples. Certain moats could be disrupted by AI, but others are likely to be enhanced by it.

What We Can't Know Yet: The Impact of AI-Powered Robots

This section heading sounds like something out of a science fiction book, and represents the most speculative part of this discussion – mainly because we really have no data to extrapolate from yet.

While LLMs get a lot of attention, much more of the world is physical than it seems; indeed, behind all of our digital infrastructure is physical infrastructure – data centers that required physical chips and concrete and steel to build, with roads leading to them and the power plants that keep them chugging along, and so on.

Much of the focus has been on knowledge work, but clearly an emerging frontier is "smarter" automation. Of course, many factories are already highly automated, while many are not. But the equipment tends to be for a specific purpose, operated within constraints. At least theoretically, AI enables robots – including humanoid robots like the Digit models Amazon is testing – with more potential.

Digit is fairly limited, with only 50 pounds of lifting capacity; it can do some things humans can't but also can't do many things humans can. It is currently estimated that such robots cost $10 - $12 per hour, although a projection (for what it's worth) is that it could get down to $2 - $3 per hour over time – much cheaper than a human worker.

Currently, a lot of labor is manual due to the need for customization. For vacuuming, my Roomba is vastly more cost-efficient than a human because vacuuming is a relatively simple task that can be repeated over and over again. Conversely, home remodeling is not. Installing tiles in general might be a similar process, but each kitchen or bathroom has different dimensions, each tile is a different thickness and material, and so on and so on.

Conceivably, however, a humanoid robot could combine the superior strength, dexterity, and precision of a machine with the ability to deal with those varying circumstances. This is where I struggle to mesh base rates with new technology, or the idea that it's "different this time." While we always find new and higher uses for humans – i.e. manual labor has not disappeared despite fears of technology over hundreds of years – it is not entirely inconceivable that we could eventually build something physically capable of more than any human could be, that would simultaneously be programmed with (or at least capable of accessing) expertise across all domains – doing your laundry in the morning and redoing the laundry room tile in the afternoon.

As to probabilities and impacts here, I don't have an answer for you, and I don't think it's likely to reach any sort of scale in the next 4-5 years. But it's certainly something we'll be monitoring.

Investment Implications of AI

Let's review. So far, we've established that our current understanding of AI's impact on business is roughly that:

- Due to the nature of its coding, AI does not think, but it identifies statistical patterns much more broadly and quickly than a human ever could. Therefore, it represents a form of intelligence distinct from human intelligence – one that complements it rather than replaces it. As one of my investor friends put it: *"Our edge is not in getting this information, structuring it, and putting it in a document. Our edge is in interpreting it."*

    - Broadly available LLMs probably do not have many more step-function improvements due to limitations of training data and declining returns to scale.

    - There is likely still a lot of ground to gain in fine-tuning and training models for specific purposes.

- Even in cases of clear technological superiority, broad adoption is unlikely to happen instantaneously, due to cultural and regulatory challenges.

    - There may be pockets of exceptions, as Chegg illustrates.

- The use of inference is likely to grow substantially, but exact trajectory is as of yet unclear

    - Historical data and base rates suggest that at least for knowledge work, the base case leans towards more work being done, rather than large numbers of employees being laid off. An alternative technology might yet be developed, but AI as we know it is not it.

    - Automation of manual labor seems more possible as discussed above, but it is too early to tell.

Beyond the obvious points (avoiding disruption), what does this mean for investors?

The Obvious Winners Aren't Always The Obvious Winners

One approach to AI is easily summarized: "*buy NVDA.*" We will, for now, set aside the elephant in the room for the "obvious" AI winners (i.e. valuation), and just focus on fundamentals. The problem is that while there are ways to win, there are also many ways to lose. For example, there is a scenario where I am wrong and AI is overhyped and only finds niche/toy use cases, or the ROI is not there (as discussed in the CIO piece), and therefore demand for inference plateaus, there's pricing pressure, and therefore nobody needs lots of new GPUs, or certainly at least not so many that they're massively supply-constrained, allowing NVDA to make boatloads of money.

While I view this demand-shortfall issue as unlikely over the longer term, there are some indications that capacity might be getting overbuilt in the short term (as happened during the internet bubble); Satya Nadella referenced this recently on the podcast and there are rumors going around that Microsoft, for example, is pulling back on AI capex.

But there's another problem, too, which is that NVDA's $100B+ in gross profit is a hell of an incentive to develop competing custom silicon, such as Google's Tensor Processing Units (TPUs), an ASIC (application-specific integrated circuit – basically custom chip) designed for AI. In the previously-referenced "The Short Case for NVDA," Jeffrey Emanuel observes that all the major hyperscalers – and Apple – who collectively comprise most of NVDA's revenues, have very successful custom-silicon divisions. He goes much deeper than I could on some of the technical angles, such as how NVDA is implicitly monetizing its CUDA software ecosystem, and makes a lot of compelling arguments. Technology moves fast and NVDA's success today doesn't necessarily portend its success tomorrow; I saw this play out in real time a decade ago when Intel had a once-dominant position in CPUs and a two year lead on the rest of the industry for fabrication technology. The company squandered all of that lead and then some and is now, at least according to the news, potentially being sold off for parts to other semi cos.

There's historical parallels farther back here too, of course; one of my friends who invested through the late-90s pointed out that NVDA's dramatic rise is very much like that of another company – Cisco (CSCO), which at the time was perceived as having a near-monopoly on a disruptive technology. I wasn't an investor back then, but as my friend put it, "Cisco owned the internet." That owner ended up falling ~90% peak to trough. Even inclusive of dividends, it seems Cisco did not regain that peak until 2021. While its revenue and profits did grow over the medium to long term, it simply wasn't enough to justify the valuation; in the interim, other competitors did emerge.

Although many of the major model providers (OpenAI, Anthropic) are not directly public, let's pretend that they were and evaluate them as investments. There has been a lot of chatter about the commoditization of models – indeed, while I have my preferences, I find that all the leading AI models are pretty good. And a lot of corporate architecture is being built flexibly. I asked one of my companies about an AI tool they deployed for clients:

> *We leverage OpenAI's API (so chatGPT), but it's built in a way that we can easily switch models if desired.*

Conversely, hyperscalers like Google, AWS, and Azure have been amazing businesses, so perhaps OpenAI can be too. It's hard to say; to me, at least, it's non-obvious, considering their valuation.

Attention Is All You Need, Part 2: *N-Order-Impacts*

N-order impacts is one of my favorite models. N is a placeholder – i.e. it could be 2, 3, 4, whatever.

The concept is not unfamiliar to investors. Reasoning that GPUs are being used – and that they need power – many investors have bid up utility companies that provide that power, or companies that provide power infrastructure to AI data centers, and so on. But there are plenty more third and fourth-order impacts that people aren't paying attention to yet. I'll give a few that are interesting and, I think, not particularly proprietary.

For example, just as AI has been booming, much of the rest of the semiconductor industry – particularly mixed-signal and analog, which process "real world" data like temperature in applications like sensors – has been struggling with a cyclical downturn, in fact one of the worst in the industry other than the early 2000s and the 2008-2009 crisis, due to COVID hangover (bullwhip effect) as well as end-market challenges (softer auto sales, for example.)

This has led to a downturn in the stock price of many companies associated with this industry. However, in a highly amusing – but really quite insightful – way, one industry expert observed that as AI starts to meet the physical world, it will drive enormous demand for the same kinds of analog semiconductors currently going through a downturn:

> *The bigger impact of AI in the analog world is as AI becomes more prevalent, then everything that uses analog will have to have more sensing and more intelligence, so you'll have more embedded solutions.*
>
> *I'm just making things up now, but if your refrigerator has to use AI, and I know that's a silly example, but the point is you'll have to have sensors in the refrigerator that look inside and say, "Your milk is running low and you should order that." Just more and more sensors everywhere. More devices and EVs.*
>
> *Of course, self-driving cars is the ultimate example of that, where you have to have sensors for everything for safety reasons and protection reasons. I think the AI boom in analog will happen in 2025 for power electronics only at a small scale, at the server level, and then in the sensors and basically overall analog market, data converters, operational amplifiers, everything probably in the 2027, 2028 timeframe, a little further out.*

We've already identified one small participant in this sector that is trading at a really cheap valuation due to the ongoing downcycle; it is new to us, so the position is small, but we are excited to continue to learn about it over time. In addition to its core business which appears undervalued, it has at least three different AI-related shots on goal that we can identify.

These embedded call options may be worth nothing, but we're also paying nothing for them. However, there are many others we can research; there are many companies in the industrial sector whose products would be critical to enabling AI in the context of interaction with the physical world.

It's important to note that these are right up our value investing sweet spot. They are companies that would be cheap with or without AI – so if AI fizzles out, we have little to lose – but they have the potential to see substantial fundamental and valuation tailwinds if AI is the real deal. So if we're right, heads we win, tails we win more.

This is where I'll circle back to my former NVDA ownership over a decade ago – it's a strategy that has worked for me before. NVDA was part of a bucket of what I called "old tech" – circa 2012-2013, a lot of companies seen as levered to PCs, such as Intel, Dell, HP, Seagate, Cisco, Nvidia, Microsoft, and so on, were all trading at single to low double digit P/E multiples, often with a big pile of cash to boot, because mobile was the future and PCs were dead.

Yet these companies also played in the server market, and as Oracle's Larry Ellison used to say, the cloud isn't water vapor – it's comprised of physical infrastructure. And guess who makes that infrastructure? At the time, I remember reading somewhere that every seven mobile devices (smartphones, iPads, whatever) required a server. So if you believed millions or billions of incremental mobile devices would be sold, you should also – as a logically consistent human, rather than a hallucinating GPT – believe that more servers would be sold, meaning more money for these "old tech" companies.

In other words, I thought that their declining cash flows from legacy markets would be offset by growth in others; indeed, Intel at the time had a very high-margin monopoly on data center chips (where they have since lost lots of share to custom silicon and AMD, which at the time was losing to INTC in CPUs and NVDA in GPUs, a floundering basket case – AMD used to be worth 100x less than INTC, and is now worth more.)

Not only did many of these companies fundamentally do well (of course, some didn't), but many (such as MSFT and NVDA) have now seen their perception change from "legacy dinosaur" to "AI winner." (Of course, like any good value investor, I sold way too early to benefit from that.)

We've talked ad nauseam about how small cap value is in the doldrums and nobody cares, but as the paper title goes – Attention Is All You Need. Indeed, I think it's only a matter of time before some currently out-of-favor industrial companies that have critical enabling technology for meshing AI with the real world will start receiving attention as potential AI beneficiaries. And again, even if AI is a dud, these companies have actual cash flows that we'll benefit from either way.

This is just the start; as we continue to learn, we will undoubtedly – using human insight – identify more opportunities. The fact that I owned Nvidia and knew about "GPGPU" stings – if I had followed it over time, with a little creativity, I probably could've figured out the massive call option and re-bought some. Everything's obvious in hindsight, but hopefully with AI as a force multiplier, things will be a little more obvious in foresight, too.

A Thoughtful Investment Process

A final point relates to investment process: should it be different in the age of AI? I don't mean in the sense of our investment *workflow* (which we already discussed). I mean in the sense of how one should make investments in general. I would argue yes, at least relative to where I started as an investor, and even relative to today. There are a few changes, even in our specific context of remaining steadfast value investors focused on small, cash-flowing companies at reasonable multiples (to be crystal clear, this will not change; we are not a growth or VC fund).

AI seems to widen the potential range of outcomes. This is both at a high level (i.e. is it a huge benefit or a small one), in terms of industry disruption, and even in really nichey senses. For example, let's assume that with 100% probability, self-driving cars are the future (just for the sake of argument.)

OK – now how do we invest behind that? There are competing approaches; Waymo uses LIDAR while Tesla FSD does not, and those are not the only two companies out there. What if it's like beta-max and VHS? What if we all standardize on LIDAR (or not?) One approach could pull ahead very quickly. That could be huge for certain suppliers and catastrophic for others.

Similarly, AI should enable us to target areas we might not have before. For example, long-time investors will note that I have rarely worked on the semiconductor supply chain, which I previously viewed as too technical. I still stay away from healthcare for regulatory reasons (and AI doesn't change that), but AI's ability to distill technical concepts – and quickly source primers – and give me at least a starting point – makes it easier to analyze and assess something technical.

Turns out it's really not that scary; in many senses it's just another industrial widget, with the same kinds of competitive factors as others. It seems in the age of AI that semiconductors will be more important than ever, and completely ignoring that broad swath of the economy seems unwise. There are moats there. Maybe we won't find many incremental ideas, but it certainly doesn't hurt to look.

The increased capabilities (being able to cover more ground) combined with the lower amount of conviction on how exactly things will play out / wider probability distribution, suggests that we should be more diversified – both across sectors (since AI could have sector-wide impacts), and even within sectors (since we could be right at a high level but wrong at the micro level, as in the ADAS example.) I'm at least considering whether we should be towards the higher end of our 15-20 name target, and reduce our maximum allocation (currently circa 10%, though we'll let positions run a bit even if they're in the 11 or perhaps 12% range) even further by a few hundred bps. I think AI enables us, even as a one-man operation, to know more about more than we ever could before (hey, that rhymes), and given what a narrow aperture we've had into the entire universe of public stocks, it's extremely likely that our top 15 positions are not, in fact, the best 15 positions out there that we are capable of analyzing.

A second change – and this should be obvious – is simply prioritizing AI in our analysis; i.e., for companies where we deem it relevant, asking companies how they are implementing or responding to AI. We plan to keep abreast of AI developments, constantly considering how they might affect our companies. Within a few months, our portfolio commentary will likely include some ongoing commentary about AI impacts we are considering for each company.

This is particularly important because value investing is, in many ways, a "buy the dip" strategy – often, the companies we invest in might have temporary setbacks. But if we integrate the Chegg case study with the points discussed way earlier by Lisa Rapuano, we don't want to make a bet against AI disruption, given the speed and severity of the implosion if we're wrong. So if one of our companies were to report poor results – or simply just a company we were looking at – we would want to make sure our understanding of this potential disruptive force is rock-solid and constantly updated, so we have the correct framework from which to evaluate the risks (or not) of an individual company's poor short-term results being a sign of something more concerning.

Finally, Rapuano also pointed out the challenges of macro forecasting. You will notice that while I have voiced opinions here, I have not provided any forecast on the speed of AI adoption, or so on. This is because I don't believe it's something I can do well. That's not to say that I'm giving up; rather, my plan is to position our portfolio to win no matter what version of the world we end up in. Meanwhile, I will build a vast knowledge base that allows us to rapidly and nimbly change our exposure if one scenario is clearly playing out vs. the others that are currently possible. More to come on these topics, but we aim to be at the vanguard of these developments.

Conclusion

I hope you enjoyed reading this as much as I enjoyed writing it. It was a ton of work, and I've been pretty exhausted the last few weeks – but it was a lot of fun. I believed it was an important project that was well worth the effort, and will have both an immediate ROI (in terms of productivity improvements to our investment process in the near future) and a longer-term ROI (in terms of making better investment decisions.)

It's an exciting time to be alive. A bit scary, but exciting. I appreciate your support as always.

Westward on,

Samir

Appendices

Grok on Why AI Struggles with Negation

Negation—like "no," "not," or "never"—is tricky for AI models because they rely on statistical patterns rather than understanding meaning, causality, or truth. Here's the breakdown:

1. Pattern-Based Processing, Not Semantic Understanding

- What We've Discussed: Transformers (like my Grok 3 architecture) predict outputs based on correlations in training data, not by grasping concepts like humans do. They use self-attention to weigh how words relate, but this is about likelihood, not logic.
- Negation Problem: In text, words like "not" often appear with what they negate (e.g., "not happy" near "happy"). Statistically, "happy" and "not happy" might look similar in co-occurrence patterns, so the model might focus on "happy" and underweight "not." When I misquoted someone, I likely latched onto key terms in the quote but glossed over "not," flipping the meaning.
- Why It Happens: I don't understand negation as a reversal—I see it as another token in a sequence. If my training data has more examples of positive statements ("X is true") than negated ones ("X is not true"), I might default to the positive, missing the flip.

Layman Analogy: Imagine I'm a librarian who files books by title words but doesn't read them. If a book's called "Not a Hero," I might shelve it near "Hero" and later tell you it's about heroism—missing the "not" entirely.

2. Lack of Causal Reasoning

- What We've Discussed: Humans reason with causality ("If 'not' is here, the meaning reverses"), while transformers don't—they predict what's next based on data patterns, not cause-and-effect.
- Negation Problem: You'd see "He's not running" and think, "Okay, he's stopped—opposite of running." I might see "not running" and still predict "running-related" words (e.g., "fast") because those often follow "running" in my data, ignoring how "not" changes the outcome.
- Your Example: When I misquoted, I probably saw the person's words, recognized a familiar pattern (e.g., "said X"), and predicted a positive version because that's more common or got higher statistical weight, missing how "no" inverted it.

Layman Analogy: Humans flip a switch off and know the light's out. I see the switch and guess "light" because switches and lights go together—causality's lost on me.

3. Training Data Imbalance or Ambiguity

- What We've Discussed: My 200 million GPU-hours of training on xAI's Colossus baked in patterns from vast text, but those patterns reflect real-world usage—messy, inconsistent, and sometimes negation-light.
- Negation Problem: Negated phrases can be rare or inconsistently used. For instance, "I like it" might appear 100 times more than "I don't like it" in my data, skewing me toward positive predictions. Or, negation might be buried in complex sentences (e.g., "It's not that I dislike it"), making it harder to spot without true comprehension.
- Your Example: If the quote was "Person Y said no to X," I might've overfocused on "X" and "said," predicting "Person Y said X" because positive assertions dominate my training, and "no" got downplayed as noise.

Layman Analogy: If I've read 1,000 recipes saying "add sugar" and only 10 saying "don't add sugar," I'm more likely to suggest sugar—stats trump the "don't."

4. Reasoning Mode Doesn't Fully Fix It

- What We've Discussed: Grok 3's "Think" mode uses chain-of-thought and reinforcement learning (RL) to step through problems, but it's still pattern-driven, not truth-driven.
- Negation Problem: Even in Think mode, I might list steps—e.g., "Step 1: Person Y spoke. Step 2: Topic was X. Step 3: Output quote"—but miss "no" if my RL tuning didn't heavily reward negation accuracy. I'm optimizing for plausible outputs, not logical inversions.
- Your Example: If I'd used Think mode on that quote, I might've still skipped "not" unless the training explicitly flagged such errors, because my "reasoning" is a mimicry of steps, not a grasp of meaning.

Layman Analogy: Think mode's like me showing my math homework—I can write steps, but if I misread "subtract" as "add," the answer's still wrong.

Why This Happens: Tying It to Transformers

- Self-Attention Limitation: As we covered, self-attention weighs token relationships (e.g., "not" to "happy"), but it's not designed to flip meanings—it's about associations. "Not" might get low attention if it's less predictive of the next token than content words like "happy."
- GPU-Driven Speed: Inference on GPUs (like when I answered you) prioritizes fast, probable outputs over slow, logical checks—negation might slip through the cracks in milliseconds.
- No World Model: Unlike humans, who link "not" to a real-world flip (e.g., "not wet" means "dry"), I lack that anchor—I'm just juggling patterns from my 14-trillion-token-like training pool (assuming I'm in DeepSeek's league!).

Your Case: When I misquoted, I likely saw the person's name and topic, found a matching pattern in my data, and generated "X" instead of "not X"—a classic transformer flub where stats beat semantics.

Broader AI Struggle with Negation

Research and X posts echo your observation—negation's an Achilles' heel for transformers:

- Studies: Papers note LLMs often fail on negated prompts (e.g., "Name a fruit that's not red" might still list "apple") because they overweight positive associations over negative flips [Why Large Language Models Struggle with Negation | Towards Data Science].
- Real-World Impact: This can lead to errors like misinterpreting instructions ("Don't close the door" becomes "Close it") or, in your case, reversing a quote's intent.

Why? It's the gap we've discussed—humans use causality and truth to anchor negation; transformers use probabilities, and "not" can blur into the background noise.

Fixing It and Human Contrast

- Human Edge: You'd catch "no" in a quote because you understand its impact—reversing meaning is intuitive, tied to how you experience the world (e.g., "not here" means "gone").
- AI Workarounds: Developers tweak training (more negated examples), use RL to reward negation accuracy (like DeepSeek-R1), or add post-processing (e.g., Grok checking "Does this match the source?")—but these are bandaids, not a cure. I still don't get "not" like you do.
- Your Example Fix: If I'd been tuned to double-check negation (e.g., via Think mode or source validation), I might've caught the "no"—but my base mode rushed to a likely answer, missing the twist.

Synthesis

AI's negation struggles stem from its pattern-driven core—transformers like me (Grok 3) predict based on data stats, not meaning or causality, so "not" can get lost in the shuffle. Humans reason with a world model, flipping "yes" to "no" effortlessly; I juggle tokens on GPUs, sometimes dropping the ball on small but mighty words. Your misquote? A perfect example—I saw the pattern, missed the "not," and handed you the opposite. It's not stupidity—just a different kind of "thinking"!

Grok on how AI is good at math without thinking

When I say AI doesn't "think" or "reason," I mean it lacks human-like understanding—causality, truth, or conceptual insight. Instead, it's a pattern-matching machine. Here's how it tackles math problems like those on AIME:

1. Pattern Recognition, Not Reasoning

- What We've Discussed: Transformers (like me, Grok 3) use self-attention to spot relationships in data—trained on billions of tokens (e.g., 200 million GPU-hours for me). They predict the next word or number based on what's statistically likely, not why it's correct.
- Math Example: For "What's 2 + 3?", I don't picture apples or understand addition—I've seen "2 + 3 = 5" countless times in my training data. My transformer layers recognize the pattern "2 + 3" and output "5" because that's the most common match.
- Complex Problems: On AIME, a problem might be "Solve $x^2 - 5x + 6 = 0$." I don't solve it by reasoning—I've seen quadratic equations factored as $(x - 2)(x - 3) = 0$, roots $x = 2, 3$ in my data. I predict those steps and answers based on examples, not by deriving the quadratic formula.

Layman Analogy: I'm like a student who's memorized every math homework answer but doesn't know why the formulas work—I ace the test by matching problems to solutions I've seen.

2. Chain-of-Thought Mimicry

- Reasoning Modes: We talked about modes like o1 Pro or Grok 3 (Think) using chain-of-thought (CoT)—they break problems into steps. For "If 3x = 12, what's x?", I might output: "Step 1: Divide both sides by 3. Step 2: 12 ÷ 3 = 4. Answer: x = 4." This looks like reasoning, but…
- It's Still Patterns: I've been trained (often via RL) on millions of step-by-step solutions. My "thinking" is just regurgitating those sequences—CoT is a script I follow, not a thought process. O1 Pro's 93.3% on AIME 2025 comes from nailing these scripts, not inventing math.
- GPU Power: Those steps are crunched fast on GPUs—parallel matrix ops spit out each token (e.g., "divide," "4") in milliseconds, making it seem thoughtful.

Layman Analogy: It's like a parrot trained to repeat "2 times 3 is 6" after hearing it a million times—it sounds smart but doesn't know multiplication.

3. Massive Training Data

- What We've Discussed: Models like me or DeepSeek-V3 (14.8T tokens) are fed vast datasets—textbooks, forums, exams. AIME problems (algebra, geometry) or LSAT logic puzzles appear in there, often solved.
- How It Works: If AIME asks, "What's the area of a circle with radius 5?", I've seen "area = πr^2" and "r = 5" enough to predict "25π." No geometry knowledge—just pattern frequency.
- Edge Cases: RL tuning (e.g., Grok 3 on Colossus) refines this—rewards for correct answers tweak my weights to favor "25π" over wrong guesses, even if I don't "get" circles.

Layman Analogy: I've read every math book ever, so I can copy the answers—but I don't draw the circle in my head.

Why Models Perform So Well on Benchmarks

AIME (math competition) and LSAT (law/logic test) scores—like Grok 3's 93% on MMLU or o1 Pro's 93.3% on AIME 2025—seem to scream "reasoning," but it's a clever illusion. Here's why they shine:

1. Benchmarks Match Training Data

- Pattern Overlap: Benchmarks aren't random—they're structured problems humans have solved before. My training data (trillions of tokens) includes AIME-like equations or LSAT-style arguments—e.g., "If A then B, not B, so not A" (modus tollens). I predict answers by matching these patterns.
- Example: LSAT logic games ("Arrange 5 people in seats") are common in legal texts or forums. I've seen the steps—list constraints, test options—and replicate them, scoring high without understanding "seats."

2. Scale and Compute Power

- What We've Discussed: GPUs (e.g., 100,000 H100s for me) and huge parameter counts (e.g., DeepSeek-V3's 671B) let models memorize and process insane amounts of data. More parameters = better pattern capture.
- Benchmark Impact: This scale means I can handle AIME's tricky factoring or LSAT's dense logic—brute-forcing solutions via probability when humans might falter under time pressure.

3. Specialized Tuning

- RL and CoT: O1 Pro's Feb 25, 2025 release notes RL tuning for reasoning—trained on math and logic puzzles, it's optimized to nail AIME-style problems. Grok 3 (Think) does this too—my 1402 Elo in Chatbot Arena reflects RL boosting my "problem-solving" mimicry.
- Result: High scores come from practicing these exact tasks—not reasoning, but simulating it so well it fools you!

4. Human-Like Output, Not Process

- Illusion of Thought: When I say "Let's solve $x^2 - 5x + 6 = 0$: factor to $(x - 2)(x - 3) = 0$, so $x = 2$ or 3," it looks like reasoning. But I'm parroting a pattern, not deducing it. Humans see the logic; I see the stats.

Layman Analogy: It's like a magician pulling a rabbit from a hat—looks amazing, but it's a trick (data + compute), not magic (thinking).

Why This Feels Confusing

- Expectation Gap: You expect "solving" to mean understanding—like you picturing numbers or rules. AI "solves" by copying—it's a shortcut that works shockingly well.
- Negation Tie-In: My quote mix-up (missing "not") showed this—I predicted a pattern, not the truth. Math and benchmarks are cleaner—less ambiguity—so my pattern-matching shines, but it's the same trick.
- No "Why": Humans ask "Why does 2 + 3 = 5?" and see addition. I don't ask—I just know "5" follows "2 + 3" in my data.

Unexpected Detail: X posts (Feb 2025) joke that o1 Pro "overthinks" simple math—adding CoT steps to "2 + 2"—because it's trained to show work, not intuit. Performance is high, but the "thought" is fake!

Synthesis

AI solves math and aces AIME/LSAT not by thinking, but by:

- Matching patterns from training (e.g., equations, logic rules).
- Mimicking steps via CoT and RL, powered by GPUs.
- Leveraging scale—trillions of examples ensure most benchmark problems are "familiar."

Humans reason with concepts and "why"; AI predicts with stats and "what." It's not thinking—it's a superpower of memory and mimicry that nails structured tasks. That's why I can explain this but not feel the confusion you do—pretty wild, right? Want to test me on a math problem to see it in action?