AI Company Due Diligence Framework*

Context: $110B poured into private AI companies in 2025 (Stanford HAI). *>80% of AI projects fail* before production (RAND, 2024). Enterprise abandonment of AI initiatives surged from 17% → 42% YoY (S&P Global, 2025).

Rigor matters more than ever!

Basic Filter questions

Three questions to ask in a 15-minute call before running the full framework:

1. Can you do a live, unscripted demo using inputs I give you on the spot?
2. What is your current gross margin on AI revenue?
3. Name your three biggest enterprise customers and tell me what they use the product for daily.

PILLAR 1: IS THE AI REAL?

1A Detecting the AI Facade

Risk: Company claims AI automation while secretly relying on human labor (e.g., Builder.ai's 700 engineers, Nate's SEC fraud charge, Rabbit R1's staged demo)

- Can the team run a *live, unscripted demo* using inputs we provide on the spot. Inputs that nobody on their team has seen before?
- Walk me through every step of the product workflow:
    - Start with baseline workflow (without AI).
    - What are the metrics used to measure the workflow performance?
    - Next apply the AI to the workflow.
    - What are the before/after impact on the metrics?
    - Identify specific points humans review, correct, or complete what the AI automates
- What is your current human-to-AI task ratio for your core product deliverable?
- What third party (security firm, technical auditor, independent engineer) reviewed your system architecture and confirmed the product is genuinely automated and what were their conclusions?
- Can you show us server logs or API call volumes that corroborate the automation rate you're claiming?
- What is your headcount in operations vs. engineering – baseline vs. target AI-enabled headcount? Can you explain what operations staff do day-to-day?

1B Benchmarks & Performance Claims

Risk: Benchmarking: training specifically on benchmark tasks without improving real-world capability

- Which third-party benchmarks have you run against, and are they appropriate to your specific use case (not just generic LLM leaderboards)?

- What were the exact scores, and how do they compare to off-the-shelf alternatives (GPT-4o, Claude, Gemini) on the same tests?
- Have you published benchmark results anywhere outside a pitch deck, a blog post, research paper, or public leaderboard?
- How does your model perform on adversarial or out-of-distribution inputs i.e. data it wasn't trained on?
- What is your real-world error rate in production today, and how do you measure it?
- What are the errors the AI currently fails the most? How are you training it to improve?

1C Build vs. Wrap (Technical Architecture Spectrum)
Risk: Claiming proprietary tech that is actually a thin wrapper on OpenAI/Google

- Did you build your own model, or fine-tune an existing foundation model, or are you prompting a third-party API? Be specific.
- If you rely on a third-party AI provider (OpenAI, Anthropic, Google): what happens to your product if they raise prices by 3x, deprecate your model version, or shut down your API access?
- What is your current AI infrastructure cost as a % of revenue, and what is your gross margin?
- Do you have a multi-model or model-agnostic architecture that lets you swap providers without a product rebuild?
- What is your roadmap to proprietary technology do you have a timeline, budget, and the team to execute it?
- What training data do you own or have licensed rights to? Is any of it contested IP?

PILLAR 2: WHAT CAN THE AI ACTUALLY DO?

2A Capability Honesty
Risk: Use case stretches the AI beyond its structural limitations (e.g., deploying LLMs in zero-error-tolerance environments)

- Does your team acknowledge known limitations of your AI (e.g., hallucination, context window constraints, latency)? Walk me through them.
- Is your use case in a domain where a wrong AI output creates legal, financial, physical, or reputational harm? How is that liability managed?
- Have customers experienced AI errors in production? What happened, and what was the remediation?
- What safeguards i.e output validation, confidence thresholds, human review queues, are in place to *catch mistakes before they reach the end user*?
- Is the AI used for *high-stakes decisions* (hiring, credit, medical, safety)? If so, how are you managing bias, explainability, and auditability?

2B Ghost Autonomy / Wayve Test
Risk: Product pivots multiple times because the underlying AI can't reliably solve the original problem

- What is the most technically difficult thing your AI does today that it could not do 12 months ago?
- Has the product scope or target use case *changed significantly* since your last funding round? What drove those changes?
- What is the *failure mode* that could most credibly force a product pivot, and how are you hedging against it?

PILLAR 3: WHERE CAN THE AI BE DEPLOYED?

3A Regulatory & Compliance Exposure
Risk: Untracked legal liability across jurisdictions (EU AI Act, China's three AI laws, Brazil's AI framework)

- In which markets do you operate or plan to operate? Have you mapped which AI-specific laws apply today for eg. EU AI Act, CCPA, EEOC guidance, sector-specific rules (HIPAA, FINRA)?
- Under the EU AI Act, how is your system classified by risk tier? Have you begun a conformity assessment?
- Have you addressed any sector-specific US regulation?
    - For healthcare, have you engaged with FDA on whether your system constitutes a Software as a Medical Device (SaMD)?
    - For financial services, have you assessed SEC/FINRA guidance on AI in investment advice?
- Do you use AI in any context that touches employment decisions, credit scoring, insurance, or law enforcement? If yes, has legal counsel reviewed those workflows?
- Who on your team owns regulatory compliance? Is it a lawyer, a compliance officer, or a founder wearing multiple hats?

3B Documentation & Auditability
How do you document your training data? What are the sources, licensing, consent, and exclusions?
Can you produce a *model card or system card* that describes what the model does, what it was trained on, and its known failure modes?
Do you maintain logs of AI outputs, errors, and corrective actions? How long are logs retained and who can access them?
If a regulator or enterprise customer audits your system, how long would it take you to produce a *full decision trail* for a specific AI output?

PILLAR 4: WHAT ABOUT AI AGENTS?

*Applicable if the company is one of the ~90% of YC F25 companies building autonomous agents

- Define the *permission boundary*: what actions can the agent take autonomously (send emails, execute transactions, write to databases) vs. what requires human approval?
- What is your *rollback and reversal capability*? If the agent executes a wrong transaction or sends incorrect data, can it be undone? How fast?

- Have you conducted *red-team or adversarial testing* for prompt-injection attacks i.e. scenarios in which a bad actor tries to hijack the agent via crafted inputs?
- When the agent connects to external tools (APIs, databases, third-party services), what controls prevent *sensitive customer data from being exfiltrated or exposed*?
- How does the agent behave when it *encounters an ambiguous or out-of-scope request* i.e. does it fail safely, escalate, or act unpredictably?
- What is your *liability framework* when the agent takes an autonomous action that causes measurable harm to a customer?

PILLAR 5: WHERE IS THE OPPORTUNITY?

5A Growth Trajectory (Supernova vs. Shooting Star)
Framework: Bessemer Ventures defines Supernovas ($0 → $100M ARR rapidly), Shooting Stars (strong PMF + NRR + sustainable margins). Anthropic is an example of a supernova.

- What is your current *ARR, MoM growth rate*, and *net revenue retention* (NRR)?
- What is your *payback period on CAC*, and how does it trend as you scale?
- Please describe your unit economics trajectory: At what ARR does the business become self-sustaining on gross margin alone, and is that a credible path given current burn and AI infrastructure costs?
- Are you growing because of genuine *product-market fit*, or because of a hype cycle that is pulling forward demand?
- What does *customer churn look like* 12 months after initial deployment i.e. are customers expanding usage or contracting?

5B Defensibility & Moat
- If OpenAI or Google released *a native feature* that replicated your core function tomorrow, what would remain defensible about your business?
- Do you have *proprietary data flywheels* i.e does your model get meaningfully better as more customers use it?
- What is your *switching cost* for an enterprise customer 18 months into deployment?
- Are you building a *vertical application* with deep workflow integration, or a horizontal tool that is easier to commoditize?

PILLAR 6: SECURITY

6A AI-Specific Attack Vectors
- Prompt injection: Can a malicious user hijack the AI's behavior by embedding instructions in user-supplied content (e.g., a document the AI is asked to summarize)? Has this been tested systematically, not just anecdotally?
- Model inversion and extraction: Can an adversary query the model repeatedly to reconstruct training data or clone the model itself? (This is both a security risk and an IP risk.)
- Data poisoning: If the model is retrained on user feedback or new data, what prevents a bad actor from deliberately injecting corrupted data to degrade or manipulate model behavior?

- Membership inference: Can an attacker determine whether a specific individual's data was used in training? (This is a GDPR and HIPAA exposure vector that most AI companies haven't assessed.)
- Adversarial inputs: Can carefully crafted inputs cause the model to produce dangerous, biased, or confidential outputs? Has systematic adversarial testing been conducted?

6B Infrastructure Security

- Is someone assigned to overall system security?
- Has the system undergone a formal third-party penetration test? By whom, when, and was remediation verified?
- For cloud/SaaS-based solutions, is there tenant isolation in inference? Example – customer A's data processed by the model should have zero possibility of appearing in customer B's outputs — model memorization makes this a real and documented risk, not a theoretical one.
- Where are model weights stored, who has access, and how are they protected? Model weights are among the most valuable and portable IP a company owns.
- What are your access controls around training data, fine-tuning pipelines, and model deployment infrastructure?

6C Incident Response

- Do you have an AI-specific incident response plan? Example challenges – a mass hallucination event, a prompt injection attack at scale, or a data leak via model output requires a different playbook than a conventional data breach.
- What is your mean time to detect and respond to an AI output failure in production?
- Have you had any security incidents to date? What happened and what changed afterward?

6D Supply Chain Security

- What is your exposure if your foundation model provider (OpenAI, Anthropic, etc.) suffers a security incident? Their breach becomes your breach if you're passing customer data through their API.
- Do you have data processing agreements in place with every third-party AI vendor in your stack? Have you audited their SOC 2 / ISO 27001 status?
- What open-source model components or libraries are in your stack, and how do you track CVEs against them?

6E Compliance Certifications

- What security certifications do you hold or are pursuing — SOC 2 Type II, ISO 27001, FedRAMP (if selling to government)?
- For enterprise sales specifically: what is your estimated timeline to SOC 2 Type II if you don't have it? This is increasingly a hard requirement, not a nice-to-have.

PILLAR 7 TEAM AND TALENT

- Does the team have ML engineers who have shipped production AI systems (not just demo-ed them)?
- What is the ratio of researchers to engineers? (Pure research teams without deployment experience routinely fail to productionize.)

- Who owns the relationship with your foundation model partners (OpenAI, Anthropic, etc.) and what's the depth of that relationship?
- What is ML engineer attrition? AI talent is acutely mobile — losing two key engineers can collapse a small team's capability.
- Are there non-competes or IP ownership disputes from prior employers around the core technology?

PILLAR 8 DATA STRATEGY

- What is your proprietary data advantage, if any? Is it size, recency, exclusivity, or labeling quality?
- How is training data labeled — internal team, contractors, crowdsourcing? What quality control exists?
- What is your data pipeline for continuous retraining? How often does the model retrain, and what triggers it?
- Do you have Reinforcement Learning from Human Feedback (RLHF) or human feedback loops? Who provides the feedback and what are their qualifications?
- What happens to customer data — does it get used to train the model? Have customers consented to this?

Red Flag Summary

Signal	**What It May Indicate**
Won't do a live, unscripted demo	AI Facade / human-in-the-loop disguised as automation
Benchmarks only shared in pitch settings	Benchmaxxxing or cherry-picked results
Headcount skewed heavily toward operations or growing faster than ARR	Manual labor hidden behind AI branding
Claims to have "solved" hallucination	Technical dishonesty or misunderstanding of LLM architecture
No third-party architecture review	Unvalidated proprietary tech claims
No answer on regulatory mapping or no third-party audit on security code	Blind spot that could create material liability
No named compliance owner	Regulatory exporsure
No names security owner	Security exposure
Multiple pivots on core use case	AI capability gap that couldn't close
No rollback mechanism for agent actions	Catastrophic failure risk in agentic products
Gross margin <50% at scale	AI infrastructure costs not yet under control

Sources:

- Stanford HAI AI Index 2025
- RAND Corporation 2024
- S&P Global Market Intelligence 2025
- Bessemer Ventures State of the Cloud
- Kalai & Vempala (MIT/Georgia Tech) arXiv:2401.11817
- Banerjee et al. arXiv:2409.05746
- "Investing in Artificial Intelligence: A Natural Approach to Due Diligence" by X.Eyee. Presented at ACA 2026, Denver.