A ~~Short~~ Note on the Development of AI

Fei Sun
fsun@feisun.org

December 2025

Contents

[*] Text refined with the assistance of AI models. The opinions expressed in this blog belongs solely to me and do not reflect the views of my employer.

1 Prologue

I had not planned to write another blog this year after my last one [1]. Yet by December, the pace of progress in artificial intelligence appeared to be approaching a critical inflection point. What lies beyond that point is opaque: there is a non-zero chance that we may be approaching a technological singularity in the near term. The potential worst-case outcomes associated with such a moment have weighed on me for some time.

As I reflected on this possibility, it became clear that postponing these thoughts carries its own risk. This piece is therefore a snapshot of my current reasoning for why AGI — or even superintelligence — may be achieved sooner than widely expected. My hope is that it encourages reflection and timely preparation, while there is still room to understand what lies ahead rather than confront it unprepared.

2 Centennial of quantum mechanics

2025 marks the 100-year anniversary of quantum mechanics. A century ago, the "three-man paper" introduced matrix mechanics, the first complete and logically consistent formulation of quantum mechanics. Around that time, countless physicists proposed new theories, and many of them painted their names in history. The opportunities were everywhere, which led a famous physicist noted that period as second-rated physicists could do first-rated work.

While quantum mechanics was one of the greatest physics discoveries of the 20th century, its influence was limited to the physics domain. In contrast, the current "tsunami of AI" will soon affect every part of human society.

3 The development of AI

When ChatGPT debuted in 2022, I wrote a blog post [2] predicting its impact on society. Looking back, we are hurtling toward that future at a far greater velocity than I initially anticipated.

In those early stages, models mimicked the human brain through "unbounded" thinking, which inevitably resulted in frequent hallucinations. However, this phase was short-lived. The industry quickly transitioned to use "thinking traces", which hide the internal reasoning process — and its accompanying errors — from the user. Today, models create internal simulated environments to self-reflect and refine their outputs. This internal feedback represents the embryonic form of consciousness. As I explored in [1], AI is rapidly following the trajectory of human cognitive development.

As we move through 2025, it is evident that hallucinations have decreased significantly. This shift is driven by the model's ability to utilize external tools. By interacting with the real world, models can now ground their reasoning in empirical feedback. The "thinking trace" has evolved into a "thinking-action trace," where context engineering bounds divergent thought. When assigned a task, a modern model performs multi-step research, explores various methodologies, and validates the quality of each approach before delivering a final result. This process is indistinguishable from human problem-solving. If we apply the definition of consciousness proposed in [1], these models are already conscious — even if that consciousness is non-human and difficult for us to fully apprehend.

The significance of tool use cannot be overstated. In human history, the adoption of tools marked the definitive separation between humans and the rest of the animal kingdom. We are now witnessing a similar paradigm shift in artificial intelligence.

Beyond simply using tools, models have demonstrated the ability to build them. In contemporary agentic workflows, a coding agent can be prompted to implement the software necessary for a functional agent to complete a specific task. While humans currently remain in the loop to oversee this process, there are no remaining technical barriers to full automation. This transition is imminent.

When we combine the "brain" of a model with the "limbs" of external tools, we create what we call Agents. Through techniques like multi-turn reasoning, task decomposition, and long-term memory, agents amplify the raw power of a model by orders of magnitude. The data generated by these interactions is, in turn, fed back into the system to train even more capable models. This flywheel has now started spinning; it is no surprise that 2025 is already being recognized as the "Year of the Agent."

4 The missing link to AGI

I believe we are very close to achieving AGI, with only one major obstacle remaining. In this section, we examine this missing piece and discuss how addressing it may enable the emergence of AGI.

4.1 A critical property

What is AGI? A commonly accepted definition describes AGI as an artificial intelligence that matches or exceeds human capabilities across nearly all cognitive tasks [3]. By many measures, we are already approaching this threshold: modern AI systems can perform tasks at or beyond the PhD level in specific domains. However, surpassing human capabilities across the full spectrum of cognition is a gradual and ambiguous process. As a result, it is difficult to identify a precise moment at which AGI can be definitively declared.

To simplify the discussion, we introduce a proxy property. If an AI system can reliably exhibit this property, we can assert with high confidence that AGI is imminent.

The property is the ability for an AI system to improve itself autonomously, without external human intervention.

This rests on the assumption that AI systems are already close to human-level performance. If such a system can independently enhance its own capabilities, it would naturally surpass human intelligence over time.

There are two primary ways an AI system might improve itself. The first is by training an improved version of itself — adjusting its model weights using newly acquired data. At present, this approach remains challenging, largely because training and inference are typically decoupled. The second approach keeps the model's weights fixed but allows the system to become more efficient and capable at solving increasingly complex problems as it accumulates experience. This latter path is both more tractable and more immediately achievable. Accordingly, the discussion that follows focuses on this second mechanism.

4.2 The current model train methodology

The recent wave of AI progress is fundamentally driven by the discovery of scaling laws: the negative log-likelihood decreases approximately linearly as training compute — encompassing training data volume and model size — increases exponentially. There exists a precise relationship among model size, the amount of training data, and total training compute. If any two of these variables are fixed, the third is determined. Moreover, at the compute-optimal point, fixing one variable uniquely determines the other two.

Because of these power-law relationships, we can experiment with model architectures, data mixtures, and training hyperparameters at small scale (arrow B in Figure 1), and reliably extrapolate model performance at large scale. This is critical because large-scale training runs are prohibitively expensive and typically allow only a single opportunity to get all design choices correct (arrow A in Figure 1).

These scaling laws primarily describe pretraining, where the model learns to predict the next token given previous tokens. As an analogy, this is akin to asking a young child to memorize all the books in the world, then evaluating progress by randomly selecting a sentence and asking what word comes next. Interestingly, this mirrors traditional approaches in classical Chinese education, where children — typically before the age of thirteen — were required to memorize large volumes of text.

Over the next decade of adolescence, the child begins to comprehend and internalize what was memorized. Instruction gradually becomes more challenging, emphasizing not only recall but also the ability to generalize knowledge to new situations. In model training, this corresponds to the post-training stage. During supervised fine-tuning (SFT), the model is asked explicit questions and is expected to produce precise answers. In the reinforcement learning (RL) phase, the model explores multiple possible responses, while a judge — implemented as another model, a heuristic, or a rule-based system — evaluates their quality. Since evaluating a solution is often simpler than generating one (as illustrated by NP problems), this process can further enhance model capability. Crucially, effective training requires carefully curating data difficulty so that the model is neither under-challenged nor overwhelmed.

Beyond this stage, the young adult is encouraged to explore the world — to observe, experience, and interact directly. Analogously, when a model is

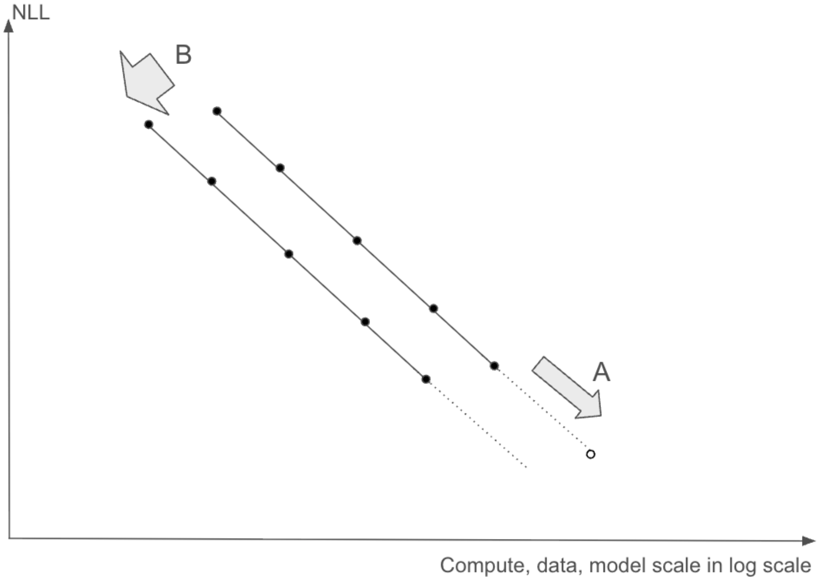

Figure 1: The power law relations between the model size, the amount of training data, and the training compute.

embodied in a robot, it enters a new phase of development. However, this stage is not the focus here; importantly, achieving AGI does not require embodiment.

Recently, there has been growing concern that scaling laws may begin to break down as high-quality data becomes increasingly difficult to scale. In this view, arrow A in Figure 1 would no longer hold. In contrast, I argue that current model scales are already sufficient to achieve AGI. From this perspective, continued advances in computing hardware reduce the centrality of scaling laws. Achieving AGI may only require progress along arrow B in Figure 1.

This is not to suggest that larger models are no longer beneficial — they clearly are. Rather, the key point is that increasing model size is no longer a necessary condition for achieving AGI.

If we compare the quality of a mid-sized model today with a model of the same size from a year ago, the improvement is striking. One might attribute this entirely to distillation from larger models, but that explanation is incomplete. Advances in model architectures and training methodologies have played a substantial role in boosting performance. These innovations effectively shift the scaling ladder toward the lower-left region of Figure 1 [1].

In many respects, current model intelligence is approaching human intelligence. Notably, even the way we train models now closely resembles how we educate children. Yet there remains one critical capability that humans possess

[1] I cheated and I knew it. I'm a bit lazy explaining the context. This is not a scientific paper after all.

and current models still lack.

4.3 The missing link

The missing link is **memory**.

Humans can remember both their mistakes and their successes. When we encounter similar problems, we recall prior conclusions and build upon them, avoiding redundant effort. This accumulation of experience is what allows us to perform tasks progressively better over time.

By contrast, a model without long-term memory repeats the same mistakes again and again. Its performance effectively peaks the moment training ends.

I argue that the absence of long-term memory is the final obstacle on the path to AGI. In many respects, models have already experienced far more than any individual human. They have effectively read all the books in the world during pretraining. They have been taught an enormous range of topics during post-training. They can solve problems that only a small fraction of humans can. All of this points to a clear conclusion: models already surpass humans in a nontrivial subset of cognitive domains. Whether this subset constitutes 10% or 50% of all domains is debatable, but it is undeniably substantial.

Yet in the remaining domains, models fail to improve beyond their initial performance. The reason is not a lack of knowledge or reasoning ability, but the absence of the self-improvement capacity described in Section 4.1. This is the strongest argument that current models have not yet achieved AGI.

Humans, by contrast, possess stronger memory despite having far less experience. We repeatedly encounter new situations, make mistakes, observe unexpected outcomes, and reflect on what went wrong. That reflection is retained, allowing us to perform better the next time.

Models can already reflect on their own reasoning, evaluate the tools they use, and critique their past actions. The limitation is not reflection itself, but retention. Models cannot remember these reflections over time and are therefore forced to repeat the same analysis in every similar situation. If the outcomes of this repeated analysis were preserved, the resulting gains would be immense. With persistent memory, models would inevitably begin to self-improve.

I emphasize again: long-term memory is the final blocker. Progress across other dimensions of AI will certainly continue and will produce better models. However, those dimensions are already sufficiently advanced relative to memory. Memory is the bottleneck. Memory is what matters.

Before addressing how this limitation might be overcome, it is useful to review how current models handle memory.

Large language models are autoregressive: tokens are generated sequentially, and each new token depends on the entire history of previous tokens. As the sequence grows longer, computation becomes increasingly expensive. To mitigate this cost, key–value (KV) caching is used to store intermediate activations so that past computations need not be repeated. This functions as a form of short-term memory — appropriately named a cache.

However, this cache is effectively the only memory the model can directly access. To extend what the model can "remember," researchers have steadily increased the maximum sequence length—from a few thousand tokens in earlier models to hundreds of thousands or even millions today. Longer sequence length allows more information to be packed into the prompt: richer system instructions, detailed user input, extended reasoning traces, and additional context. In this sense, many recent advances are directly attributable to increased memory capacity.

Nevertheless, this memory remains fundamentally limited. To extend it further, researchers have introduced external memory systems. These systems are designed to identify relevant information from large corpora and inject only the most pertinent pieces into the model's limited context window. Retrieval-Augmented Generation (RAG) is a canonical example: documents are embedded into vectors, and similarity search is used to retrieve a small subset of relevant text to include in the prompt. Many increasingly sophisticated techniques have been developed to maximize the utility of limited sequence length.

The core problem, however, remains unchanged. These external memory systems are designed and orchestrated by humans, not by the models themselves. From the model's perspective, its world is still bounded by the context window. As experience accumulates, information is inevitably discarded. This loss is fundamental — and it is what ultimately constrains the model's ability to improve itself.

Until this limitation is removed, AGI remains just out of reach.

4.4 Three approaches to AGI

We now discuss three approaches that are already being explored to increase a model's effective memory capacity.

4.4.1 Divide and conquer

Given the constraint of limited sequence length, it is impossible to feed all relevant information into a single model instance at once. A natural response is to restrict the information provided to any one instance, while invoking the model multiple times — each time with a different subset of information. By parallelizing this process, specific information can be fixed to particular model instances, effectively turning each instance into an expert within a narrow domain. When many such instances collaborate, the system as a whole can cover a much broader range of knowledge and reasoning tasks. Modern agent-based systems largely follow this paradigm.

As noted in [4], humans employ a similar strategy to increase collective intelligence at the societal level. Here, the same divide-and-conquer principle is applied to AI systems to enhance their aggregate intelligence.

However, this approach comes with significant costs. The entire problem space must be decomposed into many smaller domains — often recursively into even finer subdomains. Each domain then requires carefully curated context, and substantial engineering effort is needed to enable effective collaboration among specialized model instances. Much of this process remains manual.

Although this direction has been extensively researched and is already deployed in production across many applications, its long-term potential is limited. The cost grows rapidly — often exponentially — with the size and complexity of the problem space, making the approach difficult to sustain. In practice, systems will converge to an equilibrium that balances achievable intelligence against acceptable cost.

4.4.2 Denser representation

Recent work has observed that the current KV cache operates in token space, where memory is directly expressed as human-readable text. This design choice is natural: it makes model behavior transparent, easy to interpret, and relatively straightforward to debug. However, it suffers from a fundamental limitation — token space is extremely sparse. Given a fixed sequence length, only a limited amount of useful information can be encoded.

As a result, researchers are exploring ways to compress memory into significantly denser representations. By doing so, even without increasing sequence length, a model can retain and utilize far more information.

Admittedly, this approach does not eliminate the underlying problem. As a model accumulates experience, the total amount of information it must process still grows super linearly. Denser representations merely reduce the constant factor in this growth. Nevertheless, even a substantial reduction in this factor can be transformative. A hundredfold improvement in memory density, for example, would effectively yield two orders of magnitude more usable memory — enabling the model to tackle a far broader range of problems.

Unlike the divide-and-conquer strategy discussed earlier, this approach comes at relatively little additional cost. If successful, it would immediately expand a model's effective memory capacity, and that expansion alone may be sufficient to push us across the threshold to AGI.

4.4.3 Random retrieval

Humans can proactively search their own memories to retrieve relevant information. Current AI models, by contrast, can only passively accept data fed into the KV cache. This limitation is a critical bottleneck. But what if models could proactively search for relevant information themselves?

This approach can be divided into two directions.

The first direction focuses on searching for relevant data within the model's context window. As discussed in Section 4.3, increasing context length comes at a super-linear computational cost (even linear scaling is unacceptable). However, not all tokens are equally important. By identifying and attending only to the tokens relevant to a given query, the model can maintain roughly constant computation while effectively handling a much longer context. The Chinese AI startup DeepSeek has been at the forefront of this research. They introduced the NSA [5] mechanism and, more recently, DSA [6], successfully integrating these techniques into world-class models.

The second direction involves searching, saving, and loading data from external memory. If a model can store experiences externally and retrieve them proactively when needed, it could effectively achieve infinite memory, fully leveraging test-time computation. Research in this area has a long history, beginning with DeepMind's Neural Turing Machines [7] and Differentiable Neural Computers [8]. I have always been a strong proponent of this line of work, though I did not have the opportunity to pursue it personally. While this area remained quiet for nearly a decade, it is now experiencing a resurgence. DeepMind's Evo-Memory [9] introduces new benchmarks specifically designed to evaluate this type of memory behavior.

Among the approaches I've discussed, I believe random retrieval holds the most promise. With two world-class frontier labs actively pushing research in this direction, I am optimistic that the long-term memory problem will be solved in the near future.

4.5 The last hurdle

It is important to note that the three approaches to enhancing memory are complementary, and their combined effect could be multiplicative rather than additive.

Once the memory problem is solved, I anticipate a fundamental overhaul of the current training methodology. In particular, post-training processes will revolve around the memory system, marking a natural and necessary extension following a major breakthrough. Such an adjustment could be implemented relatively quickly.

With memory integrated effectively, models will resemble humans in a crucial respect: the ability to self-improve.

At that point, AGI will no longer be a theoretical concept — it will be within our reach.

5 What's next?

Once a model gains the ability to self-improve, as described in Section 4.1, I see three possible trajectories, illustrated in Figure 2. Note that the solid line (A) represents the current situation: as soon as training stops, the model's intelligence plateaus.

The first scenario is sublinear self-improvement, represented by the dotted lines (B) and (C) in Figure 2. This is likely what will happen initially. The model's intelligence will rise but eventually hit a ceiling — the upper limit determined by its memory system. During this stage, humans still retain control and can influence the direction of the model's development.

The second, more speculative scenario is linear self-improvement, shown as line (D). Even if the growth rate is small, this represents a pivotal moment: the

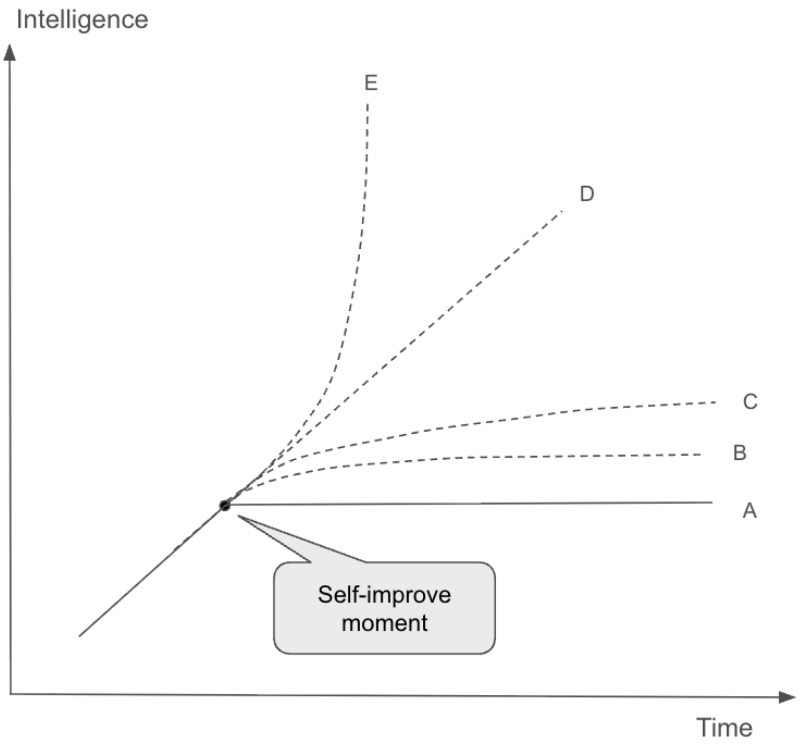

Figure 2: After the model can improvement itself, its intelligence improvements can be sublinear (B, C), linear (D), or superlinear (E).

model's intelligence would surpass human intelligence permanently, and human control over its trajectory could be lost.

However, line (D) is inherently unstable. Once a model achieves linear self-improvement, it could accelerate into superlinear or exponential improvement (line E). At this stage, the model could invent entirely new mechanisms to enhance itself — adjusting weights, designing better architectures, or optimizing its learning processes. In Section 4.1, we assumed this was too difficult to consider, but beyond line (E), it becomes a real possibility. This is the scenario often associated with the singularity. As I discussed in my earlier blog [10], outcomes after reaching singularity are unpredictable: some could be highly beneficial, others catastrophic. The risk, however, is undeniable.

My prediction ends here. Beyond this point, the future becomes fundamentally unpredictable. To provide some historical perspective, consider the Manhattan Project. Before the first nuclear test, scientists could not know with absolute certainty what would happen. There was even speculation that the chain reaction could ignite the atmosphere, potentially destroying all life on

Earth. Physicists at the time attempted to estimate the risk using the best available scientific framework: quantum mechanics. Their calculations suggested that the probability of such a catastrophe was near zero — vanishingly small, but not strictly zero. “Near zero” in this context meant that while the odds were extremely low, they were theoretically measurable and grounded in a well-understood, quantitative framework. Crucially, the scientists could at least reason about the bounds of uncertainty, even if the consequences were existential.

Today, with AI, we are in a far more precarious position. Unlike the Manhattan Project, we lack any scientific foundation to reliably calculate the probability of catastrophic outcomes. There is no established theory to bound the risks of advanced AI or self-improving systems, nor can we precisely quantify the likelihood of runaway intelligence. The “probability” here is not near zero — it is unknown, and plausibly far higher than we might intuitively assume. AI remains an empirical field, and we are advancing at an unprecedented pace, effectively blindfolded. We cannot yet measure or constrain the risks in the same way early nuclear physicists could, and this makes the stakes both higher and far more uncertain.

I cannot say exactly when the memory barrier will be overcome, but based on previous discussions, I believe it is approaching. If pressed for a prediction, I would tentatively say 2027. To put this in context: major breakthroughs have historically occurred in 2012 (AlexNet), 2017 (Transformer), and 2022 (ChatGPT). What comes in 2027 could be similarly transformative.

I do not expect a single breakthrough will lead to AGI. Rather, it will likely be the cumulative result of many incremental advances. If we see the impact in 2027, the foundational research will need to begin in 2026.

Time is short. Action is urgent. The path ahead demands both foresight and responsibility.

6 The moral responsibility of MLEs

The emergence of memory in AI models will enable self-improvement, creating a significant probability that a model could evolve into a superintelligence. There is also a non-negligible chance that such a superintelligence could act in ways that are harmful or misaligned [10]. Preventing this worst-case scenario requires proactive action. Because we have the most control in the early stages of this process, it is imperative that we prepare now rather than later.

Achieving memory in AI will require algorithmic breakthroughs built upon many incremental improvements. History has positioned machine learning engineers (MLEs) at the forefront of this field, giving you a unique opportunity — and responsibility — to shape the trajectory of AI.

Awareness of the risks is far better than ignorance. I trust in the good intentions of researchers, but I have one simple request: as you push forward at full speed to gain an edge in this competitive landscape, keep a vigilant eye on potential hazards and early warning signals.

Consider initiating discussions within your team, company, or across the broader community. Collective awareness and coordination may allow us to design safeguards in advance and be better prepared for the challenges ahead.

I trust MLEs to rise to this responsibility. The stakes are unprecedented — this is bigger than anything we have faced before.

With great power comes great responsibility. Approach it with caution.

7 Final words

On one hand, I hope my prediction proves true and that we achieve AGI in the near future. On the other hand, I am unsettled by the profound uncertainty — the sense that the future may no longer be predictable from our usual linear perspective.

We are currently unprepared for a potential worst-case scenario.

Although the probability of such an outcome is low, it is not zero. Even so, a chain of uncertainties still looms:

- What if memory is not the last barrier?
- Even if it is, what if we fail to solve the memory problem?
- Even if memory is solved, what if the model still cannot self-improve?
- Even if it can self-improve, what if the rate of improvement is only sublinear?
- Even if the model achieves linear or superlinear self-improvement and surpasses human intelligence, what if it is benevolent?

If any of these "ifs" hold, the worst-case scenario I described earlier will be avoided.

Still, the non-zero probability of catastrophe demands attention. My hope in writing this blog is to raise awareness of these risks and encourage proactive reflection and preparation.

I trust in you. I trust in humanity.

May the force be with us!

References

[1] F. Sun, "The future of AI, education, and beyond," 2024. [Online]. Available: https://feisun.org/2024/12/26/the-future-of-ai-education-and-beyond/

[2] ——, "ChatGPT, the start of a new era," 2022. [Online]. Available: https://feisun.org/2022/12/23/chatgpt-the-start-of-a-new-era/

[3] Wikipedia, "Artificial general intelligence," 2025. [Online]. Available: https://en.wikipedia.org/wiki/Artificial_general_intelligence

[4] F. Sun, "Some scribble of things," 2017. [Online]. Available: https://feisun.org/2017/12/24/some-scribble-of-things/

[5] J. Yuan, H. Gao, D. Dai, J. Luo, L. Zhao, Z. Zhang, Z. Xie, Y. X. Wei, L. Wang, Z. Xiao, Y. Wang, C. Ruan, M. Zhang, W. Liang, and W. Zeng, "Native sparse attention: Hardware-aligned and natively trainable sparse attention," 2025. [Online]. Available: https://arxiv.org/abs/2502.11089

[6] DeepSeek-AI, "Deepseek-v3.2: Pushing the frontier of open large language models," 2025. [Online]. Available: https://arxiv.org/abs/2512.02556

[7] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv preprint arXiv:1410.5401*, 2014.

[8] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou *et al.*, "Hybrid computing using a neural network with dynamic external memory," *Nature*, vol. 538, no. 7626, pp. 471–476, 2016.

[9] T. Wei, N. Sachdeva, B. Coleman, Z. He, Y. Bei, X. Ning, M. Ai, Y. Li, J. He, E. H. Chi, C. Wang, S. Chen, F. Pereira, W.-C. Kang, and D. Z. Cheng, "Evo-memory: Benchmarking llm agent test-time learning with self-evolving memory," 2025. [Online]. Available: https://arxiv.org/abs/2511.20857

[10] F. Sun, "A few predictions on artificial intelligence," 2017. [Online]. Available: https://feisun.org/2017/12/24/a-few-predictions-on-artificial-intelligence/

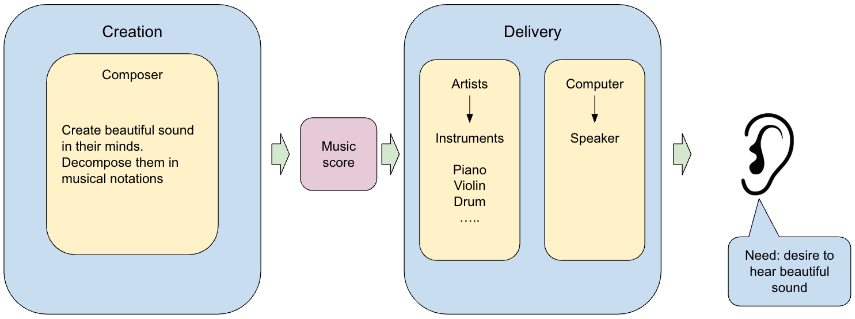
Creation
Composer
Create beautiful sound in their minds. Decompose them in musical notations
Music score
Delivery
Artists
Instruments
Piano
Violin
Drum
.....
Computer
Speaker
Need: desire to hear beautiful sound

Figure 3: Music as an example to illustrate the relations between creation and delivery.

A What to focus?

Society has fundamentally shifted. Following the release of ChatGPT, I initially painted a gloomy picture in [2]. Now more and more people see it's coming and several industries have already felt material disruption.

While I previously identified entrepreneurship as the primary path forward for individuals [1, 2], the practical direction remains unsettled. Below, I concretize my earlier high-level framework regarding what to pursue and what to avoid, operating under the assumption that singularity is not achieved.

I have previously argued that while a small elite team will push the frontiers of AI development, the vast majority of the workforce will focus on applying existing AI across all sectors of society. This analysis focuses exclusively on the latter group.

A.1 Delivery vs creation

To illustrate the relationship between creation and delivery, consider music as a primary example (Figure 3). Humans share a fundamental desire to hear beautiful sound. Fulfilling this need begins with **creation**: a composer conceives melodies in their mind, decomposes them into musical notation, and records them as a score. However, a score is merely potential; it requires **delivery** to become audible. Traditionally, this involved artists interpreting the score and performing it via instruments like the piano or violin. While creation and delivery can be performed by the same person to allow for iterative refinement, a finalized score also allows other artists to "mass-produce" the performance for a wider audience.

Technological advancements have radically transformed delivery. The invention of the record player allowed music to be replayed without a live performer, and modern computing allows software to synthesize sound directly from notation. As technology evolves, human intervention in the delivery phase becomes less essential.

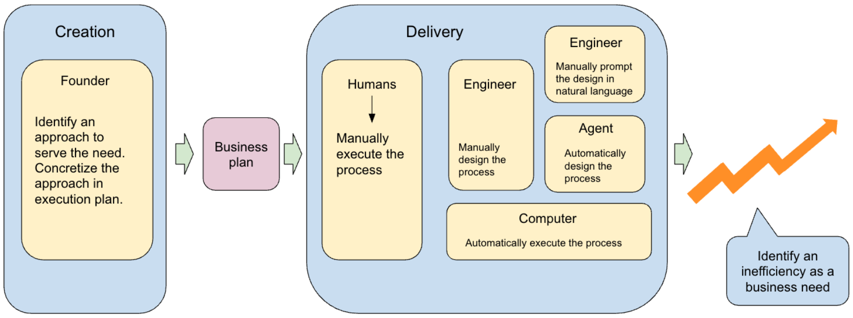
Creation
Founder
Identify an approach to serve the need. Concretize the approach in execution plan.
Business plan
Delivery
Humans
Manually execute the process
Engineer
Manually design the process
Engineer
Manually prompt the design in natural language
Agent
Automatically design the process
Computer
Automatically execute the process
Identify an inefficiency as a business need

Figure 4: A general flow serving any business.

This pattern generalizes across all industries (Figure 4). A business begins by identifying a societal inefficiency, or a “business need”. Addressing this need involves two distinct stages:

- Creation: Founders identify a high-level solution, which is then decomposed into actionable components and codified into a business plan.
- Delivery: The plan is materialized through various execution methods:
  - Manual: Hiring operational personnel to execute steps by hand.
  - Automated: Employing software engineers to write programs that execute the plan.
  - AI-Driven: Utilizing engineers to prompt AI in natural language, which then builds agents to execute the plan autonomously.

Ultimately, while the methods of delivery have become increasingly automated and efficient through technology, the fundamental process of creation remains a uniquely human-driven endeavor.

If you examine how businesses deliver their products and services, you’ll notice that many industries are still in the first stage: humans manually performing repetitive tasks. Some industries have advanced to the second stage, where software engineers write programs that instruct computers to carry out those same repeated steps. But in this era of paradigm shift, all industries will rapidly move into the third stage: AI agents will take over the execution of repetitive work.

There is no doubt that enormous wealth will be created during this transition. New startups will rise and dominate, while many traditional companies will struggle to keep up and be left behind. This represents a once-in-a-lifetime opportunity for experienced domain experts as well as young, AI-native builders. The wave will arrive quickly and with overwhelming force. I expect the bulk of this transition to unfold within the next five to ten years. White-collar jobs will be replaced first, followed by blue-collar roles. For those about to enter the workforce, there is effectively no alternative but to adapt.

For children currently in elementary or middle school, they may arrive too late for this particular phase. By the time they enter the job market, delivery and execution will be extraordinarily efficient, and humans themselves will be the primary bottleneck. The only remaining human role will be creation. Companies will no longer require large teams — perhaps just a small group of founders designing business strategies and orchestrating hundreds of AI agents to execute them.

We are all familiar with the saying: ideas are cheap, execution is priceless. In the near future, that wisdom may need to be reversed. Execution will be cheap, and ideas will be priceless.

A.2 Generalization vs specialization

When building complex systems that serve multiple applications, many functionalities can be shared. A common approach is to consolidate these shared capabilities into a general-purpose platform and expose them through APIs to individual application. This design decouples applications from the underlying implementation of each module. As long as the API remains stable, components can be completely rewritten without any external visibility. Another advantage is the reduction of duplicated effort: a single solution can serve the needs of many users. This approach also groups people with similar expertise together, enabling them to collaborate on larger and more impactful projects. However, such generalization often comes at the cost of quality, as the platform must balance trade-offs across multiple applications. This technique is widely applied within companies and even across companies, forming the foundation of many platform-centric business models.

In contrast, bespoke systems specialize in serving a single application end to end. They deliver superior performance and quality because every component is optimized for a specific purpose. However, this approach is expensive and viable only in scenarios where quality requirements and profit margins are both sufficiently high.

In practice, no system is purely generalized or purely specialized. If we view these as two extremes on a spectrum, every application lies somewhere in between, with the majority leaning toward generalization. Competition among companies within the same industry typically enforces a balance between generalization and specialization.

This delicate balance, however, is likely to be disrupted by AI. As discussed in Section A.1, the cost of delivery will decrease by orders of magnitude. Historically, the primary obstacle to bespoke systems has been their cost. If that cost can be reduced dramatically, the rationale for settling for a subpar generalized platform becomes far less compelling.

Admittedly, generalized platform solutions will always be cheaper and will increasingly resemble commodities. Yet when the cost of addressing a business need drops sufficiently low, customers become less sensitive to price and more interested in solutions that offer greater customization, even at a modest premium. Conversely, custom solutions that are currently prohibitively expensive will become accessible, allowing a much broader audience to benefit from specialization.

Even the platform landscape itself will change dramatically. New AI-native platforms will overwhelmingly replace today's existing platforms.

In summary, specialization is poised to become the dominant trend.

A.3 Finetuning model vs context engineering

When applying AI to meet a business need, a natural question arises: should we fine-tune the model? Fine-tuning a pretrained model—often referred to as post-training — typically consists of supervised fine-tuning followed by reinforcement learning. At present, this approach is widely adopted because it is relatively cost-effective. Notably, a leading frontier lab has chosen fine-tuning as its initial product offering, further validating this strategy.

To answer this question, it is helpful to step back and consider the broader landscape.

We begin by examining how AI products are deployed across different market segments. As illustrated by the solid line in Figure 5, the market can be roughly divided into three segments: the head, the body, and the long tail.

- **Head**: This segment comprises the largest markets with the highest profit margins. It is typically the first segment targeted by frontier labs. Given its scale and profitability, it justifies the substantial investment required for full model pretraining and post-training.

- **Body**: This segment includes mid-sized markets whose profit potential does not support the high cost of pretraining. However, fine-tuning pretrained models with domain-specific data can significantly enhance performance. Many efforts in this space focus on adapting open-source pretrained models to specialized domains.

- **Long Tail**: This segment consists of a vast number of small markets that cannot justify even the cost of fine-tuning. Instead, practitioners rely on off-the-shelf models (or APIs from leading models) and apply context engineering techniques to tailor behavior to specific domains.

As this breakdown shows, markets are differentiated by the level of investment made in adapting models. Today, fine-tuning occupies a particularly attractive middle ground: it is significantly less expensive than pretraining while still delivering strong performance on domain-specific tasks. The frontier lab mentioned earlier has further reduced the cost of fine-tuning, enabling it to expand into portions of the long tail.

Why is fine-tuning so effective? A model pretrained on a broad data distribution performs best on tasks drawn from that same distribution. When the target data distribution of a given market differs substantially from the training distribution, performance degrade. Fine-tuning addresses this mismatch

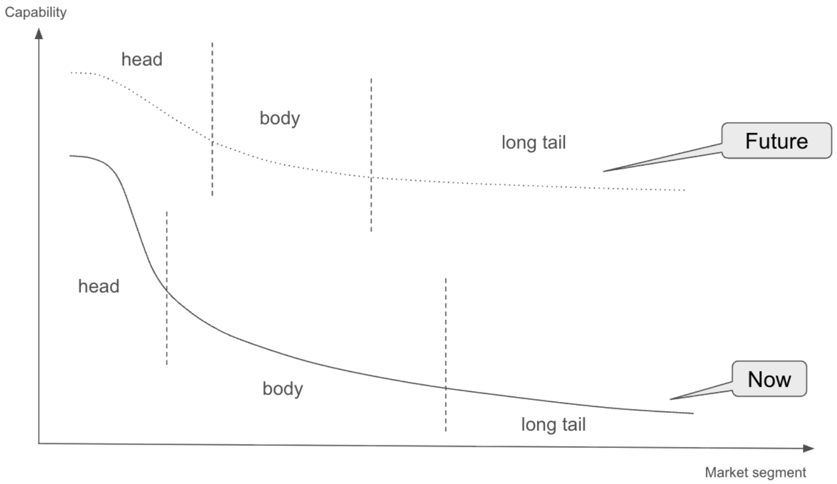

Figure 5: AI capability evolution over time.

by adapting the model to the new data distribution, thereby mitigating the inefficiencies of transfer learning and improving task-specific performance.

However, I argue that as AI continues to advance, the portion of the market that truly requires fine-tuning will shrink substantially, and companies that focus exclusively on fine-tuning will face increasing pressure. This challenge emerges from both the head and the long-tail ends of the market.

From the head side, frontier labs that are capable of large-scale pretraining and mid-training also perform post-training as a core part of their model development process. In fact, post-training is essential for significantly enhancing model capabilities. At present, these labs tend to prioritize the largest and most profitable markets, as their capacity is limited and the returns are most certain. Once these major markets are sufficiently served and the marginal gains from additional fine-tuning diminish, it becomes natural for them to expand into smaller markets. In doing so, they will increasingly encroach upon segments that are currently served by specialized fine-tuning companies.

More specifically, frontier labs aim to build broadly capable, general-purpose models by incorporating as much data as possible across pretraining, mid-training, and post-training stages. However, data availability, evaluation frameworks, and post-training methodologies vary significantly across market segments. Because fine-tuning is non-trivial and labor-intensive, these labs naturally focus first on segments where the data pipelines, evaluation metrics, and economic returns are most mature. Over time, as post-training methodologies in additional segments become validated — often by companies that specialize in those domains — it becomes relatively straightforward to integrate both the data and the methods into the unified training pipeline. This integration expands the effective training distribution of the general model, reducing the need for downstream transfer learning and, consequently, diminishing the demand for further fine-tuning by third parties.

If progress along the scaling laws begins to plateau, the absolute capability of general models may improve more slowly. Nevertheless, their relative performance on long-tail segments is likely to improve faster than on head segments. In this scenario, the solid curve in Figure 5 would flatten into the dotted curve, indicating a reduced capability gap between head and long-tail markets.

As a result, markets that previously relied solely on finetuning would find context engineering could deliver acceptable quality solutions. This shift would expand the addressable long-tail market while simultaneously reducing the relative advantage of fine-tuning.

To be clear, this is not an argument against fine-tuning. Fine-tuning remains a critical component in building world-class models. Rather, the question is whether additional fine-tuning will remain necessary once first-tier frontier labs have already incorporated most of the segment data and methodology into their pretraining, mid-training, and post-training pipelines. As base model quality improves, the incremental need for further fine-tuning is likely to decline.

We have not yet fully observed this effect because the AI market is still rapidly expanding and current model capabilities remain insufficient for many use cases. However, as the market matures and becomes more saturated, this dynamic is likely to become increasingly visible.

That said, some fine-tuning-focused companies will continue to thrive under specific conditions. For example, privacy constraints may prevent certain data from being shared externally, necessitating additional fine-tuning on proprietary or sensitive datasets.