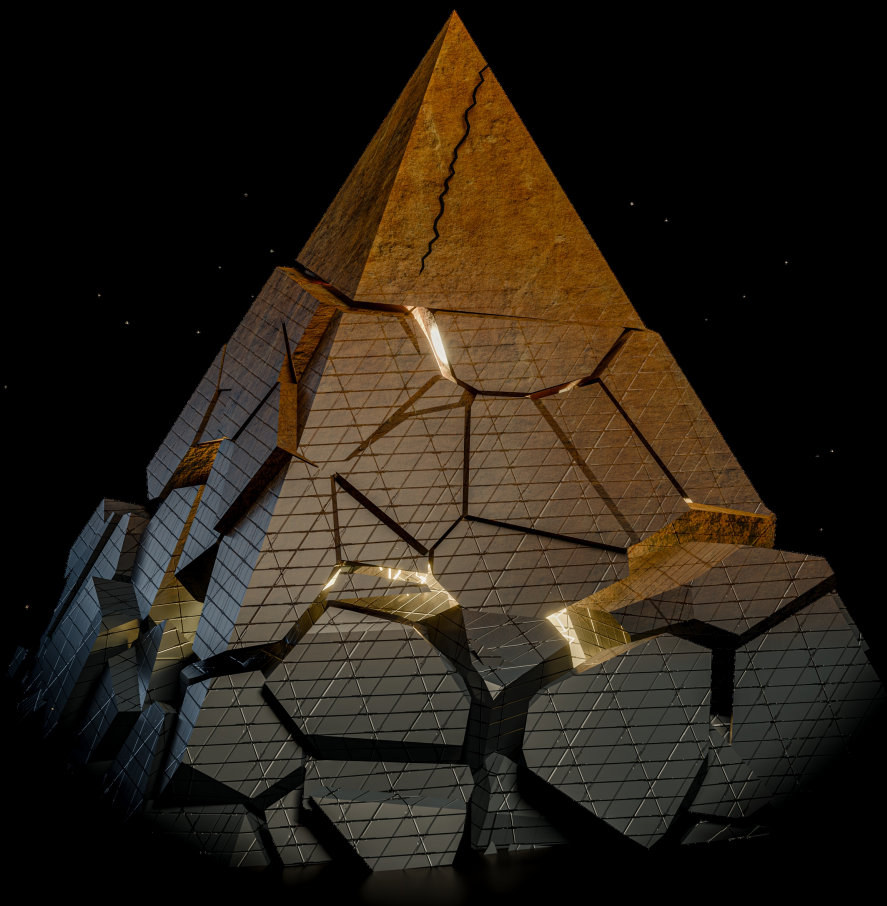
BUILDING AI-NATIVE PROFESSIONAL SERVICES FIRMS

Strategy, Economics, and Execution



DANIEL M. KATZ · MICHAEL J. BOMMARITO II · JILLIAN BOMMARITO

Chapter 4

Craft, Process, and Scale

Chapter 4 · April 2026

✓ Prologue · Four Lawyers, One Monday Morning

Part I · The Opportunity

✓ Chapter 1 · The AI-Native Professional Services Firm
✓ Chapter 2 · The Economics of Transformation
✓ Chapter 3 · Uncertainty and Structure as Strategy

Part II · The Transformation

› **Chapter 4 · Craft, Process, and Scale** YOU ARE HERE
– Chapter 5 · Thin Wrappers and Deep Systems
– Chapter 6 · Land, Expand, Compound
– Chapter 7 · The Org Chart No One Expected
– Chapter 8 · Growing Pains: Talent and Culture
– Chapter 9 · Measure Twice, Cut Once

Part III · The Execution

– Chapter 10 · Paths—Build, Transform, Combine
– Chapter 11 · Capital and Investment
– Chapter 12 · Execution

Part IV · The Value

– Chapter 13 · Value Creation and Capture
– Chapter 14 · The Future

Chapter 4: Craft, Process, and Scale

We are what we repeatedly do. Excellence, then, is not an act, but a habit.

— Aristotle (as paraphrased by Will Durant)

Sarah landed at Phoenix Sky Harbor on a Tuesday evening in the late spring of 2025, the desert heat hitting her the moment she stepped off the jet bridge. After weeks in the Bay Area—the mild mornings, the redwood-scented hike with Alex, the polished coffee shops where term sheets were negotiated over $7 lattes—Phoenix felt raw by comparison. Hotter. Less forgiving. More real. She had spent the trip closing investors, but it was Rachel's parting words from the hike that would not leave her alone: *Find lawyers who will refactor how they do the work.* Sarah had the capital commitments. She did not yet have the operating system. Life as an associate was easy by comparison—running an "AI Native" law firn was serious all-encompassing work.

The round was committed. Handshakes all around. Alex Hawthorne's anchor had attracted the other investors exactly the way he said it would, and the term sheets were signed. But committed was not closed, and closed was not funded. The lawyers were still papering the deal—Alex's counsel in Palo Alto negotiating protective provisions with her own attorney in Phoenix, both sides marking up documents as though three million dollars in seed capital required

the same diligence as a leveraged buyout. The money would come. It had not come yet.

On the drive from the airport, Sarah passed the exit for her old firm's office tower. The glass building caught the last of the sunset, glowing amber. She had left that building not that many months ago with another associate, a laptop, and an unshakeable conviction that AI would reshape legal services. She had, to a minimal extent at least, proven the concept—a handful of clients, promising unit economics, enough traction to convince Alex and the seed investors. However, she wanted to move things beyond mere conceptual validation. She wanted a firm that could perform work at scale. And waiting for lawyers to finish papering a seed round was not building. But she tried to put that out of her mind.

Now the real work begins. Those were Maya's words. Sarah had understood them intellectually at the time. Sitting in her car in the airport parking garage, she understood them viscerally. The capital was coming—she believed that—but she could not afford to wait for it. Every week spent watching lawyers redline subscription agreements was a week not spent building the platform, hiring the team, acquiring the clients, and proving the unit economics that would justify a Series A at a materially higher valuation.

Bright and early, she pulled out of the garage the next morning and headed toward downtown Phoenix. Her apartment was less than ten minutes from the office. She would be there by 6:40am the first morning back. There was no time to waste.

4.1 It is Hard to Automate Chaos

The first two weeks back were a blur of hiring conversations, technology discussions, and client pitches. Sarah's three-person team—herself, Elena, who had followed her from the old firm, and Joshua Thornton, whom she had recruited away from a well-regarded Phoenix firm with the promise of building something entirely new—had been running on instinct and adrenaline. They experimented with a number of the frontier models such as Claude, GPT, Gemini and a patchwork of other AI tools to handle work that would have required

twice the headcount at a traditional firm. The AI results were promising albeit far from perfect. But the process was total chaos.

Sarah knew it. She just did not know how to fix it.

It had been quite a few weeks since their initial hike when Alex called late on a Thursday afternoon. "How are things progressing with my new favorite law firm?"

"Honestly? We are delivering good work, but we are flying by the seat of our pants. Every matter is a little different. Every workflow is improvised. Joshua handles intake one way, Elena handles it another. I review everything myself because I do not trust any process I do not personally oversee."

"That is exactly what I expected to hear and it is okay. Remember I was a lawyer once upon a time." Alex paused. "Okay but I want you to meet someone. His name is David Park. He runs operations consulting for a healthcare system in Phoenix—process engineering, Six Sigma, Lean methodology. He has zero legal background but he has turned three hospital networks from operational disasters into models of efficiency. Before getting into healthcare, he cut his teeth as an engineer at Honeywell."

"Alex, I need lawyers. I need developers. I am not sure I need a—"

"You need David more than you need another lawyer. Trust me on this. I have seen a dozen startups with brilliant founders and no operating system. Half of them are dead. The other half wish they were."

Sarah hesitated. She trusted Alex's judgment—he had been right about everything so far. But the idea of bringing in someone from healthcare to tell lawyers how to practice law felt wrong in a way she could not quite articulate. Ultimately, Alex had a reputation for providing more than just capital and this was the advice that at some level she knew she probably needed.

"One meeting," Alex said. "Coffee. If you do not see the value, I will never bring it up again."

"Sure, let's have the meeting," said Sarah.

Of course, Alex was not being completely truthful. He was going to press her if she resisted. He had money on the line and a reputation to uphold with the

other investors. He invested in Sarah not really because she had it all figured out—she did not. He invested in her because he thought she had the right mixture of drive, curiosity, and humility to put the pieces together. In both his time as a lawyer and as an investor he had seen that the wrong type of ego was an enterprise killer. Success was not guaranteed here but Sarah passed an important test in his mind by taking the meeting. Alex was intrigued to see where things might go from there.

David Park arrived at Sarah's office on a Monday morning at 7:45, fifteen minutes before their scheduled meeting. Sarah was unavoidably late as an impromptu client call pushed her back until about ten minutes past eight. David was unfazed; from the minute he arrived he was walking around the office assessing and actively taking notes.

David was in his early fifties, compact and precise in his movements, wearing khakis and a button-down shirt with the sleeves rolled to the elbows. He had a pen behind one ear and a Moleskine notebook open on the conference table. His handshake was firm but brief—the handshake of someone who did not waste time on ceremony.

"Sarah. Thanks for making time." He gestured at the office around them—the open floor plan, the whiteboards covered in scrawled notes, the stack of client files on a side table. "Alex gave me some background on what you are building. Before we talk, I spent the last twenty-five minutes observing. May I share what I see?"

Sarah raised an eyebrow. "You have been here twenty-five minutes."

"Twenty-five minutes is enough for a first-pass assessment." He flipped open his notebook. "You have no standard operating procedures. None. I watched your associate—Joshua?—take a client call, open a new matter, and begin work without referencing any documented process. He made three decisions about how to structure the intake based on what appeared to be personal judgment. Your other associate walked in, saw the same type of matter on her desk, and handled the intake differently."

Sarah felt a flash of defensiveness. "We are a small startup which is less than a year old. We have been focused on delivering quality work, not writing manuals."

"I understand. And I am sure that your team is very strong lawyers who care about their work and their clients. I know all of you are moving very fast everyday." David set down his pencil. "But you are running a craft workshop, not a scalable business. Every deliverable is artisanal. Every matter is a snowflake. You cannot scale artisanal. And you certainly cannot build AI-native workflows on top of processes that do not exist."

The words stung because they were true. Sarah had spent months thinking about AI capabilities, investor economics, and regulatory structures. She had not spent equivalent time thinking about how work actually moved through her firm.

"Alex said you turned around hospital networks," she said. "How does that translate to a law firm?"

David leaned forward. "I bet you a hospital emergency department and a law firm have more in common than you might think. Both handle complex knowledge work. Both rely on highly trained professionals making judgment calls under time pressure. Both have traditionally resisted standardization because practitioners believe every case is unique." He paused. "And both suffer from the same operational disease: process variation that masquerades as professional judgment."

From Toyota to Process Driven Modern Legal

The principles David brought to Sarah's firm trace a lineage far older than most people realize. The intellectual origins reach back to Adam Smith, whose description of pin manufacturing in *The Wealth of Nations* (1776) first demonstrated that dividing complex work into standardized, repeatable steps could multiply output by orders of magnitude. The resistance of craftsmen who believed their skill made standardization unnecessary would echo through every industry for the next two and a half centuries. More recently, W. Edwards Deming pioneered the modern application of

these ideas, bringing statistical process control and continuous improvement to Japanese manufacturing in the 1950s. Deming's insight that quality was not inspected into a product but built into a process became the intellectual foundation for everything that followed. The Toyota Production System, developed by Taiichi Ohno and Eiji Toyoda with Deming's principles as a catalyst, demonstrated that eliminating waste and standardizing repeatable processes did not reduce quality—it elevated it. Motorola formalized a complementary approach in the 1980s with Six Sigma, and General Electric's adoption under Jack Welch proved the principles worked beyond the factory floor. Healthcare followed: Virginia Mason Medical Center in Seattle modeled its improvement system directly on Toyota's, sending physicians to Japan to study production lines.

Law's engagement with process engineering began earlier than most practitioners realize. In 2001, Thomas Sager, DuPont's general counsel, launched the DuPont Legal Model, consolidating the company's outside counsel from 350 firms to 38 and demanding that those remaining adopt process metrics and efficiency standards (Sager 2009). The initiative saved DuPont hundreds of millions in legal spend and sent a signal that sophisticated clients would no longer accept the billable hour's perverse incentives without question.

The most visible law firm response came from Seyfarth Shaw. In 2005, under the leadership of J. Stephen Poor and Lisa Damon, the firm launched SeyfarthLean—the first comprehensive adoption of Lean Six Sigma principles by a major law firm (Poor and Damon 2008). Seyfarth brought in consultants from manufacturing, trained hundreds of lawyers in process mapping, and eventually developed more than 500 documented process maps for common legal workflows. The results were concrete: flat-fee mergers and acquisitions work delivered at roughly 15 percent below market rates, with margins that exceeded the firm's hourly-rate engagements. Harvard Business School wrote it up as a case study (Bower and Kaplan 2011). Other firms noticed but few followed.

Catherine Alman MacDonagh founded the Legal Lean Sigma Institute in 2007, building a certification program that trained legal professionals in process improvement methodology (MacDonagh 2014). By 2012, firms including Clifford Chance had launched their own continuous improvement programs, with Clifford Chance reporting annual savings exceeding $1.5 million from process redesign alone. Yet the movement remained a niche pursuit. Most firms treated Lean and Six Sigma as interesting experiments rather than existential necessities—the billable hour still worked, clients still paid, and the urgency to change had not yet arrived. David's contribution to Sarah's firm was recognizing that AI had finally created that urgency.

4.2 The Path to Muda is Paved with Good Intentions

"When I walked into my first hospital, surgeons told me that every patient was different and therefore every surgical workflow had to be different. They were half right. Every patient is different. But the prep process, the instrument layout, the post-op checklist, the handoff protocol—those can and should be standardized. Standardizing the repeatable parts freed the surgeons to focus their judgment on the parts that actually required it."

"We are lawyers, not a factory," Sarah said. The words came out before she could stop them, and she heard the echo of every partner at her old firm who had resisted change with exactly the same refrain.

David smiled. It was not a condescending smile—more like the smile of someone who had heard this objection a hundred times and had a ready answer.

"I understand this is a different way of thinking about the work you are doing. Toyota does not make worse cars because they have a production system. They make better ones. The Toyota Production System is as much or more about quality as efficiency. When every bolt is tightened to spec, every weld is inspected, every component meets tolerance, the engineer can focus on the design choices

that actually differentiate the vehicle." He met her eyes. "When your intake is standardized, your AI processing follows defined workflows, your review has clear criteria, and your delivery meets documented specifications, your lawyers can focus on the judgment that actually differentiates your firm. Right now, they are spending a significant amount of their cognitive energy on logistics."

Sarah thought about her own day. How much time she spent deciding how to structure a new matter versus actually analyzing the legal issues. How much time Joshua spent reinventing the intake process for each client. How much time Elena spent formatting deliverables because there was no template.

"Okay," she said. "I am listening."

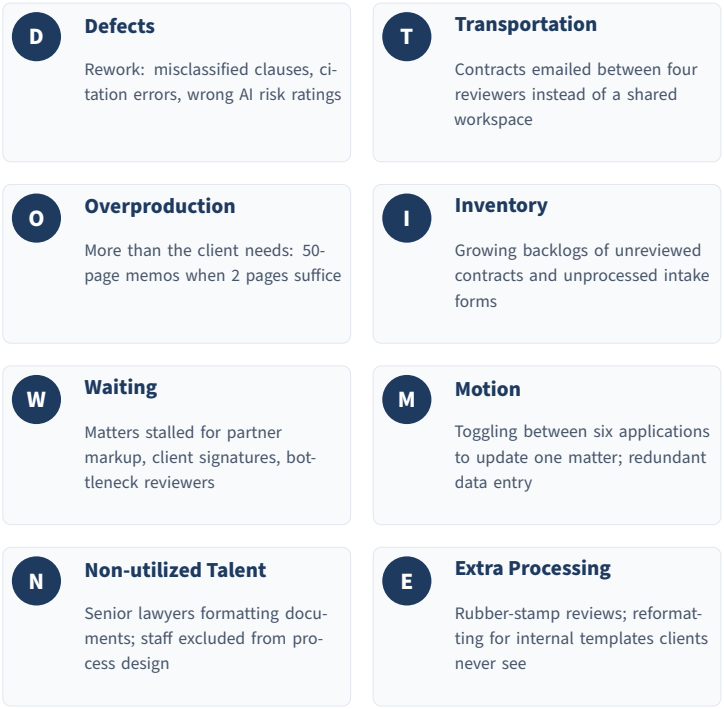


Figure 4.1: The eight categories of waste (Muda) from Lean methodology, applied to legal and knowledge work. The mnemonic DOWNTIME identifies these eight wastes in any process.

David did not ease into things. He showed up the next morning with a roll of butcher paper, a group of sticky notes, a pack of colored markers, and a request that stopped all three lawyers in their tracks.

"I need you to map every service you deliver. End to end. From the moment a client contacts you to the moment you send the final deliverable. Every step, every decision point, every handoff."

Joshua looked at Sarah. Elena looked at Joshua. Sarah looked at the butcher paper.

"All of them?" Joshua asked.

"Start with the ones that generate the most revenue."

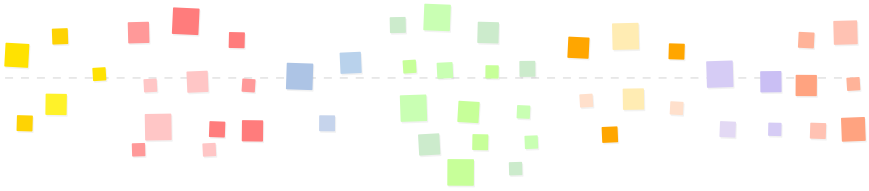


Figure 4.2: One process mapped by Team Candor during the impromptu workshop run by David Park.

Sarah thought about their current portfolio. Contract review was the largest—it accounted for roughly 40 percent of the firm's billings. Regulatory compliance analysis was second. Entity formations and corporate filings were third. A handful of advisory engagements filled out the rest.

"Contract review," she said. "That is our bread and butter."

David taped the butcher paper to the conference room wall and uncapped a blue marker. "Walk me through it. A client calls and says they need 450 assorted contracts reviewed. What happens next?"

What followed was one of the most uncomfortable exercises Sarah had experienced since law school. As she, Joshua, and Elena described their contract review process, David mapped it on the wall with sticky notes and arrows. The gaps became visible in a way they never had been when the process lived only in their heads.

The intake was inconsistent. Sometimes Sarah handled it, sometimes Joshua did. They asked different questions. They captured different information. They made different assumptions about scope.

The AI processing was ad hoc. They used Claude for first-pass analysis, but the prompts varied by matter and by lawyer. Joshua had developed his own prompt library. Elena had a different one. Sarah used a third approach entirely. The outputs were formatted differently depending on who set up the workflow.

The human review had no defined criteria. It was just vibes. Each lawyer reviewed AI output based on their own sense of what mattered. There was no checklist, no sampling protocol, no systematic way to catch the kinds of errors that AI systems made predictably.

The quality assurance was Sarah. She reviewed everything personally. Every deliverable passed through her before it went to a client. This was not a quality system—it was a bottleneck wearing a quality system's clothing.

The delivery was manual. Elena formatted the final reports. Joshua handled the client communication. Sarah did the follow-up. No templates. No standard format. No defined timeline.

David stood back from the wall when they were done. "Listen," David said. "This may not feel like it matters as much now but as you scale this is going to be the backbone of a winning organization." The butcher paper was covered with arrows, boxes, question marks, and a disturbing number of spots where he had written "varies" or "ad hoc" or "depends on who."

"This," he said, pointing at the entire map, "is not a production system. It is three talented people doing their best to deliver quality work without any infrastructure to support them. You are succeeding despite your process, not because of it."

The room was quiet. Sarah felt the truth of it settle into her chest.

"The good news," David continued, "is that you have the raw material. Your people are excellent. Your AI tools are powerful. Your clients are satisfied—for now. But if you try to scale this to even three times the volume, you will break. Not because the AI is not currently good enough or the lawyers are not smart

enough, but because the system connecting them is held together with duct tape and good intentions."

"So what do we do?" Elena asked.

David picked up a red marker. "We build the production system."

4.3 Standardize, Supplement, Specialize, Stop

In his first weeks, David introduced a framework that would reshape how Sarah thought about every service her firm offered. He called it portfolio analysis, but the concept was deceptively simple: before you can build workflows, you have to decide which services deserve workflows in the first place.

"Not everything you do should be done the same way," David said. They were in the conference room, the butcher paper from the mapping exercise still on the wall, now joined by several additional sheets. "And some things you are doing should not be done at all."

He drew a two-by-two matrix on a fresh sheet. On the vertical axis he wrote "Volume" and on the horizontal axis he wrote "Standardization Potential."

"Every service your firm offers falls into one of four categories. I call them the four S's."

He wrote in each quadrant.

The 4S Framework

Standardize: High-volume, repeatable services where AI handles 70–90% of production. Fixed-fee, per-unit pricing.

Supplement: Moderate-volume services where AI assists but human judgment is essential at multiple points.

Specialize: Low-volume, high-complexity work where the value is almost entirely human judgment. Value-based pricing.

Stop: Services that do not fit the firm's strategic position. Refer out or discontinue.

"Standardize. These are your high-volume services with clear, repeatable processes. Contract review. Entity formations. Regulatory filings. These should be productized—defined scope, defined process, defined output, defined price. This is where your AI investment pays off the most."

"Supplement. These are moderate-volume services where AI can handle a significant portion of the work but human judgment is still essential at multiple points. Compliance assessments. M&A due diligence synthesis. Litigation document review with strategic analysis. AI supplements the lawyer; it does not replace her. Think roughly 60 percent AI, 40 percent human—but the human 40 percent is where the real value lives."

"Specialize. These are low-volume, high-complexity services where the value is almost entirely human judgment. The bet-the-company advice. The novel regulatory question. The negotiation strategy. AI might help with research or drafting, but the core deliverable is your brain. Price these on value, not volume."

"Stop. These are services that do not fit your strategic position. Maybe the volume is too low to justify AI investment. Maybe the margins are terrible. Maybe someone else does it better. Stop doing them."

Sarah studied the matrix. "This feels like triage."

"It is exactly triage. You have limited capital, limited people, and limited time. You cannot build world-class AI workflows for every service simultaneously. So

	Standardization Potential	
	High	**Low**
Volume: High	**STANDARDIZE** AI-native production Per-unit pricing	**SUPPLEMENT** AI-assisted delivery Enhanced fixed fees
Volume: Low	**STOP** Insufficient ROI Refer or discontinue	**SPECIALIZE** Human-led expertise Value-based pricing

Table 4.1: The 4S Framework: Service classification by volume and standardization potential

you prioritize." David tapped the "Standardize" quadrant. "This is where you start. Pick one service, productize it completely, prove the model works, then apply what you learn to the next one."

"Contracts," Sarah said immediately. "All things involving contracts."

"Why?"

"Highest volume. Most consistent inputs—we are usually looking at similar document types with similar clause structures. Clear criteria for what constitutes a risk or deviation. And it is the service where our AI already performs best."

David nodded. "Good. Now let us talk about what productizing actually means."

Productizing contract review turned out to be the hardest thing Sarah had done since leaving her old firm. Not because the concept was complicated, but because it required a kind of discipline that lawyers were not trained for.

David started with the intake. "What information do you need from a client before you can begin a contract review engagement?"

Sarah rattled off a list: the contracts themselves, the client's standard positions on key terms, any specific concerns or focus areas, the timeline, the desired output format.

"Good. Now write that down as a formal intake form. Every field required. Every option defined. No ambiguity."

"We do not want to make it feel bureaucratic—"

"You want to make it feel professional. Right now your intake process communicates that you are winging it. A defined intake form communicates that you have done this a thousand times and know exactly what you need. Which message do you want your clients receiving?"

He had a point. Sarah spent an afternoon with Elena building the intake form. It was more difficult than she expected. Every field required a decision about what information was truly necessary versus what was nice to have. Every option required a definition. By the time they finished, the form was two pages long and captured everything the AI system would need to begin processing—client information, document specifications, review parameters, risk tolerance thresholds, output preferences, and timeline commitments.

Next came the AI processing workflow. David's core principle was simple: AI first, human second. Every workflow should begin with AI processing before humans engage. This reversed the traditional model where humans did the work and AI assisted at the margins. In the AI-native model, AI produced first-pass work and humans reviewed, refined, and added judgment. He also insisted on designing explicit paths for the 5 to 10 percent of matters that would not fit standard processes—exceptions that, if ignored, would become dangerous.

"Walk me through what the AI models do when they receive a batch of contracts," he said.

"It is all about the data model," Sarah said, and walked him through the pipeline in detail. The system ingested the documents, classified them by type, extracted key terms—parties, dates, obligations, termination provisions, indemnification clauses, limitation of liability, governing law—then compared each contract against the client's standard positions and flagged deviations. It assigned a confidence score to each extraction and each risk flag. The outputs were structured into a standardized report format.

"Good," David said. "Now, what happens when the confidence score is low?"

"The lawyer reviews it manually."

"What is 'low'? What is the threshold?"

Sarah paused. "We... have not defined a specific threshold. We sort of eyeball it."

David wrote on the whiteboard: *No undefined thresholds. Every decision point needs explicit criteria.* "Listen, I am not an AI expert but I know that as a general matter the more narrow the task the less surface area for errors, human or machine. We need serial decomposition of tasks into smaller and smaller units—this is going to be key."

They spent a full day calibrating the confidence threshold. Too low, and the AI would flag too few items, letting errors through. Too high, and the AI would flag everything, defeating the purpose of automation. They settled on 92 percent after reviewing historical data from their first seven months of engagements. At that threshold, the AI's flagged items captured 98 percent of genuine issues while keeping the false positive rate manageable.

What mattered was not the number itself but the act of measurement. Traditional firms never set thresholds because they never measured error rates—operating under the unstated assumption that human review was error-free. It was not. But firms like Sarah's old one were prisoners to a convenient fantasy: that because machines make errors, human-led processes must be superior. Nobody had ever bothered to check.

Candor would be different. Where Sarah's old firm treated quality as an article of faith, David treated it as an engineering problem. "I am not interested in vibes," he said. "I am interested in science. You cannot improve what you do not measure, and you cannot measure what you have not defined." That willingness—to quantify what traditional firms left to assumption—was itself a source of competitive advantage. He pushed the team to define and formalize exactly what the reviewing lawyer was looking for.

"It is not enough to say 'review AI output for issues.' That is what you would tell a summer associate, and you know what happens—they either review

everything with equal intensity or they skim the whole thing and miss the important parts." David was relentless on this point. "Define the review criteria. What specific things is the human checking? What is the decision at each checkpoint?"

They built a review protocol with explicit checkpoints. The reviewer verified that the AI had correctly identified all parties and key dates. The reviewer assessed flagged deviations against the client's risk tolerance. The reviewer evaluated any items where the AI's confidence was below threshold. The reviewer checked for cross-reference integrity—clauses that interacted with each other in ways the AI might miss. And the reviewer applied judgment to ambiguous provisions where the AI had identified a potential issue but could not determine its significance.

"Now quality assurance," David said. "How do you know the review was done correctly?"

"I review the reviewer's work."

"That does not scale. What happens when you have ten reviewers?"

Sarah did not have a good answer. David built one. A sampling protocol where a senior lawyer—initially Sarah, eventually others—reviewed a random sample of completed matters. The sample rate started high, at 25 percent, and would decline as the process matured and error rates dropped. Every error caught in sampling was logged, categorized, and fed back into both the AI system and the reviewer training process.

"This is a feedback loop," David said, drawing a circle on the whiteboard. "The AI produces output. The reviewer catches errors. The errors ultimately improve the overall AI led process. The improved AI produces better output. The reviewer catches fewer errors. The sample rate drops. The system gets more efficient over time. But only if you capture the data. From what I am reading about the Future of AI, some of the fixes are going to come from LLMs and some will come from the application of traditional AI/NLP. Eventually, it will be a layer cake of agents, subagents looking, checking in various ways slowly building higher fidelity outputs."

Finally, delivery. David insisted on standardized output formats, defined delivery timelines, and a client communication protocol.

"Your client should know exactly what they are getting, when they are getting it, and what it will look like," he said. "No surprises. No variation. The deliverable from your firm should be as consistent as a product from a factory—because it is a product from a factory. A very sophisticated law factory staffed by AI systems and expert lawyers, but a factory nonetheless."

Sarah felt a twinge of resistance. She had become a lawyer to exercise judgment, not to run a factory. But she could see the results taking shape. The mapped workflow was clear, logical, and—she had to admit—far superior to the ad hoc chaos they had been running.

Building the production system consumed the rest of the summer. David had initially estimated a few weeks for a fully industrialized contract review workflow. It took many more weeks than anticipated.

The intake form went through four revisions before clients stopped calling with clarifying questions about the fields. The confidence threshold that felt right at 92 percent had to be recalibrated twice after edge cases—procurement agreements with nested amendments, multi-party frameworks with cascading obligations—revealed that certain contract types needed different baselines. The review protocol broke down during a mid-August test run when Joshua and Elena discovered they had different interpretations of how to handle a flagged deviation the AI had scored at 91.5 percent, just below threshold. David documented the disagreement, revised the protocol with an explicit decision rule, and made them run the test again.

There were weeks when the system felt close to ready and weeks when a new failure mode set them back. Sarah grew impatient in July, then resigned herself to the pace in August, then felt something shift in September—not perfection, David never used that word, but stability. The workflows held. The quality metrics converged. The dashboards told a consistent story instead of a different one every morning.

By September, David declared the system version one-point-zero. Not because nothing remained to improve—his list of refinements ran to three pages—

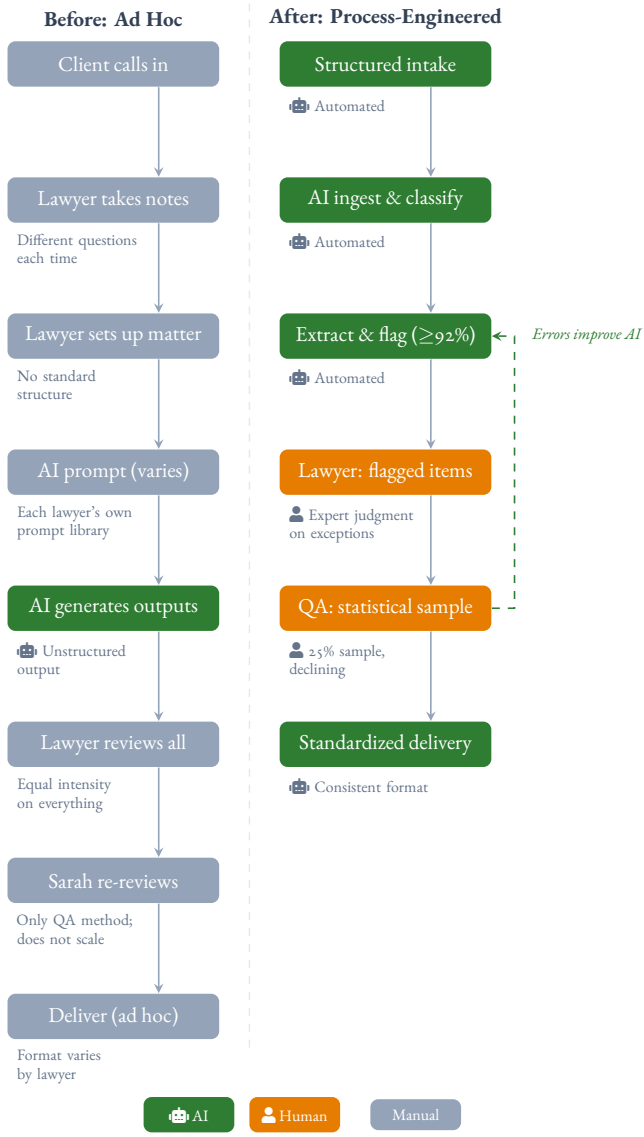


Figure 4.3: Contract review process: before and after process engineering. The ad hoc workflow requires human involvement at every step. The engineered workflow automates repeatable tasks and focuses human judgment on exceptions, with a feedback loop that improves the AI system over time.

but because the production system was stable enough to begin to trust with a real engagement at full scale.

4.4 The Need for Speed

The first real test of the production system came that September. A mid-market technology company needed 350 assorted contracts reviewed before a financing round. The complexity of the agreements varied but there were quite a few complex agreements in the set. Their existing law firm had quoted $105,000 and a six-week timeline. Sarah quoted $52,500—$150 per contract—and committed to a two-week turnaround.

The general counsel, a woman named Linda Torres, was skeptical. "How is that possible? We have used Brennan and Associates for years. They are a good firm. They quoted us $300 per contract." She glanced at the proposal cover page. "Candor. That is an unusual name for a law firm."

"We have productized the service," Sarah said. She walked Linda through the workflow: standardized intake, AI-powered analysis, human review of flagged items, quality assurance through sampling, structured deliverables. "Same outcome, different production model. Our AI system handles the volume work—reading every contract, extracting key terms, identifying deviations from your standard positions. Our lawyers focus on the items that require human judgment. The result is faster, more consistent, and significantly less expensive."

"What if the AI misses something?"

"Our confidence threshold is set at 92 percent, which means any extraction or risk flag below that level gets mandatory human review. On top of that, a senior lawyer samples 25 percent of completed reviews. And we track every error we find, which feeds back into improving the system." Sarah paused. "I will be direct with you. No review process—human or AI—catches everything. But our system is designed to measure its own performance and improve. Traditional firms do not measure it because they do not have to."

Linda was quiet for a moment. "You are telling me your process is more reliable because you actually track reliability."

"That is the theory. This will be our first full-scale engagement with the new production system. I want to be honest about that."

Linda studied her for a long moment. "I appreciate the 'candor.' Let's try it. For us—honestly—it is not even really about the money. We like spending less, of course, but our leadership has told us repeatedly that the legal department needs to move at the speed of the business. If you can deliver quality results on a significantly faster timeline, then yes, you have my attention."

She signed the engagement letter that afternoon.

The 350-contract engagement became the proving ground for everything David had helped the team build. It proved, above all, that version one-point-zero was version one-point-zero.

The intake form worked—mostly. Linda's team uploaded the contracts through the secure portal, but forty-three of them arrived as scanned PDFs rather than native digital files. The system's OCR layer handled most of these, but eleven came through garbled—poor scan quality, handwritten amendments in margins, faded ink on older documents. Joshua spent most of the first day just getting the inputs clean enough to process.

The AI processing ran overnight. By 7:00 the next morning, the AI models had classified the 339 readable contracts, extracted key terms, compared them against Linda's standard positions, and flagged 762 deviations. But the results were uneven. For straightforward agreements—standard NDAs, simple service contracts—the analysis was crisp and reliable. For the more complex instruments—procurement agreements with nested amendments, multi-party frameworks with cascading obligations—the AI's confidence scores were erratic. It flagged a routine indemnification clause as high-risk in one contract and missed a genuinely problematic limitation of liability in another. The severity ratings were inconsistent: what the system called "critical" in one agreement it labeled "moderate" in a nearly identical provision three contracts later.

Joshua and Elena divided the review work, but the protocol that had seemed clear during testing buckled under the volume and variety of real-world con-

tracts. Elena kept finding edge cases the decision rules did not cover—what do you do when the AI flags a deviation but the client's own standard position is internally contradictory? Joshua, working faster, made judgment calls without documenting them. By Wednesday, David discovered they were applying different standards to the same types of provisions.

He stopped the line. Pulled them both into the conference room, documented every inconsistency he could find, revised the protocol with three new decision rules, and made them re-review the previous day's work. They lost a full day to the rework.

Sarah's quality sampling made things worse before they made things better. She pulled her 25 percent sample and found problems she had not expected—not just classification errors from the AI, but human errors layered on top. In her first batch of 40 reviews, she found seven that needed correction. Three were AI classification issues where the system had mischaracterized modified standard language as boilerplate. Two were cases where Joshua had approved a flagged deviation without adequate explanation. Two were formatting errors in the deliverable template that made the output confusing to read. An error rate of nearly 18 percent. David's production system had been designed to catch errors. It was catching them. The problem was how many there were to catch.

The eleven garbled contracts from the initial upload were never fully resolved. Sarah's team extracted what they could and manually reviewed the rest, but three of the contracts had provisions that remained partially illegible. Sarah called Linda to explain the issue and recommend that her team locate the original executed copies.

The deliverable went out on day sixteen—two days past the committed deadline. Sarah sent it with a cover memo she had agonized over for an hour, documenting every known limitation: the three partially illegible contracts, the provision types where the AI's analysis had been least reliable, and the error rate her quality sampling had revealed. David had insisted on the transparency. Sarah had resisted, then relented, then realized he was right.

4.5 Kaizen and More Kaizen

Linda called two days later. Her tone was measured, not warm.

"I have reviewed the report. There are issues. My team found four contracts where the risk ratings do not match what we would have expected based on our knowledge of those vendors. The formatting on the portfolio summary is inconsistent—some sections have the severity color coding and others do not. And three of the contracts in the high-risk category appear to have the same boilerplate analysis language, which makes me wonder whether anyone actually reviewed each one individually."

Sarah felt her stomach drop. "You are right about the formatting inconsistency—that was a template error we identified during our quality review but did not catch in every instance. The repeated language is a different issue. Let me investigate and get back to you today."

She investigated. The three contracts Linda had flagged were all in a batch Joshua had reviewed late on the Thursday before delivery, rushing to meet the deadline. He had used shorthand in his analysis notes that the deliverable template had rendered as identical text. The underlying analysis was actually different—the notes were in the system—but the client-facing output made it look like cut-and-paste work.

Sarah called Linda back within two hours. She explained exactly what had happened, showed her the underlying analysis notes, and acknowledged that the deliverable should have caught the discrepancy in quality review.

"I am not going to tell you this was the experience I promised," Sarah said. "It was not. We hit problems we did not anticipate, and some of the ones we did anticipate we did not handle well enough. What I can tell you is that we know exactly where every error occurred, why it occurred, and what we are changing so it does not occur again. That is not something your previous firm could tell you after six weeks and $105,000."

Linda was quiet for a moment. "The turnaround was fast, even with the delay. And I will say this—the cover memo documenting your limitations was something I have never received from a law firm. Brennan would have delivered with a confident cover letter and let me find the problems myself."

"We cannot afford to hide problems. Our model only works if we see them clearly."

"I am not exactly ready to serve as a reference," Linda said. "But I am not firing you either. I see promise. I see the possibility here. Fix the issues you have identified and send me the corrected deliverable. And then let us talk about what another engagement might look like—because if you can get the execution to match the concept, the economics and particularly the speed are genuinely compelling."

That afternoon, Sarah sat with David in the conference room. The butcher paper from their original mapping exercise was still on the wall. She did not feel like celebrating.

Unit Economics: The Truth Detector

Four metrics reveal whether an AI-native service scales:

Revenue per unit—the price charged for a single standard delivery.

Direct cost per unit—professional time, AI compute, and third-party costs.

Gross margin per unit—revenue minus direct cost.

Contribution margin—gross margin minus allocated overhead.

If gross margin per unit is healthy, the service scales. If it is not, no amount of volume will fix the problem.

"Walk me through the numbers," David said.

Sarah pulled up the spreadsheet. "Revenue: $52,500. That is the one thing that went according to plan. Professional time: Joshua and Elena spent a combined 80 hours at a fully loaded cost of $200 per hour—$16,000. That includes the rework day and then some. AI compute and platform costs: $5,500, much

higher than projected because we had to reprocess the garbled documents and rerun several batches. My review and remediation time: 18 hours at a higher fully loaded cost, call it $5,200. Total direct cost: $26,700." She paused. "That is about $76 per contract in direct cost. Gross margin: $25,800. That is 49 percent. And that cost per contract should decline as the system improves."

"Forty-nine percent on a first engagement with a version-one system," David said. "We are in striking distance of 50 percent and we are only going to get better from here. How does that compare?"

Contract Review: First Engagement Reality

Traditional model (Brennan & Associates): 4 associates, 6 weeks, 500–600 hours of labor. Revenue: $105,000. Cost: $60K–$75K. Gross margin: 29–43%.

AI-native model, first engagement: 2 lawyers + AI, 16 days, 98 hours of professional time (including rework). Revenue: $52,500. Cost: $26,700. Gross margin: 49%.

Messy, imperfect, and already better margin percentage than the traditional alternative. A significant improvement with turnaround time. Ideally, the gap will only widen as Candor's systems and processes improve.

"Brennan quoted $105,000 for the same work. Their gross margin would have been 29 to 43 percent. We are at 49 percent, and that is with a full day of rework, an error rate I am not proud of, and a client who is decidedly not singing our praises."

David nodded. "The goal is to approach 70 percent eventually if we can get there. Tell me where the 49 starts climbing."

Sarah had been thinking about this since Linda's call. "Three places. First, the intake process. Half our problems started with those scanned PDFs. If we add a validation step—check document quality before the engagement clock starts, reject files that do not meet the processing standard—we eliminate the garbled-document problem entirely. Second, the review protocol. Joshua and

Elena need tighter calibration. The inconsistencies were not because they are bad lawyers. They are excellent. But they were interpreting ambiguous rules differently because the rules were ambiguous. We fix the rules, we fix the inconsistency."

She paused.

"And third?" David asked.

"The AI." Sarah leaned forward. "The models are getting better, David. Not incrementally—fundamentally. We are actually behind on using the most current frontier models, and there are new agentic tools built on top of them—things like Claude Code—that we have not even begun to explore. In the future, we might actually combine multiple models, using one model to audit the output of another—what some people are calling 'LLM as a Judge.' As long as we can manage the token costs, we have a lot of leeway over time. What I know is that we are not just going to be using these models to flag deviations. Our orchestrated AI systems will reason about deviations—agentic architectures that can check their own work, cross-reference findings across a portfolio, and catch the kind of inconsistency that tripped us up with the severity ratings. Things on the technical front are moving in the right direction for us as an organization. When we combine that with what we have built—the process discipline, the quality loops, the measurement rigor—the improvement will compound. Better models feeding into better processes yielding better and better outputs as time moves forward."

David smiled. It was the first time she had articulated his framework back to him in her own words.

"What is the Japanese word?" Sarah said.

"Kaizen," David said. "Continuous improvement. Small, relentless, compounding gains." She looked at the error log David had printed out—three pages of documented failures, each one annotated with a root cause and a corrective action. A month ago, that document would have felt like an indictment. Now it felt like a roadmap.

"Kaizen, Kaizen, Kaizen," she said, and meant it.

With contract review productized, David turned his attention to the rest of the portfolio. He and Sarah spent a week applying the four S's to every service the firm offered.

Entity formations and corporate filings were clear candidates for Standardize. High volume, consistent inputs, rule-based processing. David designed a workflow similar to the contract review system: standardized intake, AI-generated documents from templates, human review for accuracy, automated filing submission. The target was a fully productized offering with fixed pricing and a 48-hour turnaround.

Regulatory compliance assessment fell into Supplement. The work was more complex—each client's regulatory environment was different, and the analysis required genuine judgment about risk tolerance and strategic priorities. But the research component, the regulatory mapping, and the initial gap analysis were all amenable to AI processing. David designed a workflow where AI handled the first 60 percent—research, mapping, preliminary analysis—and lawyers handled the remaining 40 percent, focusing on judgment, recommendations, and client communication. Pricing was fixed but higher, reflecting the human judgment component.

A small number of advisory engagements—novel regulatory questions, strategic counsel on corporate transactions—fell into Specialize. David's advice was counterintuitive: do not try to systematize these. "This is where your lawyers earn their keep as lawyers, not as workflow managers. The AI can help with research and drafting, but the value is in the conversation, the judgment, the relationship. Price it on value. Deliver it with the personal touch that justifies the premium."

And there were services to Stop. Sarah had been doing small-scale litigation support—document review for a few clients who had legacy relationships. The volume was too low to justify AI investment, the margins were thin, and it distracted from the firm's core positioning. David recommended stopping it.

"But those clients—" Sarah began.

"Refer them to someone who does it well. You cannot be everything to everyone. Every hour your team spends on low-margin litigation support is an

hour not spent on building the machinery to support the higher quality work that will make this firm truly valuable."

Sarah made the referrals the next week. It felt like a loss at first—turning away revenue always did. But within a month, the freed capacity had been absorbed by two new contract and regulatory review clients, generating three times the margin of the work she had let go.

The pricing conversation was one of the most difficult adjustments. Sarah had been trained in the billable hour. Her old firm priced everything by the hour because the hour was the fundamental unit of the traditional production model. You tracked time. You billed time. The client paid for time.

David challenged this from the beginning. "You are not selling time. You are selling outcomes. Your client does not care how many hours Joshua spends reviewing contracts. Your client does not really care if you take up 9 percent or 90 percent of the task with AI. Your client cares that the review is accurate, complete, delivered on schedule for a competitive price. You are selling quality assurance 'as a service.' "

"I know that intellectually. But when I sit across from a general counsel and quote a fixed fee, there is always this moment where they ask, 'How did you arrive at that number?' And if I cannot explain it in terms of hours, they get nervous."

"Then explain it in terms of what they are actually buying." David sketched on the whiteboard. "Per-unit pricing. You charge $150 per contract. The client knows exactly what they are paying. They can budget for it. They can compare it against their current provider. The transparency is the selling point, not the opacity."

Sarah thought about it. Her old firm's invoices were deliberately opaque—pages of time entries that obscured the relationship between effort and value. Clients hated them but accepted them because they had no alternative. Fixed-fee, per-unit pricing was the alternative.

"There is also the subscription model," David continued. "For clients with ongoing needs—companies that sign dozens of vendor agreements every month, that have consistent ongoing regulatory matters—offer a monthly subscription. Fixed fee for a defined volume. Overages priced per unit. The client gets budget certainty. You get recurring revenue. Investors love recurring revenue."

Sarah filed that away. The seed fund that had invested explicitly asked about recurring revenue during due diligence. A subscription contract review service would address that concern directly.

"And for the specialized advisory work?" she asked.

"Value-based pricing. What is the outcome worth to the client? If your regulatory advice helps a client avoid a $5 million fine, what is that advice worth? Not the hours you spent, not the cost of your AI compute—the value of the outcome."

"That is harder to sell."

"It is harder to sell and more profitable when you do sell it. But start with per-unit pricing on the standardized services. Prove the model. Build the track record. The value-based pricing conversations become easier when you have data showing that your work consistently delivers results." He paused. "And whatever you do, avoid hourly billing entirely. Hourly billing rewards inefficiency. Every workflow improvement, every AI enhancement, every process optimization reduces your revenue under hourly billing. Fixed pricing rewards efficiency. Every improvement increases your margin. Align the pricing model with the behavior you want to incentivize."

Over the course of the fall, as the production system would further take shape and the firm began to feel different. Not just in the workflows and the documentation—though those were transformative—but in the way the team operated. There was a rhythm now, a cadence that had not existed before.

Mornings started with a ten minute stand-up. David had imported the practice from agile software development. Each team member reported what they had completed, what they were working on, and what was blocking them.

Sarah resisted at first—it felt like unnecessary overhead for a five-person team. But she quickly saw the value. Problems surfaced faster. Work was distributed more evenly. Nobody was siloed in their own matters with no visibility into the rest of the firm's pipeline.

The metrics dashboard was David's other contribution. He built a simple tracking system that displayed the firm's key performance indicators in real time: matters in progress, average turnaround time, error rate, client satisfaction scores, revenue per professional hour, gross margin by service line. Sarah checked it every morning over coffee.

"You cannot improve what you do not measure," David said when he first presented the dashboard. "And you cannot scale what you do not track."

Joshua had been the most skeptical of David's approach. He was a gifted lawyer—meticulous, creative, and deeply committed to getting every detail right. The idea of standardizing his work felt like an insult to his craft.

"I did not go to law school to follow checklists," he told Sarah one evening after David had left.

"Neither did I," Sarah said. "But think about what you actually spent your time on last month versus this month. Last month, you spent more than twenty percent of your time on logistics—formatting documents, chasing client information, reinventing the intake process for each matter. Bet you did not go to law school to do that either. This month, the production system handles the logistics. You spent that twenty percent on actual legal analysis. On the judgment calls that nobody else—no AI, no checklist—can make."

Joshua was quiet for a moment. "The quality is better," he admitted. "I hate that the quality is better."

"Because it means the old way was not as good as we thought?"

"Because it means David is right. And he is not even a lawyer."

Sarah laughed. "He does not need to be. He is an engineer. He engineers the system. We practice the law. The system makes us better at the law. That is the whole point."

Joshua nodded, but Sarah noticed he did not smile. The concession was intellectual, not emotional. She filed that away.

Several months after David joined, Sarah had a conversation with Alex that crystallized how far they had come.

"Give me the numbers," Alex said. He was calling from his car, driving between meetings in the Bay Area.

"Monthly revenue run rate is roughly $110,000—annualizing to about $1.3 million—up from about $80,000 when I closed the round. Gross margin on standardized services is averaging 56 percent—better than the 49 percent on our first engagement but nowhere near where we need to be. EBITDA is running around 25 percent after technology costs and overhead. Turnaround time on contract review has dropped from an average of sixteen days to twelve. Error rate is down to 2.4 percent from our initial estimate of 3 to 4 percent. We have eight active clients, up from four."

"Client acquisition cost?"

"We are tracking that now. Average of $8,000 per client, mostly my time on pitches and proposals. Lifetime value is looking like $120,000 to $180,000 based on early retention data, but it is too soon to have confidence in those numbers."

"Net revenue retention?"

"One hundred and fifteen percent. Clients are expanding scope. Linda Torres started with one engagement and has now signed a subscription for ongoing review. Two other clients have added regulatory compliance assessment after starting with contract review."

Alex was quiet for a beat. "David was worth it."

"David was worth it," Sarah agreed. "He is not a lawyer and he is not a technologist. But he understood something that I did not: that the gap between a good team and a great firm is the production system. We had the team. He gave us the system."

"Listen, you are going to need more of him before it is all said and done. My suggestion is that you give him real upside somehow in the structure so he is aligned in the continued growth of the enterprise."

"Yes, I agree," said Sarah. "He has some other requests as well."

"What does he need?" asked Alex.

"He wants to hire a junior process engineer. Someone to help him document workflows, build training materials, and manage the quality assurance program as we scale. He also wants to start mapping the supplement-category services for AI workflow integration."

"Well, it is up to you but if it were me—I would 100 percent approve those hires," said Alex. "And Sarah—start thinking about the Series A story. The numbers you just gave me? Those are heading toward Series A numbers. You are months ahead of where I expected you to be. Keep cooking."

Sarah felt a surge of something she had not felt since leaving her old firm: not just confidence, but momentum. The firm was not just surviving. It was working. The production system David had built with the team was delivering exactly what the economics promised—higher quality at lower cost with expanding margins. And the flywheel was turning: each engagement improved the process, which improved the next engagement, which attracted the next client.

As was the nearly daily ritual, Sarah drove home that evening through downtown Phoenix, past her old firm's glass tower. She imagined the counterfactual world. The path she took and the path she was once on.

The building lights were still on—associates billing late hours, reading contracts page by page, the way they had always done it. She felt a pang of something that was not quite pity and not quite nostalgia. Those were good people doing good work. But they were doing it the hard way, and the economics of the hard way were about to get much harder.

4.6 A Saguaro Blooms in the Sonoran Desert

Long after everyone else had likely gone to bed, Sarah sat wide awake at the desk in her home office and drafted a memo to herself. Not for investors, not for clients—for her own clarity. She titled it "What We Learned" and wrote until midnight.

Building an AI-native firm is not primarily a technology problem. The AI is necessary but not sufficient. What differentiates us is not that we use the latest AI

offering—anyone can in principle do this. What differentiates us is the production system around AI. The workflows, the quality protocols, the feedback loops, the measurement infrastructure. David taught me that a firm is not its people or its technology. A firm is the system that connects people and technology to produce outcomes.

Industrialization is not the enemy of professionalism. It is the foundation of scalable professionalism. When every legal workflow follows the same intake, the same AI processing, the same review criteria, the same quality assurance, and the same delivery format, the result is not cookie-cutter work. The result is consistently excellent work that happens to also be efficient. The judgment remains. The craftsmanship remains. What disappears is the waste (muda).

Not everything should be productized. The four S's framework saved us from the trap of trying to automate everything. Some work should be standardized. Some should be supplemented. Some should remain specialized. And some should be stopped. The discipline is in knowing which is which and having the courage to act on the distinction.

Unit economics are the truth detector. You can tell any story you want about AI transformation, but the unit economics do not lie. Revenue per unit, direct cost per unit, gross margin per unit. If the numbers work at the unit level, the business scales. If they do not, no amount of storytelling will save it.

The flywheel is real. Each engagement improves our ability to harness AI to the mix and in turn reduces our error rates. Lower error rates reduce human review time. Less review time improves margins. Better margins enable further investment. The system compounds. But only if you capture the data. Only if you measure the errors. Only if you feed the corrections somehow back into the process. The flywheel does not turn by itself. It turns because you built the infrastructure to capture and use information. While we have some proprietary alpha and real differentiation in our supporting process, we still need some meaningful differentiation in the tech itself. That is our next frontier.

She saved the memo and closed her laptop. Through her apartment window, she could see the lights of Phoenix spreading toward the horizon—a city built in the desert, improbable and relentless. Not so different from what she was building, she thought. Something that conventional wisdom said should not

work, thriving anyway because the people who built it were willing to engineer solutions rather than accept constraints.

Tomorrow there would be more matters to process, more workflows to refine, more clients to pitch. The Series A conversation would eventually be here. The team needed to grow. The production system needed to evolve. More than anything, they needed to find real differentiation in the tech itself.

Coming Next Week

Chapter 5
Thin Wrappers and Deep Systems

Sarah's production system is running, but it is built on borrowed technology—frontier models accessed through APIs, off-the-shelf tools stitched together with custom prompts. Chapter 5 confronts the build-versus-buy decision: when should an AI-native firm build proprietary technology, when should it buy, and when should it partner? The answer determines whether the firm owns a defensible moat or rents a commodity.

Subscribe at `https://theainativefirm.com` to receive new chapters as they release.

About the Authors

Daniel Martin Katz, PhD, JD is Professor at Illinois Tech–Chicago Kent College of Law and Academic Director of the Bucerius Center for Legal Technology & Data Science. Named by the *Financial Times* as one of the top 20 legal market shapers of the past twenty years, his research focuses on legal analytics, legal technology, and the future of the legal profession.

Michael J. Bommarito II, MSE, MA is a serial entrepreneur, researcher, and adjunct professor with twenty-five years of industry experience. He has founded and led multiple companies at the intersection of artificial intelligence and professional services.

Jillian Bommarito, CPA, CIPP/US/E is an advisor, risk and governance expert, and one of the first certified AI auditors in the world. She brings deep expertise in compliance, privacy, and the governance challenges of deploying AI in regulated industries.

Get the Full Book

`https://theainativefirm.com`