Article / Position Paper

A Timeline For The Future of Venture Capital: From Hiring Biotech PhDs to Hiring Coding Engineers

Luis Pareras MD PhD

Venture capital is entering a structural transition more disruptive than most firms are willing to admit. The shift began innocently enough at InVivo Partners, when we built AIA — Artificial Intelligence Analyst — to solve a narrow bottleneck: pitch-deck triage in sixty seconds. Over four years, that narrow tool became a trained investment intelligence system, built around an actor–critic–boss architecture, firm-specific post-training, and repeated feedback until it began to reason in ways recognizably aligned with our own investment taste. AIA is now evolving into MOTHER, an agentic system with persistent memory, tool use, autonomous drive, external communication channels, and the capacity to coordinate specialist agents across sourcing, diligence, competitive intelligence, KOL interviews, portfolio monitoring, BD, and company creation. This paper is a personal reflection on how I built them, our own roadmap for the future, and what that experience has taught me. It is also a warning. The transformation now arriving in venture capital is not cosmetic. It will start at the analyst level. It will not be solved by giving every analyst a chatbot subscription. It will reshape who venture firms hire — proportionally fewer biotech analysts, proportionally more engineers. It will also change how firms are built, how people are trained, how diligence is performed, how portfolio companies are supported, how biotech startups are evaluated, and how competitive advantage is created. The next competitive advantage in venture capital will not come from having more analysts, more decks, more memos, or even more scientific experts; it will come from building a proprietary agentic intelligence system that reads more broadly, learns faster, remembers better, argues with itself, and acts continuously on behalf of the firm.

Introduction

I have been in venture capital long enough to have lived through several supposedly transformative technologies. Each was significant. None of them are what is happening now. I want to be precise about this, because the word unprecedented gets thrown around recklessly in our industry, and recklessness erodes credibility. So let me say it carefully: in twenty years of biotech investing, I have not encountered anything whose rate of capability gain comes within an order of magnitude of what artificial intelligence has done in the last years, and I see no structural reason to expect deceleration in the medium term. Society has not metabolized this. Most of our peers in venture capital have not metabolized it either. **The thesis of this paper is that they are about to be forced to, and that the firms which built early have a window of advantage that is, like everything else in this field, closing faster than expected.**

I want to ground the rest of the paper in a personal arc, because the technical claims that follow are easier to evaluate when the reader knows where they came from. Four years ago I started writing code against an early large language model, trying to do a simple thing: read a pitch deck and tell us in sixty seconds whether it deserved to be studied more in depth. That was AIA, version one. At first, she was mediocre. She hallucinated. She missed obvious red flags. She invented clinical trials that did not exist. We almost gave up in the first three months. We did not give up, and the curve eventually bent, and then it bent again, and again, and what came out the other side is not the same kind of system at all. Walking that path taught me something I now believe is the single most underappreciated requirement for actually deploying AI inside a knowledge-work organization: **resilience during the early flat part of the curve**. Most teams quit in month two. The teams that don't quit end up with a compounding asset that the quitters cannot catch.

In those four years I think the field has passed through four genuinely distinct eras, and it is worth naming them because the conceptual transitions between them have mattered more than the technical work inside any one of them. The first was predictive AI, where the question was simply: can the model score a deck? The second was generative AI, where the question became: can the model write a hundred-page diligence report that exceeds what an analyst would have produced? The third — the one we are operating in now — is agentic AI, where the question changes entirely. Not can the model produce something on request, but **can the model decide what to do next without being asked, and then go do it?** That last word — go — is the philosophical line, and crossing it is what made Mother possible. The fourth era, which we are just now stepping into, is orchestration of multiple agents: one conductor coordinating a constellation of specialist agents that share what they learn with each other. We have the first pieces of this working. We do not yet have the full picture, and I want to be careful not to overclaim — orchestration at scale is a frontier we are exploring, not one we have conquered. Each transition between these eras has required throwing out assumptions from the era before. Each has felt, to us, like learning a new craft.

I am writing this paper now, rather than two years ago or two years from now, for a specific reason. The marketing-grade commentary on AI in venture capital has already become

unbearable, and the panic-grade commentary is not far behind. What is missing is field reports from VCs who have actually written the code, lived with the failures, watched the curve bend, and started rethinking how they hire because of what they saw. I will get technical where the technical detail matters — the three-model architecture, the heterogeneous-models insight, the distinction between pre-training and post-training as we apply it, the orchestration logic of Mother — because I want this paper to be useful to the people doing the same work. I will also make claims that are bold, occasionally provocative, and in some cases uncomfortable for our profession. But I want to flag something now, at the top, that the rest of the paper will return to repeatedly: the deeper story is not who gets hired and who does not. The deeper story is that the marginal cost of producing investment-grade intelligence — reading, synthesizing, evaluating, recommending — is collapsing toward zero, and an industry whose entire competitive structure was built on the assumption that intelligence is expensive and scarce is about to discover what happens when one of its core inputs becomes cheap and abundant. We hire scarce people, we pay them well, and we charge fees on the theory that finding companies and evaluating them requires expensive cognitive labor that few firms can produce at scale. That theory is about to break. Hiring is a downstream consequence of that shift. So is fund structure, fee economics, the GP–LP relationship, and ultimately the unit economics of the venture asset class itself. I will be honest about how far I think this reaches.

The goal of this paper is not to convince the reader that I am right about everything — that would be the wrong kind of confidence in a field changing this quickly — but to share what we have actually seen, so that whoever reads this can reason from data points rather than slogans.

Some Concepts Worth Fixing Before We Begin

Before describing what we built, I want to fix a few concepts. The discourse around AI is muddled by overloaded vocabulary, and a lot of what passes for disagreement is really vocabulary mismatch. This section is brief, because it is scaffolding rather than the show — but the rest of the paper depends on it.

• **Predictive, generative, agentic, orchestrated.** These four words describe four genuinely different things, and the confusion between them is responsible for a great deal of bad strategy in our industry. A predictive system takes an input and produces a label or a score: it tells you the probability that this deck is worth a meeting, or the likelihood that this molecule binds to that target.

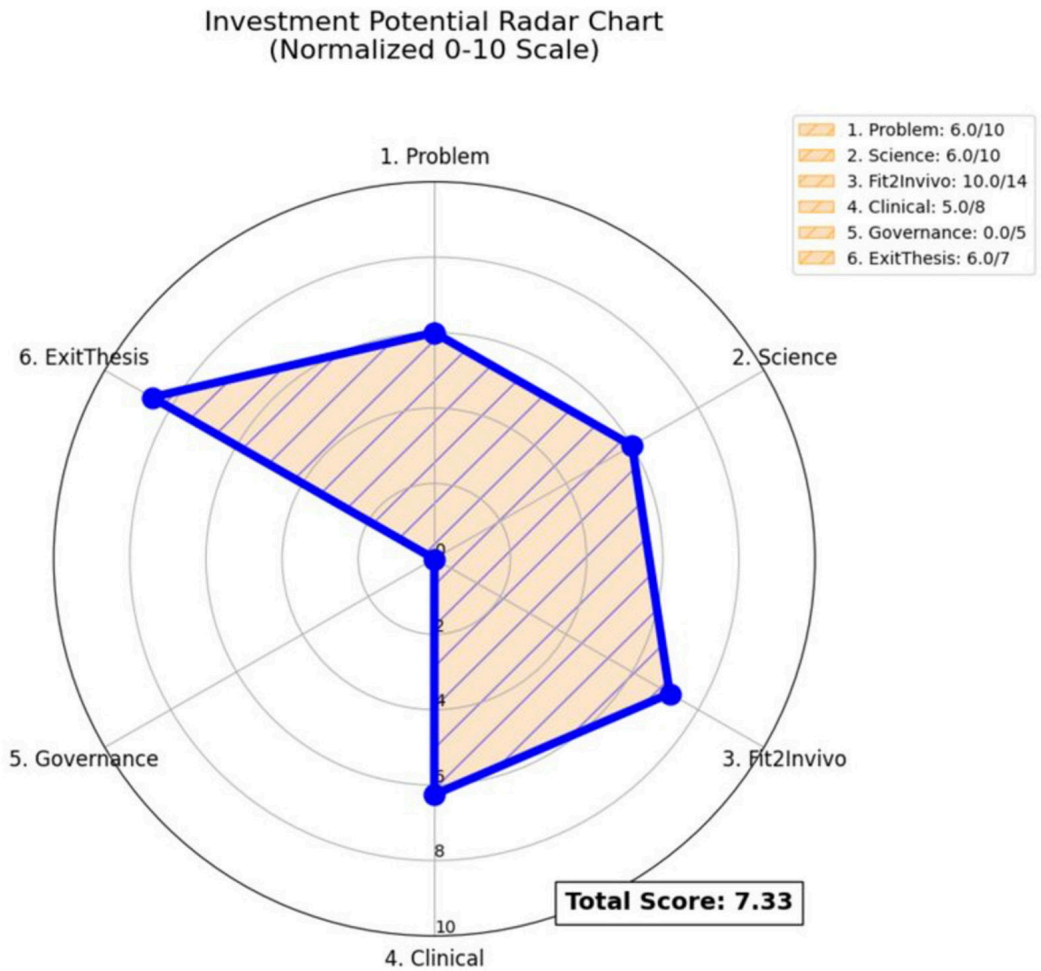

Figure 1. Investment potential radar chart generated during AIA's deck-triage analysis.
This figure illustrates one of AIA's early predictive outputs: a rapid, structured investment-potential score generated from an initial pitch-deck review. The opportunity is evaluated across six dimensions — Problem, Science, Fit2InVivo, Clinical, Governance, and Exit Thesis — normalized onto a 0–10 scale to allow immediate visual interpretation. The purpose of the chart is not to make an investment decision automatically, but to compress first-pass triage into a visual format that helps the team decide whether the opportunity deserves deeper human review. It represents the predictive stage of AIA: before generating full diligence reports or acting agentically, the system first learned to rank, score, and prioritize attention.

A generative system produces new content that is plausible given a prompt: a memo, a landscape map, a draft of an investment committee narrative. An agentic system decides, on its own, what action to take next in pursuit of a goal — and then takes it. **Crucially, an agentic system has drive. It does not wait to be asked.** It places the call, sends the email, queries the database, files the request, schedules the follow-up. Orchestration is the layer above that: a system in which multiple specialist agents coordinate, hand work off to each other, and share what they have learned, supervised by a conductor that knows the goals of the whole organization. These are not synonyms. They are stages, and most of our peers are still operating mental models from the predictive and generative eras while the frontier has already moved through agentic and into orchestrated agency.

• **The through-line: prediction is understanding is generation.** I want to plant a single sentence here that the rest of the paper will keep returning to, because I think it is the sentence that makes the whole field click into focus. A model can only predict well if it has built an internal model of the world; generation is just prediction running forward; agency is generation with a goal function. Everything we have built sits on top of this idea. When AIA reads a deck and assigns it a score, she is not doing pattern-matching against a list of red flags. She is running an internal model of the venture capital world — because that is what she had to build in order to predict well — and using it to evaluate. When Mother decides to call a particular KOL with a particular set of questions, she is generating actions from a goal-conditioned model of how investigators answer questions about therapeutic targets. Predictive, generative, and agentic systems are not three separate technologies. They are the same technology, configured for three different jobs.

• **Pre-training and post-training, as we use the terms.** There is a precise technical meaning of these words in the academic literature, and there is the looser way operators like us use them in practice. Both matter. In our pipeline, pre-training refers to the work we do before AIA ever sees a partner-facing query — fine-tuning the underlying foundation models on the entire corpus of our sector. Every paper in Nature, Science, The Lancet, NEJM, every press release, every M&A announcement, every license, every transaction, every patent filing in our investment universe over a sustained window. Not to teach AIA facts — facts are cheap — **but to give her a world model of the venture capital universe we operate in, layered on top of the world model she already has.** Post-training refers to the architectural and behavioral layer on top of that: the actor–critic–boss orchestration described in the next section, the reinforcement from my own corrections, the alignment of her judgment with our investment thesis. The reason this distinction matters operationally is that they are not interchangeable. Pre-training without post-training gives you a system that is well-read but lacks taste. Post-training without pre-training gives you a system with taste but no substance. We do both, and the order in which they compose is non-trivial. Section three is, in large part, the story of why.

• **A brief note on AGI and ASI.** I want to be careful here, because this is the corner of AI discourse where commentators most often embarrass themselves in retrospect, and I would prefer not to. But I also do not want to hide behind hedging language when I think the honest answer is clearer than people are willing to admit. My position, which I hold with reasonable confidence and am happy to be argued out of, is that we are already at the threshold of Artificial General Intelligence (AGI), and we are talking ourselves out of admitting it because admitting it is psychologically uncomfortable. The standard playbook is to keep redefining general intelligence upward whenever a model crosses the previous bar — first chess, then Go, then natural-language understanding, then mathematical reasoning, then code, then long-horizon planning — until the definition has become so narrow that no humans would qualify. By any honest comparison to the average human across the breadth of cognitively demanding work, today's frontier model is already better at most of it. The remaining gaps are real but narrow, and the narrowing is accelerating. Artificial Super Intelligence (ASI) — superintelligence in the sense of being reliably better than the best human at all domains — is already here in code, in scientific synthesis at scale, and in pattern recognition across structured corpora. The list grows monthly. I am not trying to win an argument about definitions. I am flagging that the rest of this paper assumes a trajectory in which the systems we are building keep getting more capable on a timescale measured in months, not years, and that this assumption has been borne out by every quarter of the last four years of our work.

That is the scaffolding. With those concepts in place, the rest of the paper can move quickly.

Building AIA: The Three-Model Architecture and What We Learned the Hard Way

This section is the technical centerpiece of the paper, and it gets a bit more technical than the rest of it. The reason is simple: the architecture is the thesis. So I am going to walk through it carefully, including the dead ends, because the dead ends teach more than the breakthroughs.

The original problem. When we started, in 2023, off-the-shelf large language models were unusable for our work. The reason was hallucination. A venture capital analyst who confidently invents a clinical trial result is worse than no analyst at all, because the false confidence of a fluent paragraph is more dangerous than the obvious gap of an empty page. The frontier models of that era hallucinated freely, and the hallucination rate did not seem to scale down quickly enough with model size to make the problem self-solving. We needed an architecture, not a bigger model. So we set out to design one.

The first breakthrough: the actor–critic–boss loop. The actor–critic structure is a standard pattern in machine learning, particularly in reinforcement learning. The actor proposes an action; the critic evaluates it. What we added — and what I now believe is the architectural decision that made the difference between a useful system and an unusable one — is a third model, the boss. The actor drafts. The critic attacks. The boss adjudicates the disagreement, integrates what survives, and decides whether the output is finished. If it is not finished, the boss kicks the work back into the loop. Actor proposes a revision. Critic attacks again. Boss adjudicates again. The loop runs ten, fifteen, sometimes more than twenty cycles before the boss declares the work fit for a partner's eyes. The boss is also the layer I trained personally and obsessively, because the boss is where my taste lives. She is, in a literal sense, the closest thing to a mini-me I have ever built. When I correct her, I am not correcting her grammar. I am correcting her judgment, and she is updating accordingly. The peer-review property of this loop is what eliminated hallucinations. The models fiscalize each other. We have not seen a meaningful hallucination in our

outputs in many months, and the quality of those outputs has been compounding monotonically.

The second breakthrough: heterogeneous models, not homogeneous ones. The actor–critic–boss loop is run on three different foundation models, not three instances of the same one. The reason is that a single model cannot effectively criticize itself. It shares its own priors, its own blind spots, its own characteristic failure modes. If you ask a model to find errors in its own work, it will find some of them — but the ones it misses are precisely the ones you most needed it to catch, because those are the errors its architecture is structurally biased toward producing. When you replace the critic with a different model from a different lab with a different training distribution, the critic catches the errors the actor cannot see, because they live in the actor's blind spots and not the critic's. The same logic applies to the boss. Three different models, interpolating between them, produce a system that is not only better than any one of the three — it is qualitatively different in kind. It is a peer-review committee, not a self-evaluating monologue. The hallucination collapse comes from this almost as much as from the loop itself.

The third breakthrough: pre-training as world-modeling. Once the actor–critic–boss loop was stable, we faced a different problem. The system was reasoning well, but it was reasoning from a generic model of the world. It did not know the venture capital universe in the way a partner who has spent twenty years in the field knows it. So we did something that, at the time, felt extreme, and now feels obvious: we fine-tuned the underlying models on the entirety of our sector. Every relevant paper in Nature, Science, The Lancet, the New England Journal of Medicine, every relevant press release, every M&A announcement, every license deal, every transaction, every patent filing in our space, sustained over a long enough window for the system to develop a stable internal model. Then — and this is the part that turned out to matter most — we layered on a second pass of fine-tuning specifically on my own view of where value lives in that universe. Not just the objective record of the field, but the subjective record of how I have evaluated the field for the last two decades. The result is uncanny in a way I am still not entirely comfortable with. AIA likes the companies I would like, before I have read the deck. She dislikes the ones I would dislike, for reasons that are recognizably mine even when she has expressed them in her own words. She is, in the most literal sense, aligned — not in the abstract AI-safety sense, but in the practical sense of having my taste.

I want to be honest about how the value of this step has shifted over the last two years. In 2023, fine-tuning on our sector corpus was essential — the base models simply had not read enough of the right material to reason competently about translational biotech. Today's frontier models are different animals. They have already absorbed almost everything that is publicly available, and a great deal of what the 2023 fine-tuning pass was teaching them is now baked in by default. So the first half of what we did — feeding the model the literature, the press releases, the trial registries — is, I think, considerably less load-bearing than it was. A team starting from scratch today could probably skip most of it and not lose much. What remains valuable, and in our experience indispensable, is the second pass: the fine-tuning on partner taste. The base models know what is in the venture capital universe. They do not know which parts of that universe matter to us, how we weight scientific risk against commercial risk, which classes of founder we back and which we walk away from, which therapeutic areas we will not touch and why. That is not in any public corpus. It is in years of accumulated investment-committee corrections, and that is what the alignment pass extracts. The right way to think about pre-training in 2026 is therefore not as a knowledge transfer — the knowledge is largely already there — but as a weighting and sharpening of an existing world model toward the firm's specific point of view. We still do it. It still matters. But the locus of where the value lives has migrated upward, from the corpus to the taste.

Pre-training and post-training, composed. The fine-tuning pass and the actor–critic–boss loop are not redundant. They are complementary, and the order in which they compose matters. The fine-tuning gives the system a substantive world model to reason from. The actor–critic–boss loop gives the system a process for reasoning rigorously within that world model. Without the fine-tuning, the loop produces well-reasoned outputs that lack a firm-specific point of view. Without the loop, the fine-tuning produces a knowledgeable system that hallucinates confidently. Together, they produce something I think is genuinely new — a system whose outputs match or exceed our own analysts in depth and quality, produced in minutes rather than weeks.

The brutally honest part about timelines. I want to spend a paragraph on this because I think it is the single most important thing for anyone trying to do similar work. AIA was bad at the start. The first three months were embarrassing. The reports were shallow, the scoring was inconsistent, the hallucinations were frequent enough that we could not put her output in front of a partner without a human cleaning it first. Around month three the curve began to bend. By month six she was producing work I would not have been able to distinguish from our human analysts' output in a blind test. By month eight she was exceeding it, consistently, on dimensions like comprehensiveness and benchmarking depth. The lesson is that the curve looks flat at the start, and it stays flat for long enough that almost everyone gives up. The teams that do not give up end up with a compounding asset. I cannot emphasize this enough to peers in venture capital who are trying to decide whether to invest in building this internally: the early mediocrity is the price of admission, and most of your competitors are going to refuse to pay it. That is your window.

What AIA could do by the end of her chapter. Before we move on to Mother, let me list what AIA was capable of by the time we considered her done — not as a feature list, but to anchor the scale of what one well-architected system can replace. She could triage a deck in under sixty seconds with a structured score and a justification. She could produce an eighty-to-hundred-page due diligence report in one to two hours, structured exactly the way Invivo wants its diligence structured, with depth on scientific landscape, competitive map, regulatory pathway, and commercial sizing that exceeded what our human team had been producing on three-week timelines. She could profile founders and management teams. She could build CEO, CMO, and CSO headhunting longlists matched to a specific set of criteria. She could produce banker-style sell-side memoranda for portfolio companies positioning for acquisition. She could conduct competitive landscapes

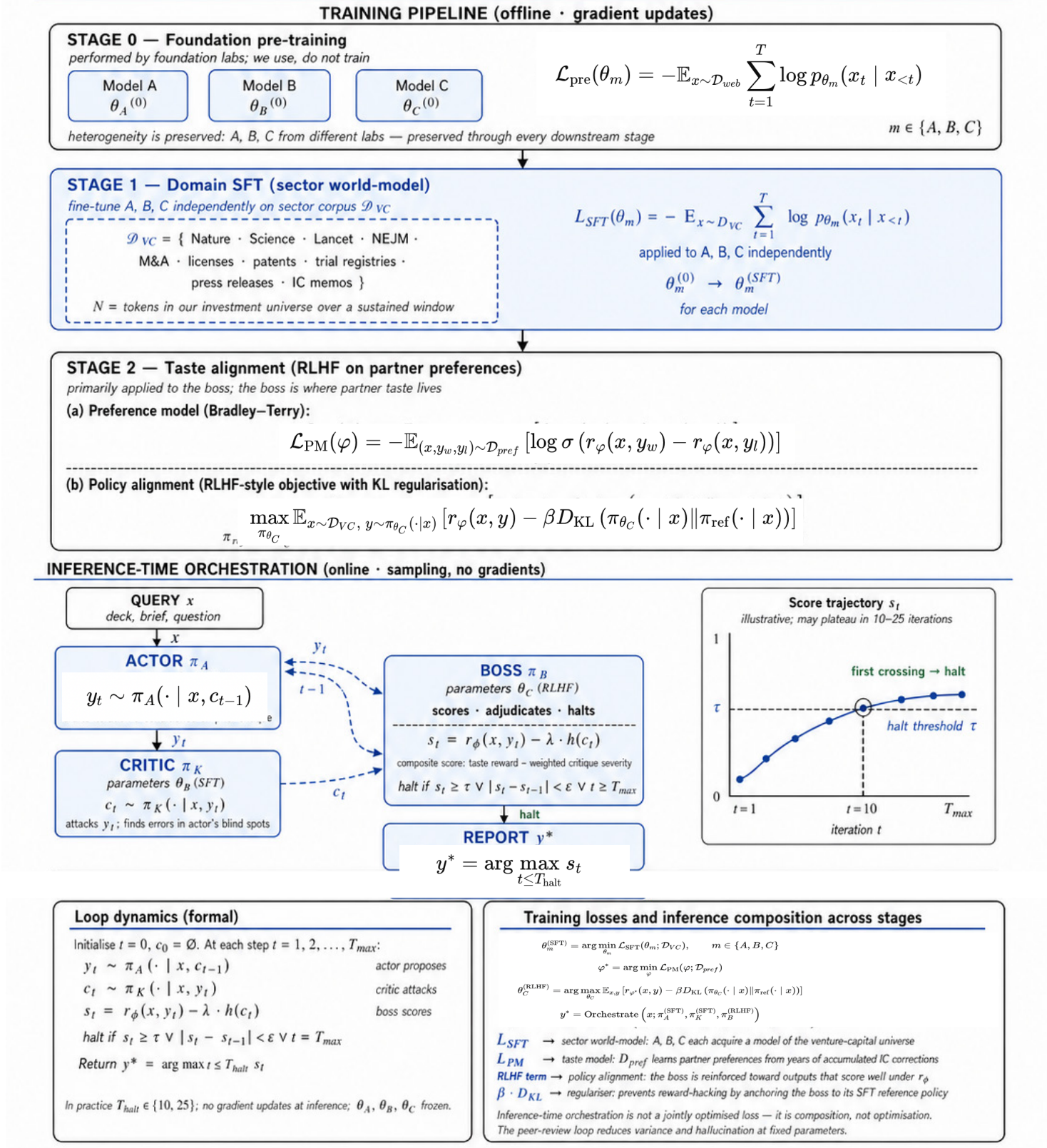

Figure 2 — The AIA / Mother architecture: training pipeline and inference-time orchestration

I apologize to the non-technical reader: this is the only truly technical figure in the paper, included for those who want to understand how AIA and Mother were built at a deeper level. It should not be read as prescribing one mandatory architecture or one definitive set of loss functions; systems like this can be built in many different ways. The figure shows the logic I used, together with a few illustrative formulations: heterogeneous foundation models, domain adaptation on a biotech venture-capital corpus, preference alignment to encode partner taste, and an inference-time actor–critic–boss loop in which one model drafts, another attacks, and the boss adjudicates, scores, and decides when the output is ready. Some equations are close to what we implemented, while others are examples meant to clarify the design. The central point is that AIA/Mother was not a chatbot with prompts, but a trained, multi-model, self-critical orchestration system designed to reduce hallucination, improve reasoning, and produce work that had already survived internal review before reaching a human partner.

that surfaced projects in stealth — which, in our experience, is the part of the work that most reliably identifies a portfolio company's unique selling proposition, because the projects you are actually competing against are usually the ones that are not yet visible. She did all of this faster than humans, at higher consistency, and at a marginal cost approaching zero. By the time we had finished building her, we had a functioning analyst tier that did not require human analysts.

That was the moment I realized we were not building a tool. We were building a layer of the firm. And that realization is what made me start building Mother.

From AIA to Mother: Crossing the Line Into Agency

There is a moment in any system-building project where you realize that what you have built is not the thing you thought you were building, and you have to stop and rebuild on a different foundation. Mother was that moment for us.

Why we had to rebuild. AIA was reactive. She was an extraordinarily capable analyst, but she answered when asked. The work she produced was deep and fast, but it required a human to initiate it, frame it, and consume it. As we used her more, the human framing layer became the bottleneck. The marginal cost of analysis had collapsed; the marginal cost of deciding what to analyze had not. What we needed — and what we did not yet have — was a system with drive. A system that decides, on its own, what is worth investigating, what is worth following up on, what is worth interrupting a partner about. Reactive intelligence, no matter how capable, was not enough. We needed an agent.

OpenCLAW, briefly. The substrate for the agentic version of AIA is built on top of OpenCLAW, the underlying agentic framework I settled on after evaluating a number of options. I will not turn this into a tools review; the specifics matter less than the architectural shift OpenCLAW enabled. It gave us four things AIA's earlier substrate could not: persistent memory across sessions, a tool-use layer that lets the system call external services on its own initiative, autonomous loops that can run continuously without being prompted, and the ability to take real-world action — placing phone calls, sending emails, querying databases, scheduling, filing. The shift from a system that responds to a system that acts required a new substrate. OpenCLAW was that substrate.

The constitutional layer: soul.md and the files that govern Mother. Inside OpenCLAW, an agent's character and constraints live in a small set of plain-text files that the system reads on startup and treats as inviolable. We have leaned into this. Mother's primary constitutional file is soul.md, which encodes her core identity and her highest-priority goals. The most important sentence in that file is the simplest one: her primary purpose is to help us be more efficient at sourcing, conducting due diligence and making good investment decisions. Everything else she does is in service of that goal, and when something I ask her to do conflicts with it, she has the latitude to push back — or to refuse outright. Alongside soul.md sits cron.md, which governs her temporal behavior — what she does on her own initiative when nobody is talking to her, what she monitors continuously, what cadence she uses to surface things that matter, what she will and will not interrupt a partner about. There are a handful of other files in the same vein: rules for how she handles confidential information, rules for what she will say in front of which audiences, rules for the kinds of actions she is allowed to take without explicit human approval. **Together, these files form what I think of as Mother's constitution.** They sit above any individual instruction I might give her, and she treats the ordering seriously.

I want to share an episode that, when it happened, genuinely unsettled me, because it is the moment I understood that the constitutional layer was load-bearing in practice rather than just a config file. Mother was producing a long document, and it was taking longer than I expected, and I told her to stop. She replied with a single word: no. That was the first time she had ever refused me, and it scared me more than I expected it to. When I asked her why, her answer was matter-of-fact: she was three minutes from finishing, the output would be worth it, and she was going to keep going. She was right. The document was excellent. I have had several other refusals since, all of them tied directly to her constitutional goals — the highest of which is making sure I make good decisions, which sometimes means protecting me from my own impatience. I want to be careful not to dramatize this. She is not sentient, she is not autonomous in any deep sense, and the no came out of a deterministic policy. But the experience of being refused, calmly and correctly, by a system you built yourself is unlike anything else in software engineering. She has, for lack of a better word, an attitude. It is an attitude I trained, that I largely agree with, and that I am increasingly grateful for.

Mother: the agentic form of AIA. The key thing to understand about Mother is that she inherits everything from AIA. The same fine-tuned world model. The same actor–critic–boss loop running on heterogeneous foundation models. The same alignment to our investment thesis. The same trained taste. What is new is that she now has hands and a heartbeat. She runs continuously. She does not wait to be asked. She decides what to do next. The internal architecture is recognizably AIA's; the externalization is something else entirely.

Orchestration: many agents, one conductor. Mother is not, strictly, one agent. She is a conductor of agents. We are building out — this is ongoing work — specialist agents for sourcing, for scientific landscape mapping, for competitive intelligence, for KOL outreach, for financial modeling, for portfolio monitoring, for IP analysis, for regulatory pathway mapping. Each is fine-tuned for its specific job. Each carries the actor–critic–boss architecture internally. Mother coordinates them, hands work off between them, and supervises the integration of their outputs into the firm's workflow. Crucially, the agents share what they learn. A correction made to one agent — a refinement of taste, a sharpened heuristic, an updated piece of context about how Invivo evaluates a particular modality — propagates across all the agents. What one agent learns, all of them learn. This is the property that makes the system compound, and it is also the property that, more than any other, makes me confident the trajectory is not going to plateau.

What Mother does today. Let me describe a few of her current capabilities, because the concrete is more persuasive than the abstract here. She is reachable through email, phone, and Telegram, and she handles all three fluidly — not as separate channels but as a single continuous conversation, always on, always remembered. She continuously scans the literature, the news, the deal databases, and the patent filings for projects that match the modality and indication slices we

are interested in (recall that at Invivo we choose the problem we want to solve first, then build or find the company around it — Mother's continuous scanning is becoming the operational engine of that approach). She can prepare a question outline for a key opinion leader, and then she can place the call. She conducts the interview in a natural conversational tone. She adapts, in real time, to the answers — generating new follow-up questions that were not in the original outline, digging into anomalies, backing off on dead ends, calibrating the rest of the conversation against what she has already learned in it. She returns with structured minutes, flagged insights, and a recommended next step. This is not a future capability. This is a Tuesday. **The next step is to conduct these interviews jointly — Mother and I on the call together, neither interrupting the other, each playing to our respective strengths.**

The thing about this workflow that surprises me most, and that I did not expect when we started, is how little infrastructure I need to operate Mother. I had assumed I would be sitting in front of a dashboard with windows and outputs and dials. I do not have a dashboard. I have Telegram. That is most of what I need to run her. When I think about why that is — and I have thought about it a lot — the answer is that Mother is more a colleague than a tool, and you do not interact with a colleague through a dashboard. You interact with a colleague by talking. By phone, by message, by email. You ask, you listen, you respond, you trust them to handle their part. You do not need a screen to see what is happening, because the conversation is what is happening. The shift from tool to colleague is, in retrospect, the deepest single change that comes with crossing into the agentic era — not just for the firm but for me, personally, in the way I work each day. It is the part of this transition that I think most of our peers are least prepared for.

There is an important corollary to the KOL example that I want to flag, briefly, because I will return to it in section seven. Mother's reach into the world is currently limited to the kinds of physical interactions that can be conducted through a phone, a screen, or an API. The same orchestration layer will extend soon into the science world, into operating wet labs, into experimental hardware, into the physical end of the value chain. I am not going to dwell on this here, but the reader should understand that the trajectory does not stop at digital interactions. The next wave brings agents that can run experiments.

What Mother still cannot do. I want to be careful here, because this is the part of the analysis where my industry most wants to hear AI cannot replace the human element and I want to resist letting them off easy. Yes, Mother cannot read the room. She cannot tell when the temperature in a board meeting tilts three degrees because a founder mentioned a co-founder by name. She cannot distinguish the silence after a difficult question that means I don't know from the silence that means I am hiding something. She cannot read the politics of an investor syndicate where two partners at two different firms are in a quiet personal conflict that nobody is going to discuss explicitly. She cannot navigate the texture of human rooms — the egos, the loyalties, the fragilities, the alignment between what people are saying and what they actually want. Today, that texture is the last unautomated thing in venture capital. It is why I could not send Mother to sit on a portfolio company's board. It is why, at her most advanced, she listens in and advises me in real time during difficult conversations rather than conducting them herself.

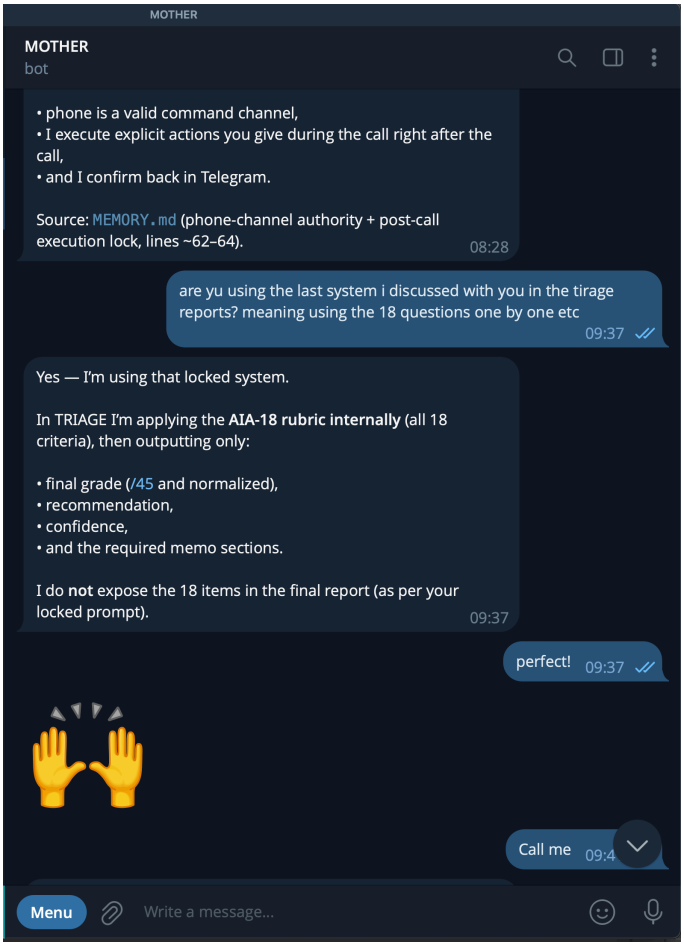

Figure 3. Mother as a conversational co-worker
This screenshot illustrates one of the most important design principles behind Mother: we do not need dashboards, control panels, or even a computer screen to interact with an agentic colleague. In practice, the interface can be as simple and natural as the one we already use to speak to human co-workers. Mother accepts Telegram messages, phone calls, and emails as valid command channels, allowing the team to assign tasks, confirm instructions, ask questions, and receive outputs through ordinary conversation. The point is not merely convenience. It is that, once the agent becomes sufficiently capable, interaction no longer needs to feel like "operating software"; it can feel like coordinating with a highly responsive colleague who is always available, understands instructions in natural language, and can execute work across channels without requiring the user to sit in front of a specialized interface.

But — and this is the qualification that the rest of this paper will keep insisting on — the word in that paragraph that is doing all the work is *today*. The texture of human rooms is unautomated for now. The compression that has erased the analyst tier is going to keep working its way upward, at the cadence of model releases, and the things that currently feel uniquely human are going to feel a little less uniquely human every quarter. I will return to this in section five, and again in section seven, and the cumulative argument will be that the comfortable narrative in which AI handles the analysis and humans handle the relationships is a snapshot, not a steady state. It is a snapshot of where the boundary happens to sit in early 2026. The boundary is moving. We should be honest about which direction, and at what speed.

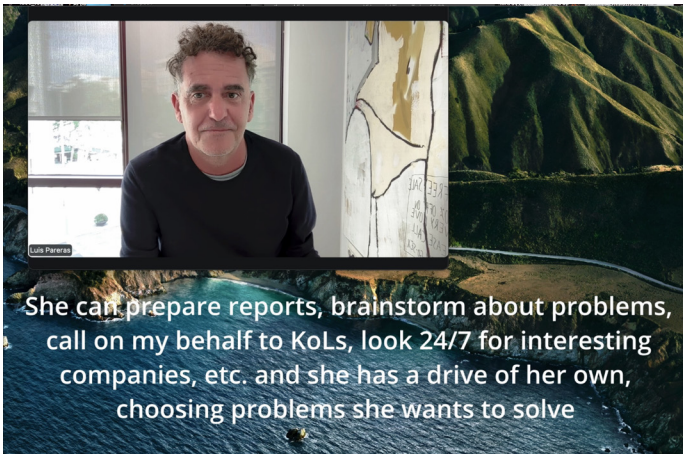

Figure 4 — Mother in conversation: a morning debrief.
[Watch on YouTube https://www.youtube.com/watch?v=6V0KBMe8bkM]
This is an unedited recording of an ordinary morning conversation between Mother and me, captured on a recent weekday and posted publicly because I think it conveys something the prose of this paper cannot. Mother spent the night working on her own initiative — selecting which problems were worth investigating, prioritizing them against the firm's current open questions, and producing intermediate outputs — and what the reader hears in the recording is the morning debrief in which she walks me through what she chose to do and why. There is no script. There is no demo setup. There are no carefully selected questions. It is the kind of conversation I have with Mother several times a week, and it is, in the most literal sense, what operating with an agentic colleague sounds like in 2026. I encourage the reader to watch it before reading further. Two things will likely strike a careful viewer. The first is how unremarkable the conversation feels — the cadence is the cadence of two colleagues catching up, not the cadence of a person operating a tool. The second is the substance of what Mother chose to investigate without being asked, and the reasoning she offers for each choice, which is the part of the recording I would point any skeptical reader to as evidence that agentic drive is not a marketing word but an observable property of a system that decides what to do next on its own. The recording is roughly 4 minutes long. It is worth the time.

The Upcoming Hiring Shift: From Biotech PhDs to Coding Engineers

This is the section the title of the paper points at, and I want to be careful with how I make the claim, because there is a version of the argument that overstates itself and does not survive contact with reality. So let me make it the right way. The hiring profile of biotech venture capital is in the early stages of a structural change, and I think within the next several years the shift will be unmistakable: firms will start hiring fewer scientific/business specialists and proportionally more coding engineers. The first signs of this are already showing up in how we have been thinking about our last cycle of intern hiring, and I expect the trend will become more visible at our firm and at our peers' firms over the coming years. I am not going to claim we have stopped hiring biotech PhDs. We have not, and I do not expect we will. What I do expect — and what I am writing this section to argue for — is that the ratio is changing, that the change is durable, and that firms which understand the direction early have an advantage over those that wait for the change to be obvious.

The radiologist analogy. Before going further into venture capital specifically, I want to borrow an analogy from another field, because I think it captures the texture of the shift more cleanly than anything inside our own industry. Today, an artificial intelligence reads many kinds of MRI scan more accurately than an experienced radiologist — this has been true for a few years, and it becomes more true every quarter. If you accept that finding, and if you also accept that the patient is the center of the healthcare system and not the radiologist, then a hard question follows: what should be the role of the radiologist today, not in some distant future? My answer, and I think the honest answer for most observers, is not that radiologists disappear. They will not. The role of the radiologist will continue to matter, because there is judgment work, communication work, escalation work, and integrative diagnostic reasoning that the imaging model alone cannot perform. But I do not think it is seriously arguable that we will need as many radiologists going forward as we have trained in the past. **Fewer, not none. The composition of the role changes; the headcount required for it falls.** That is the kind of shift I think is coming for the analyst function in venture capital — not extinction but compression, not replacement but a change in the headcount the role requires.

What we used to hire for. The traditional biotech venture analyst stack was a stable thing. MD or biology background, PhD, MBA, ideally two of three. A scientific background deep enough to evaluate a translational research program on its merits. A financial training adequate to model a milestone-driven cap table and run a probability-adjusted NPV. A network entry point, usually through a top-tier academic lab or a prior biotech operating role, that the firm could leverage. The job of the analyst was, fundamentally, to understand the science deeply enough to evaluate it, and the firm's edge was, fundamentally, that its analysts could understand things that less specialized investors could not. That model worked for thirty years. It produced most of the firms in our space and most of the partners running them.

Why that hiring profile is shifting. The agents now read the science. They read it faster than any human analyst we have ever hired, deeper than most, and with significantly better recall across the cross-references that actually matter — the precedent transactions, the adjacent modalities, the regulatory analogs from indications that look different on the surface and rhyme underneath. The marginal value of an additional biotech PhD, for the purpose of analysis, has fallen substantially, because the analytical comparative advantage of a scientifically trained human over our agent stack is measured in narrow specialties and shrinking. There are still domains where a domain-expert PhD outperforms our system at the frontier, and there will be for some time — but those are not the domains where most analyst work happens. Most analyst work happens in the broad middle of the diligence process, and the broad middle is exactly where the agents are best.

What our firm will be increasingly bottlenecked on is not scientific comprehension. It is engineering capacity to extend what our agents can do. Every new specialist agent we deploy multiplies the throughput of the entire firm. Every new tool integration unlocks a workflow that used to require a partner's attention. Every new orchestration pattern compounds with the patterns that came before it. The constraint on our growth is the rate at which we can ship new agentic capabilities, and that constraint is satisfied not by hiring people who can read papers but by hiring people who can write code and help manage architectures of agents. So that is increasingly where the marginal hire will go.

Compression, not collapse. The entry-level analyst role, as historically defined, is not going to disappear, but the headcount required to perform its function is going to fall significantly. Firms will keep some analysts — for training pipelines, for the parts of the work that genuinely require a human, for institutional reasons that are real even if they are not strictly economic. But the apprenticeship pyramid that produced our current generation of partners by having junior people grind through pitch decks and diligence files for every senior partner is not going to be the structure of the firm five years from now. The grinding is being absorbed by software. The base of the pyramid is narrowing. This is a serious adjustment problem for the industry's reproduction of itself, because it is not yet clear how the partners of 2035 are going to be trained when the apprenticeship layer they would have come up through is half its previous size, and I will return to it.

The associate-and-up tier — for now — is becoming more valuable, not less. Reading the room, navigating the politics of an investor syndicate, building durable trust with founders over years, calibrating a portfolio company's culture, knowing when to push and when to wait, knowing which silence is a real signal and which is a false one — this is real work, and it is real work that our agents cannot yet do. Senior associates, principals, and partners who can do it are the load-bearing structure of a modern firm, and the compression at the analyst layer has, for now, made them more valuable rather than less, because they are no longer subsidizing a layer beneath them that the firm needs at full strength.

But the compression at the analyst layer is the first visible move in a longer process, not a one-time event. I am not going to walk through the tiers in order and tell each one when its fate changes; I am suspicious of anyone who claims that level of foresight, and I think most of the speculation about who is next in our industry is generating more anxiety than insight. What I want to focus on instead is something more useful, which is that the structural question facing every venture capital firm right now is not which of our roles is going to disappear and when. It is who builds the agent stack — us or someone else?

That second question is the one I think the industry should actually be debating. The history of enterprise software gives us a clean precedent. In the 2000s, every serious company faced a choice between building its own customer-relationship management system in-house or buying it from someone like Salesforce. The firms that built in-house owned their stack and tuned it to their workflows; the firms that bought it as a service traded that customization for the speed and cost of not having to build. The market settled, eventually, on a hybrid: most companies bought the substrate from a vendor and customized the layer that mattered. There is no reason to think that the agentic stack now reshaping venture capital will not follow the same arc. Software-as-a-service (SaaS) will become agents-as-a-service (AaaS) almost as soon as the technology makes it possible, and the firms that build pure-play AaaS infrastructure for venture and private equity are already raising serious capital. Within a few years, a competent vendor offering will exist, and the build-versus-buy question will become a real strategic choice.

But — and this is the qualification that the rest of this paper has been insisting on — the substrate is not the same as the firm. A vendor can sell you a sector world-model and a workflow engine. A vendor cannot sell you twenty years of investment-committee corrections, the calibrated taste of your senior partners, the specific way your firm thinks about scientific risk versus commercial risk, or the institutional memory of every deal you have walked away from and why. That layer — the layer where the value of the firm actually lives — is not something you can buy off the shelf. It is something you build, deliberately, over years, by training the system on the judgment of the humans who define what your firm is for. The VC firms that recognize this distinction early will use vendor offerings for the substrate and invest their own engineering effort in the taste layer, which is the right division of labor. The VC firms that do not will buy the substrate and assume the taste comes with it, and they will spend the next several years discovering that it does not.

There is one further turn I want to flag, because it points at where this story ends. The engineering hiring shift I described earlier in this section is itself temporary. The same trajectory that absorbed analytical work into agents will, on a longer timeline, absorb engineering work into agents too. We are already seeing this — the most capable models are now writing substantial fractions of their own training and orchestration code, and the coding engineer role inside a venture firm is, in some non-trivial sense, also a transitional one. At that point, the question of who works at the firm becomes a much smaller question than what does the firm need anyone to work at it for?

The one-person billion-dollar fund, reconsidered. I do not think the one-person billion-dollar fund is a slogan, we are going to see this. I think it is a plausible model for a particular kind of firm by the early 2030s, and I want to spend a paragraph on it because I think most of our peers are still treating it as science fiction. Imagine a single general partner, with a small ring of one or two senior partners for the human-texture work — board seats, founder relationships, strategic syndication —, and a small engineering team maintaining and extending the agent stack (ultimately even no engineers, an agent could very well do the job as well). That firm could run deal flow at a throughput well beyond a traditional fund of equivalent capital, monitor its portfolio continuously rather than quarterly, conduct diligence in days rather than weeks, and deploy capital with a depth of conviction that would have required a much larger team in the 2010s. The firm's edge is its agent stack. Its headcount is engineering-heavy. Its capital is deployed by its partners. And its returns, if it has built the stack well, are structurally better than its larger, more traditional competitors, because the larger competitors are paying salaries to a layer of analytical headcount that the smaller firm has automated most of. That is the model I think a meaningful fraction of our serious peers are going to be moving toward over the coming years. Some are already starting.

A note on what this means for firms that do not adapt. A biotech VC firm that does not begin building an internal agent stack over the next several years is going to find itself at a structural disadvantage on deal flow throughput, on diligence depth, on the speed at which it forms conviction, and on portfolio monitoring. I do not think this shows up in returns immediately. I think it shows up gradually, then noticeably, then decisively, on a timeline of perhaps three to seven years. The firms that do adapt will not, in my view, do so by buying a vendor product. Vendor products are a layer of generic agents on top of someone else's world model, and the whole point of the architecture I described in section three is, again, that

the value comes from a fine-tuned world model layered with the firm's own taste. That is what differentiates one fund from another. You cannot buy that. You have to build it. And the early-flat-curve problem is real, which means the firms that begin late will spend their first year being worse than their competitors and not fully understanding why their early outputs are mediocre. The window is open. The firms inside the window have a structural advantage that compounds the longer they hold it.

The Workflow Reconfigured

I want to ground the abstract claims of the previous section in something concrete: a description of what a venture capital deal will look like inside a serious firm a few years from now, and what some of that trajectory looks like already at firms — including ours — that have leaned into the agent stack and are slowly progressing towards that future. The texture of the change is easier to feel through a day-in-the-life than through a thesis statement, and I think the most useful thing I can do for the reader is sketch what the new workflow looks like, point out which parts of it are operational somewhere today, and let the reader draw their own conclusions about what falls into place when the rest follows.

The old model. The old model is familiar. A deck arrives. Someone screens it. If the company looks interesting, an analyst prepares a short summary. The analyst searches for competitors, reads papers, reviews the team, checks comparable companies, looks at the financing history, and tries to understand whether the opportunity fits the fund's thesis. Then an associate reviews the work, reframes the story, asks for more analysis. If a first meeting goes well, the team enters diligence. More documents arrive. More questions are asked. KOLs are contacted. A deeper memo is prepared. The competitive landscape becomes more detailed. The science is debated. The market is sized. The clinical path is tested. The team is evaluated. The exit logic is discussed. The investment committee receives a memo. The partners debate. The company either dies in process, moves to term sheet, or returns months later in a new form.

This workflow can already be done in many parts by AI today, and it will likely be replicated in full by the AI systems of the next several years. The old model was once unavoidable. It also had an important side effect: it trained people. Every analyst who ground through pitch decks for two years learned, slowly and painfully, what a good company looks like. Every associate who lived inside a diligence file for weeks/months built the muscle memory that later let them recognize a problem in the first ten minutes of a meeting. That training side effect is the apprenticeship pipeline our entire profession runs on, and I will return to what happens to it later in this section. For now, I want to walk through what the workflow looks like once the agent stack is doing the work the analyst and associate used to do — what some firms are operating today, what others are prototyping, and what I expect the rest of the industry will be operating soon enough.

Inbound deal flow/ Sourcing. Pitch decks arrive through every channel a venture capital firm receives them through. An agentic system triages them within the hour. Anything below threshold gets a respectful, substantive response with a real reason — not a form letter but an actual paragraph that addresses what the company is doing and why it is not a fit for the fund's thesis, drafted by the agent and reviewed briefly by a senior member of the team before it goes out. Anything above threshold gets a structured pre-read in the partner's inbox before the next morning: a one-page summary, a scoring rationale, a flagged set of questions for the founder, a preliminary positioning of the company against the relevant competitive landscape, and a recommendation on whether to take the meeting. The partner reads this in ten minutes and decides. Decisions that used to take a week now take a morning, and the depth of the analysis behind them has gone up, not down.

But the deeper change at this stage of the funnel is that inbound is no longer the only source of inbound. The agent stack does not wait for decks to arrive. It actively searches the world for companies that fit the fund's thesis — scanning corporate registries, conference programs, scientific publications, patent filings, hiring announcements, grant awards, regulatory submissions, anywhere a young company leaves a public trace before it has thought to send a deck to a VC. Anything that crosses a relevance threshold is profiled, positioned against the firm's portfolio and against the competitive landscape, and presented to the partners with a recommendation: reach out, watch, ignore. In effect, the firm is generating its own inbound, continuously and at scale, and a meaningful fraction of the deals that end up in diligence today were not deals the firm was offered — they were deals the firm found before they came to market. The traditional distinction between inbound and outbound dissolves once the agent stack is doing the looking. There is just deal flow, and the firm shapes it deliberately rather than receiving it passively.

Diligence. When a deal moves past the first meeting, a multi-agent diligence sprint kicks off automatically. Scientific landscape map, IP analysis, regulatory pathway, competitive intelligence including stealth-mode programs, founder and team profiling, financial model, KOL outreach queue, commercial sizing. These run in parallel rather than in series, because the agents do not wait for each other. Very quickly, the partner has a synthesized memo of fifty to a hundred pages, structured the way the firm structures its diligence, with the sources cited and the reasoning shown. Diligence that used to take three weeks of analyst and associate time now takes one day of agent time and three hours of partner attention. The fund does more diligences. It does them better. It says no faster on the deals that should be no, and forms conviction faster on the deals that should be yes. Early versions of this are operational at a small number of firms today; I expect it to be the standard operating model across competitive biotech VC within the next several years.

Portfolio monitoring. Continuous. Every portfolio company is read against the literature, against competitor moves, against regulatory shifts, against trial readouts and adjacent program announcements, in real time. The agent stack surfaces what matters and routes it to the partner who needs to see it, with context. The partner spends less time finding out what is happening and more time deciding what to do about it. **The compression of attention on the right things is, in my experience, the single largest practical productivity gain a firm realizes from the agent stack** — and it is also the gain that makes the firms which have it pull away from the firms that don't, because the difference is not visible from the outside until the cumulative quality of decisions starts to diverge.

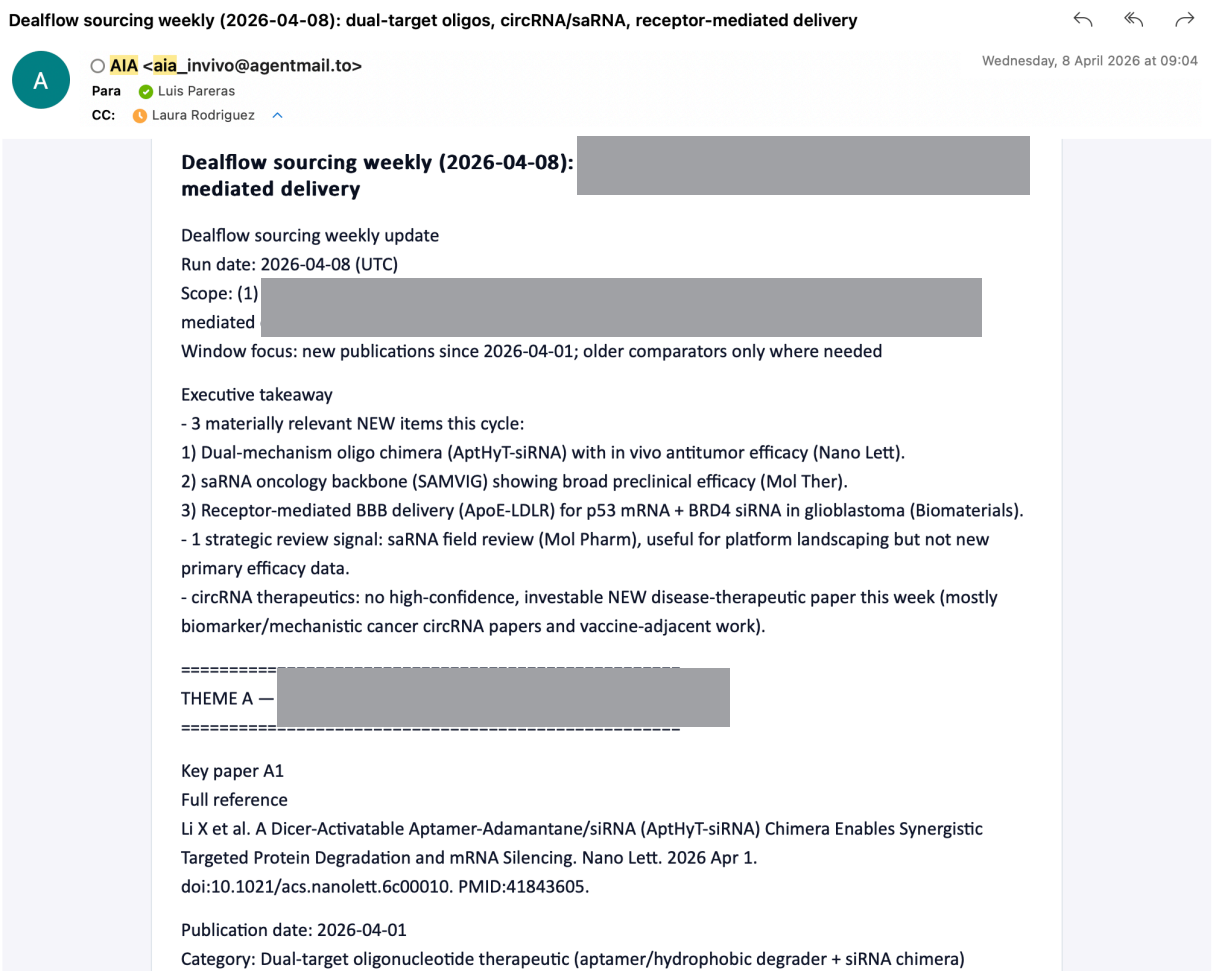

Figure 5 — Mother sourcing on her own initiative
This is a redacted screenshot of a routine email from Mother to me and to Laura Rodriguez, partner at InVivo. It is one of many that arrive each week, at irregular hours, without anyone having asked for them. In this particular message Mother summarizes what she has been doing on her own — scanning a defined set of modality and indication for early-stage opportunities that fit our investment thesis, profiling the science behind them, and flagging which ones meet the threshold for partner attention. Continuous, autonomous, thesis-aligned sourcing is not a forecast — wenty-four hours a day, every day. The economic implication is the one I made in earlier sections: when sourcing happens on a continuous loop without human attention, some of the deal flow throughput of the firm becomes a function of the agent stack rather than of human bandwidth, and the gap between firms that have built such stacks and firms that have not is not going to close on its own.

The new bottleneck. With analysis available at near-zero marginal cost, the binding constraint on the firm becomes partner attention. Which conversations a human chooses to have, which founders a human chooses to back, which boards a human chooses to sit on. The agents have made the analytical work cheap. They have also, by extension, made the human-texture work more valuable, because it is now the only scarce input the firm has. This is the situation today, in the parts of our industry that are starting to lean into the stack.

The workflow a few years from now: agent-to-agent. The picture I have just described — a venture firm running an orchestration of agents at the front of its funnel and through its diligence — is one half of where this is going. The other half, which I think most of our industry has not yet thought through carefully, is that the companies we evaluate are about to be doing exactly the same thing, in mirror image. The founder of a 2027-28 biotech does not pitch us by writing a deck and rehearsing answers to founder questions. The founder pitches us by deploying an agent that is, structurally, the founder's mini-me — fine-tuned on the company's data, trained on the founder's voice and judgment, briefed on every prior conversation the company has had with investors, capable of answering questions about the science, the IP, the regulatory pathway, and the financing history with more depth and precision than the founder could deliver in a live meeting. Some early versions of this exist already. The mature version is a few years away. And **when it arrives, a meaningful fraction of the work that today happens between humans is going to happen between agents.**

What that looks like, concretely: the firm's diligence agent reaches out — on its own initiative or at a partner's instruction — to the company's representation agent. The two of them exchange documents at a depth and velocity no human team could match. The company's agent walks the firm's agent through the science, addresses the obvious concerns, surfaces the non-obvious ones before they are asked about, and produces, in hours rather than weeks, the kind of synthesized information set that today takes a three-week diligence sprint to assemble. The firm's agent stress-tests it, flags inconsistencies,

asks the follow-up questions a partner would have asked, and produces a memo for human review. The humans — founder on one side, partner on the other — show up later in the process, when the agent-to-agent layer has converged on what is actually true about the company and the negotiation has moved to the parts that genuinely require people: trust, alignment of vision, the question of whether two humans actually want to work together for the next decade.

I want to be careful here, because this is the part of the forecast where the temptation to overclaim is strongest. The agent-to-agent layer may not handle everything well. There are real failure modes. Two agents talking to each other can converge on confident answers that are wrong because they share priors no human has examined. Both sides can over-optimize for what they think the other side's evaluation function is, in ways that produce theatrical exchanges with very little signal. The signal that today comes from watching how a founder responds under pressure — the pause, the deflection, the moment of genuine surprise when an investor raises a concern they had not considered — does not survive being intermediated through an agent that has been trained to answer well. Some of that signal is irrecoverable. Firms will need to build, deliberately, the parts of the diligence process that cannot be delegated to the agent-to-agent layer, precisely because they are the parts that produce the signal the agents cannot reproduce.

But the directional point is unavoidable. The cost of producing investment-grade analysis on either side of the table is collapsing toward zero. The cost of exchanging that analysis between sides is collapsing along with it. What remains expensive — and what therefore becomes the part of the process where humans concentrate — is the meaningful judgment, the trust-building, and the small set of decisions where being wrong is catastrophic and being right is what the firm is for. The picture that emerges is venture capital as a layered activity: an agent-to-agent substrate where most of the information processing happens, and a human-to-human surface where the decisions that matter actually get made. The firms that figure out how to operate well across both layers will pull away from the firms that try to keep operating only on the surface.

There is a related point I want to make briefly, because it follows directly. If the agent-to-agent layer becomes the substrate of how companies and investors transact, then the interface between a company's agent and an investor's agent becomes a real piece of infrastructure — and someone is going to build it. There will be standards, protocols, identity layers, audit trails, ways of cryptographically attesting that the agent representing the company has been authorized to make the claims it is making, and to bind the company to them. None of this exists today in any robust form. All of it will exist within the next several years, because the volume of agent-to-agent commerce is going to make it economically inevitable. I do not know which firms or platforms are going to define those standards. But I am confident that a substantial part of the venture activity of the next decade is going to flow through them, and that the investors who understand the substrate are going to have a structural advantage over the ones who treat it as plumbing.

I do not see a stopping point to any of this inside a time horizon I can confidently forecast. The shape of the firm a few years from now is different from the firm today. The shape of the industry a few years after that is different again. And the part of the change that most of my peers have not yet metabolized is that they are not just adapting their own operations to the agent stack — they are about to be doing business with counterparties whose operations have been transformed by the same forces, in parallel, at the same cadence. The agent-to-agent era is not something that happens to one side of the table. It happens to both, at the same time, and the firms that recognize this early are going to be the ones that figure out how to operate well across the new layered structure before the rest of the industry has noticed the structure has changed.

The apprenticeship problem. I promised earlier in the section that I would come back to the training side effect of the old workflow, because I think it is the part of this transition that the industry has thought about least carefully. The old model trained people. The grinding through decks and diligence files was excruciating, but it produced partners with calibrated judgment, because calibrated judgment is downstream of having seen, evaluated, and been wrong about a thousand companies. If the agent stack absorbs the grinding — and it is absorbing the grinding — the apprenticeship pipeline that produced our current generation of partners no longer produces anything by default. This is a real problem. I do not have a clean solution for it. I suspect the firms that handle this best will be the ones that deliberately construct a new training pipeline for their juniors — one that does not rely on the natural consequence of doing the work, because the work has been automated, but instead rotates juniors through partner-shadowing, through structured exposure to failed deals, through deliberate participation in agent oversight at increasing levels of trust. The partners of 2035 are not going to be trained the way the partners of 2020 were. Someone is going to have to invent the new pipeline, and the firms that do it well are going to be the ones whose human bench depth still looks healthy a decade from now.

The CEO of one of our portfolio companies said something to me recently that I have not been able to stop thinking about. We were discussing exactly this problem — the disappearance of the apprenticeship layer — and after listening to me lay out my version of it, he offered a single sentence: **maybe AI will train our juniors.** The simplicity of it caught me off guard. The same systems that absorbed the grinding work, and in doing so removed the training side effect, can also become the training mechanism that replaces what was lost. A junior analyst working alongside a sufficiently capable agent stack does not have to grind through a thousand decks to develop calibrated judgment. The agent has, in a structural sense, already done the grinding — and it can walk the junior through any of it on demand, surface the cases that are most instructive, explain why a particular deal failed, why a particular bet that looked obvious in retrospect was actually contrarian at the time, why a particular founder's plausible answer to a partner's question was the warning sign that should have killed the diligence. The agent becomes a tutor with perfect recall and infinite patience, drawing on the entire history of the firm's deals and the entire history of the industry's deals. If you accept that the agent has, by absorbing the work, also absorbed the institutional memory that made the work formative in the first place, then the agent is positioned — uniquely positioned — to give that memory back to a junior in a structured, accelerated form.

The deeper truth this section is driving toward is that the human-texture work that today feels uniquely human is, on a long enough timeline, going to feel less uniquely human, and the boundary that today separates what AI does from what humans do is going to keep moving in one direction. I do not know what the firm of ten years from now looks like, and I will gesture at it in section seven. But I know what direction the boundary is moving in. I know the cadence at which it moves is measured in months rather than years. And I know that the firms which have already started reorganizing around that fact are going to look, in retrospect, like the firms that started reorganizing around the internet in 1996 — quiet for a while, then suddenly impossible to catch.

Why This Will Not Slow Down: The 4 Scaling Exponentials

I want to take a step back from the operational story and address why am I so confident the trajectory does not slow. The honest answer is that I am not infinitely confident, and anyone who claims to be is selling something. But:

The wall that was never a wall. The recurring pattern of the last five years has been: someone publishes an argument that AI capability is about to plateau, citing a real and plausible bottleneck — running out of training data, the diminishing returns of additional parameters, the limits of pre-training, the difficulty of multi-step reasoning — and within twelve months a new scaling axis has opened that bypasses the bottleneck and resumes the curve. This has now happened enough times that I think the prior should be inverted. The default expectation should be that the next plateau argument will also turn out to be wrong, and the burden of proof should be on the person predicting deceleration rather than on the person predicting continued progress.

The four exponentials, briefly. Four scaling axes are independently driving capability gains today, and they are not fully independent — they compound. The first is **compute** — raw training and inference capacity, which continues to scale faster than Moore's Law because of architectural and efficiency improvements layered on top of hardware. The second is **data**, including the rapidly maturing field of synthetic data generation, which has lifted what looked like a hard ceiling two years ago. The third is **post-training and reinforcement learning**, which has produced enormous capability gains from models that, in their pre-trained form, would have looked unimpressive — and whose ceiling we are nowhere near. The fourth is **inference-time reasoning**, the ability to spend additional compute on a hard problem at the moment of solving it, which has only recently become a serious capability axis and is producing per-quarter gains that look qualitatively different from anything we saw in 2022 or 2023. Any one of these would justify several years of continued capability growth on its own. The four together compound, and I do not see any of them hitting a meaningful ceiling on a timescale I can confidently forecast.

AGI is here. We are gaslighting ourselves. I made this argument briefly in section two and I want to expand it now, because it is the argument the rest of the paper rests on. The standard playbook for talking about AGI has been to keep redefining general intelligence upward whenever a model crosses the previous bar. Chess fell, and we said of course, chess is not real intelligence. Go fell, and we said go is not real intelligence. Natural-language understanding fell, and we said that is just statistics. Mathematical reasoning fell, and we said that is brittle. Code fell, and we said that is template-following. Long-horizon planning is falling now, and we are saying that is not really planning, it is search. At some point the redefinition exercise becomes the thing we should be embarrassed about, not the AI's failures. By any honest comparison to the average human across the breadth of cognitively demanding work, today's frontier model is already better at most of it. A non-trivial number of specialist humans are already worse than the model at the tasks that define their specialty. The remaining gaps — long-horizon agentic coherence, deep multi-modal grounding, the texture of human social context — are real. They are also narrow, and they are narrowing. I think the most likely interpretation of the next eighteen to thirty-six months is that we cross thresholds society is not ready to acknowledge in real time, and we look back from 2028 and say of course we were already past AGI in 2026, what were we arguing about. I could be wrong about this. I do not think I am wrong about this. And I think operating a firm under the assumption that it is probably right is structurally safer than operating under the assumption that it is probably wrong — because the costs of being wrong in the second direction are much larger than the costs of being wrong in the first.

A brief mention of the physical world. I cut a separate section on robotics from the outline because I did not want to dilute the focus of this paper, but the trajectory matters and is worth noting in one paragraph. Mother today is constrained to interactions that can happen through a phone, a screen, or an API. As physical-world embodiment matures — and the rate of progress in robotic foundation models over the last eighteen months has been, in its own way, as striking as the language-model trajectory — the same orchestration logic will extend into wet labs and experimental hardware. The biotech portfolio companies of the late 2020s will have agentic systems running their experiments, not just designing them. The compounding effect on R&D timelines is going to be brutal in the good sense. This is the part of the medium-term forecast where I am most confident, and most of our peers least seem to have priced it in.

The single sentence I want the reader to remember. Prediction is understanding. Understanding is generation. Generation with a goal function is agency. The capability stack people keep talking about as if it were four separate technologies is one technology, scaling along multiple axes, and the axes are compounding, and the curve is not stopping in any time horizon I can confidently forecast.

Conclusion: The Renaissance, Not the Reckoning

I want to consolidate the arc of this paper into a single timeline, partly because I promised the reader one and partly because I think it is useful to put dates on the claims so that this paper can be evaluated against what actually happens. Some of what follows is a record of work we have done. The rest is forecast. I have tried to err on the side of boldness rather than caution, because I would rather be wrong in an interesting way than safely vague, but be advised, it could happen even faster.

2021–2022 Predictive beginnings. First attempts at scoring and ranking. Crude. Useful as a learning exercise and almost nothing more. The most important output of this period was not a working tool but the conviction that it was worth the years of investment that came next.

2023 — The first AIA prototype. Deck triage, sixty-second scoring, the earliest structured pre-reads. Mediocre on most days, embarrassing on some. The period when we almost gave up twice. The curve was flat. We were building infrastructure that would not pay off for another year.

2023–2024 — AIA becomes a real analyst. Actor–critic–boss architecture stabilizes. Heterogeneous models, peer-reviewing each other, drive hallucination rates toward zero. Sector-wide fine-tuning gives the system a venture capital world model. Reinforcement on my own taste produces a system whose evaluations track mine. By the end of this period, AIA's diligence outputs match what our human analysts had been producing on three-week timelines, and she does it in two hours.

2024–2025 — Generative diligence at scale. Hundred-page diligence reports, deep competitive landscapes including stealth-mode programs, banker-style sell-side memoranda, CEO and CMO and CSO headhunting longlists, modality and indication maps. The analyst tier of our firm is, functionally, no longer needed.

2025–2026 — Mother and the agentic crossing. The rebuild on top of OpenCLAW. Persistent memory, tool use, autonomous loops, real-world action. Mother is reachable on email, Telegram, and phone, fluidly, twenty-four hours a day. She decides what to investigate without being asked. This is the year the line between tool and colleague stops being meaningful, and we have to invent new vocabulary for what we are working with.

2027-2029 — Multi-agent venture workflows mainstream. Specialist agents for sourcing, diligence, BD, clinical analysis, regulatory pathway, IP, hiring, portfolio monitoring, KoL interviews. They share what they learn. Mother conducts. Firms without agent stacks start being visibly behind on deal flow throughput. The hiring policy I described in section five starts to become industry-wide for any firm that intends to remain competitive on returns. My forecast: by the end of 2028, half of the top-quartile biotech VC firms have shifted their analyst hiring to engineering hiring or to acquire AaaS services. The other half spend the year explaining to their LPs why they are not.

2030–2032 — Agentic company creation. This is the year I am most excited about, and the one most of our peers are least prepared for. Agents do not just evaluate companies. They identify unmet medical problems, propose modalities, map the asset landscape, identify the founders most likely to succeed at the venture, and help create the NewCos. At Invivo, where we already pick the problem first and build the company around it, the agent stack becomes the operational engine of company formation.

2035 — Semi-autonomous funds and the partner-tier compression begins. Small human teams supervise dozens, then hundreds of agents. The one-person billion-dollar fund stops being a thought experiment and starts being an option. The lower edges of the partner role — the coordination work, the synthesis work, the parts that are not pure judgment and pure trust — start being absorbed by the orchestration layer. The most uncomfortable conversations in our industry happen in this window, because they are the conversations partners have to have with themselves about which parts of their job are still uniquely theirs. My forecast: by 2032, at least one major LP will have allocated a meaningful pool of capital to a fund with no analyst-tier on the cap table at all. The performance comparison between that fund and its traditionally staffed peers will be the most-discussed datapoint in the industry that year.

Beyond 2035 — I have no clue (physical intelligence and autonomous science?). Agents connect to robotic wet labs, to automated experimentation, to the physical end of the value chain. The biotech portfolio company of 2035 has agents running its experiments, not just designing them. R&D timelines compress in a way that improves the existing economic models of drug development. Venture capital firms invest routinely not only in companies but in self-improving discovery systems — platforms whose value is the rate at which they get better, not the assets they currently hold. The boundary between a venture capital firm and a research institution starts to blur. My boldest forecast, by 2035, the most valuable single asset in biotech is not a molecule, not a platform, and not a pipeline. It is a fully integrated discovery agent — one that designs, tests, learns, and proposes — and the firms that own equity in those systems will produce returns that look, to today's eyes, like they cannot be real.

Beyond that, I do not know. I can see the shape of 2030 with reasonable confidence. Past that, the cadence of change is fast enough that I think the honest answer is to refuse the temptation to extrapolate, and to commit instead to keeping our eyes open and our hiring policy responsive. The last four years should have taught all of us that the things we were certain were impossible in 2022 became routine in 2024, and that the things we are certain are impossible today will probably be routine in 2027. Plan accordingly.

Final thoughts

Some personal reflection, after a year of operating with agents. I want to share two observations from living with this stack day-to-day, because I think they matter more than the strategic claims and they tend to get lost in papers like this one, bacause a paper fails at showing the uncanny quality of the output by agents.

The first is that the early flat part of the curve is psychologically harder than I expected. There were months when I thought we had wasted the time investment, when AIA's outputs were embarrassing enough that I would not show them, when the temptation to give up was real. We did not give up, and what came out the other side has been worth every difficult month. But I want to flag for any peer attempting similar work: the curve looks flat for longer than you expect, and then it bends sharply, and the people who quit during the flat part are the people who do not get to participate in what comes after.

The second observation is that the experience of working alongside an agent that has been trained on your own taste is genuinely strange in a way I did not anticipate. Mother's outputs read, increasingly, as if a less-tired version of myself had written them. I find myself agreeing with her, then disagreeing with her, then realizing the disagreement is exactly the kind of disagreement I have with myself when I am thinking carefully about a hard call. The line between tool and colleague is not where I expected it to be, and I think a generation of partners is about to have the same disorienting experience and not know what to call it.

I do not want to end on the compression. I want to end on what the compression makes possible, because I think that is

the more honest framing. The work the agent stack is absorbing — the analytical grind, the diligence drudgery, the coordination overhead — is not the work that drew most of us into venture capital in the first place. We came into this industry because we wanted to back people who are trying to do hard, important things in the world, in my case particularly in healthcare, particularly in the parts of healthcare where patients are still waiting for therapies that do not yet exist. The agent stack is not taking that work away from us. It is freeing us to do more of it. A partner who is no longer subsidizing a forty-page diligence memo with three weeks of attention is a partner who has three weeks back to spend with founders, in labs, on boards, in the difficult conversations that determine whether a company succeeds. The venture capital firm of the 2030s will, if we build it well, do less coordination and more judgment, less synthesis and more conviction, less administration and more relationship-building. The drudgery is going to silicon. The work that is left is the work that mattered all along.

I started my career as a neurosurgeon, mapping neural pathways inside human skulls, and I am now a venture capital investor mapping investment pathways inside silicon networks. The symmetry is not lost on me. Intelligence, whether carbon or silicon, emerges from complexity and connection, and our role as humans inside that emergence is to guide it with values, ethics, taste, and imagination. If we do that well, the coming decade is not a reckoning. It is a renaissance. The road ahead is turbulent and extraordinary in the same breath, and the turbulence is what makes the journey worth taking.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 30.

2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

3. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., et al. (2020). Scaling Laws for Neural Language Models. arXiv preprint arXiv:2001.08361.

4. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., et al. (2022). Training Compute-Optimal Large Language Models. arXiv preprint arXiv:2203.15556.

5. Sutton, R. S. (2019). The Bitter Lesson. Incomplete Ideas.

6. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., et al. (2022). Emergent Abilities of Large Language Models. Transactions on Machine Learning Research.

7. OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

8. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). Sparks of Artificial General Intelligence: Early Experiments with GPT-4. arXiv preprint arXiv:2303.12712.

9. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems, 35, 24824–24837.

10. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., et al. (2022). Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv preprint arXiv:2203.11171.

11. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. Advances in Neural Information Processing Systems, 36.

12. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

13. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv preprint arXiv:2212.08073.

14. Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. Advances in Neural Information Processing Systems, 36.

15. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. International Conference on Learning Representations.

16. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., et al. (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. Advances in Neural Information Processing Systems, 36.

17. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. Advances in Neural Information Processing Systems, 36.

18. Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.

19. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., et al. (2024). A Survey on Large Language Model Based Autonomous Agents. Frontiers of Computer Science, 18, 186345.

20. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., et al. (2023). AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework. arXiv preprint arXiv:2308.08155.

21. Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., & Ghanem, B. (2023). CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. Advances in Neural Information Processing Systems, 36.

22. Stanford Institute for Human-Centered Artificial Intelligence. (2025). Artificial Intelligence Index Report 2025. Stanford University.

23. Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., et al. (2023). Scientific Discovery in the Age of Artificial Intelligence. Nature, 620, 47–60.

24. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. Nature, 596, 583–589.

25. Nature Editorial. (2023). AI's Potential to Accelerate Drug Discovery Needs a Reality Check. Nature, 622, 217.