Private AI for the Boardroom

Generate, by Iterate.ai, is a private, governed AI + Agent environment for board directors.

THE PROBLEM

Boards have entered a dangerous gray zone.

Right now, a board director, somewhere, is pasting next quarter's earnings draft into ChatGPT, Grok, or Claude to summarize it. It's not malice — it's productivity. But it creates exposure that no general counsel would sign off on if asked directly.

- **Data leakage.** Sensitive board materials — earnings, M&A, strategy, legal memos — are being pasted into public LLMs with no audit trail and no control over retention.
- **Zero visibility.** Companies cannot see what directors upload, which tools they use, or where data flows after it leaves the boardroom.
- **Legal exposure.** AI prompts and outputs may become discoverable in litigation. Third-party AI use typically falls entirely outside corporate data policies.
- **Insurance gap.** Cyber insurance coverage for AI-related breaches has collapsed — only a handful of states remain covered. A Director's use of public AI is an insurability gap.
- **Agents gone wild.** As AI usage scales, uncontrolled automation introduces decision risk, unintended actions, and downstream data exposure.

For directors and executives who want to go deeper on the risks of public AI, see:

- (a) **Understanding the AI Revolution** and (b) **An AI Field Guide: Taking Action** — IterateOn.ai/books
- **The World Has Changed** and **AI Risk Disclosure and the Regulator in the Room** — IterateOn.ai/board

Beyond AI for employees. Beyond AI for developers. ***AI for the Board.***

High-risk surface. Clear value. Easy budget conversation. Close the boardroom AI gap

THE SOLUTION

Private AI for the Boardroom: Generate

Generate, Iterate's secure AI + Agent application, provides every director with a private AI environment, governed by company policy, with full observability into how AI is used with confidential materials. It is not just another productivity tool. It is a governance system, a risk mitigation layer, and a board intelligence platform — built specifically for the place where data is most sensitive, and controls are weakest.

PRIVATE AI
Per director — runs locally on your company's hardware or on a dedicated (fully private) Equinix server. No external API calls to Claude, ChatGPT, or Gemini.

PRE-MEETING INTEL
Board materials uploaded and indexed. Directors summarize, compare quarters, and model scenarios.

GOVERNANCE
Every interaction is logged for audit and e-discovery. Advanced governance, policy enforcement, and detection are available via AgentWatch.

Most products marketed as *"private AI"* **still depend on external model APIs. Generate closes all three loops: your data, your model, your hardware.**

	Public AI	API-Dependent AI	Private AI *(Iterate.ai)*
Data Control	✗ Shared servers	✓ Your environment	✓ Your environment
Model Control	✗ Shared LLM*	✗ Same shared LLM*	✓ Your infrastructure
Hardware Control	✗ Shared hardware*	⚠ Split*	✓ Yours
Policy Control	✗ None	✗ None	✓ Complete
The Reality	Zero control	Marketed as "private" but API-dependent and shared LLM	**Truly private**

** Frontier models like ChatGPT, Claude + Gemini are too large to host privately. Every customer queries the same shared LLM + GPU instance. This is not the shared infrastructure of old IT — LLMs hold prompts in working memory, where data exposure operates by entirely different rules. Memory is what makes AI governance categorically different from the IT era we are leaving. See appendix.*

USE CASES

What a director actually does with your company data.

- **Pre-meeting prep.** "Summarize the key risks in the CFO's report and compare to last quarter."
- **M&A analysis.** "What assumptions drive the synergy estimates? How does this compare to our 2023 acquisition?"
- **Governance & risk oversight.** "Does this new data privacy policy conflict with our existing retention policy?"
- **Strategic decision support.** "How does the proposed international expansion align with our stated risk tolerance?"
- **Legal & e-discovery readiness.** "Show every interaction this director had with the M&A materials in Q3."

DEPLOYMENT & PRICING

Options.

Deployment Options	Location	Setup	Monthly	Users / Privacy
Single Tenant *Recommended for boards*	**Datacenter***	**$20,000****	**$2,500 / mo**	Up to 30 directors, officers — 100% private, dedicated

** On-premise is available upon request when the customer supplies and owns the hardware. The setup fee is waived when the customer supplies compatible RTX Pro hardware.*

*** Setup of dedicated hardware, including GPU, CPU, and load balancing.*

ROLLOUT

Pilot to production in weeks.

Phase 1 — Pre-Rollout (days). Management uploads recent board materials. Validate workflows and confirm governance policies with general counsel.

Phase 2 — Full Board Deployment (14 days from initial request). Happens quickly — mostly logistics. Simple onboarding. Establish retention and policy controls. Connect financial, legal, and data systems, as needed. Extend to audit committee workflows and risk management dashboards.

APPENDIX

A bit more about Private AI.

The technical reality. Frontier models like ChatGPT (GPT-5), Gemini, Grok, and Claude.ai have hundreds of billions to over a trillion parameters. Running one instance requires racks of H100/B100 GPUs and tens of millions in capital infrastructure. No enterprise — Fortune 10 included — runs its own private GPT-5. They can't. The economics don't work, and the weights are not available for licensing. OpenAI (ChatGPT), xAI (Grok), Google (Gemini), DeepSeek, and Anthropic (Claude) each operate a single shared instance of their flagship models, and millions of users access the same model.

That is why the data exposure problem is not a bug — it is a structural requirement of how frontier AI is delivered today.

Vendors like OpenAI share your space with your competitors because it is the only way for frontier models to be economically viable. The moment a prompt enters that shared instance, control over where it goes, what it trains, and who can subpoena it is in the hands of your suppliers (third parties). Not yours.

More importantly, the enterprise does not control how that data is *used* once it enters the AI system. With modern AI, the model itself is intertwined with the memory.

In traditional IT, access control answered the question: *who can see the data?*

In AI systems, the more important question becomes:

> *"Should the system be allowed to use this data in this context, for this purpose, at this moment? What if competitors gain advantages from the fingerprints we leave?"*

That is not a data storage problem. It is an execution problem.

And in shared frontier models, the enterprise does not own that execution path.

Generate's structural advantage. Iterate builds smaller, efficient models, fine-tuned for business use and right-sized to run on a single air-cooled RTX Pro workstation GPU. Patented and patent-pending KV cache and runtime inventions — including Lifeboat — push that hardware far further than off-the-shelf inference engines.[1] That is what makes private AI possible at boardroom and enterprise economics. Big enough to be useful, small enough to be yours, fast enough to compete with anything in the market — and strong on margins and on ESG.[2]

A real conversation: when a public company CDO asked the right question.

[1] ***Lifeboat*** is one of many runtime inventions patented or patent pending by Iterate.ai. Most of our patents target infrastructure level processing that makes private AI possible and economically feasible for the enterprise. It's why world-class technologies companies like NetApp, Equinix, AMD, and Qualcomm have strategic partnerships with Iterate. See Iterate's IP: https://iterate.ai/company/intellectual-property. And Iterate's partner page: https://iterate.ai/partners

[2] In benchmarks against best-in-class SGLang, Generate delivers 2x to 6x performance gains. In customer deployments — even with frontier models — inference costs drop by up to 95% versus hyperscaler delivery, image generation falls from 80¢ to 4¢ per image, and render time drops from 12 seconds to 2. Iterate's runtime can even run voice and text-based LLMs on tiny Qualcomm 6490 chips — inside Zebra-type handheld computers — not originally designed for language models. Equinix ($100B), Qualcomm ($175B), and NetApp ($25B) all rely on Iterate for private-environment inference at scale.

In a recent working session, the Chief Digital Officer of a publicly traded consumer brand pushed past the usual AI capability questions and asked the one her board and investors will ask:

> *"What does this do to net earnings? And what is my sustainability narrative? If I am using AI for everything, am I quietly increasing our energy footprint? Am I quietly creating legal exposure by asking questions of third-party-owned LLMs?"*

That is a CEO and audit-committee question, not a technology question. It reframes AI from a productivity story into a margin, ESG, and governance story — and exposes a risk most enterprises have not priced in: poorly architected AI can increase compute costs, legal risks, energy consumption, and ESG exposure faster than it produces measurable productivity gains.

Frontier models make this worse. Trillion-parameter models like GPT-5 and Claude Opus consume enormous compute for every query, including simple ones. The energy is spent regardless of whether the task warranted it.

> *"Using a trillion-parameter model to summarize a one-page CEO note is*," as one of Iterate's engineers put it, *"like using a bazooka to swat a fly."*

Generate is built on the opposite premise. Iterate.ai uses smaller, efficient models — fine-tuned 400-billion to 36-billion-parameter models — to handle the majority of board-level analysis tasks with comparable quality, far less legal exposure, and a fraction of the compute. They run on air-cooled NVIDIA RTX Pro workstation GPUs rather than liquid-cooled hyperscale clusters. The architecture is itself a sustainability and cost story:

- **Lower compute per query —** fewer GPU hours, lower energy draw.
- **Air-cooled hardware —** no liquid-cooling infrastructure, lower facility energy overhead.
- **Right-sized models for the task —** not a trillion-parameter model summarizing a memo.
- **Optimized Inference** — Lifeboat uses 2x more efficient compute than worldclass benchmark SGLang.
- **Predictable, flat economics —** *no per-token billing* that scales with usage.

For a public company board, this becomes a defensible investor narrative:

> *"Our AI strategy is designed to improve director preparedness and decision quality while minimizing compute cost and energy consumption by utilizing right-sized models and private infrastructure."*

Defending unbounded API spend against frontier vendors, legal exposure due to prompting third-party LLMs, and a rising ESG carbon-intensity question is not defensible.

The deeper structural argument for private AI: It is not just a privacy and governance choice. It is a margin, sustainability, and **control-of-execution** choice.