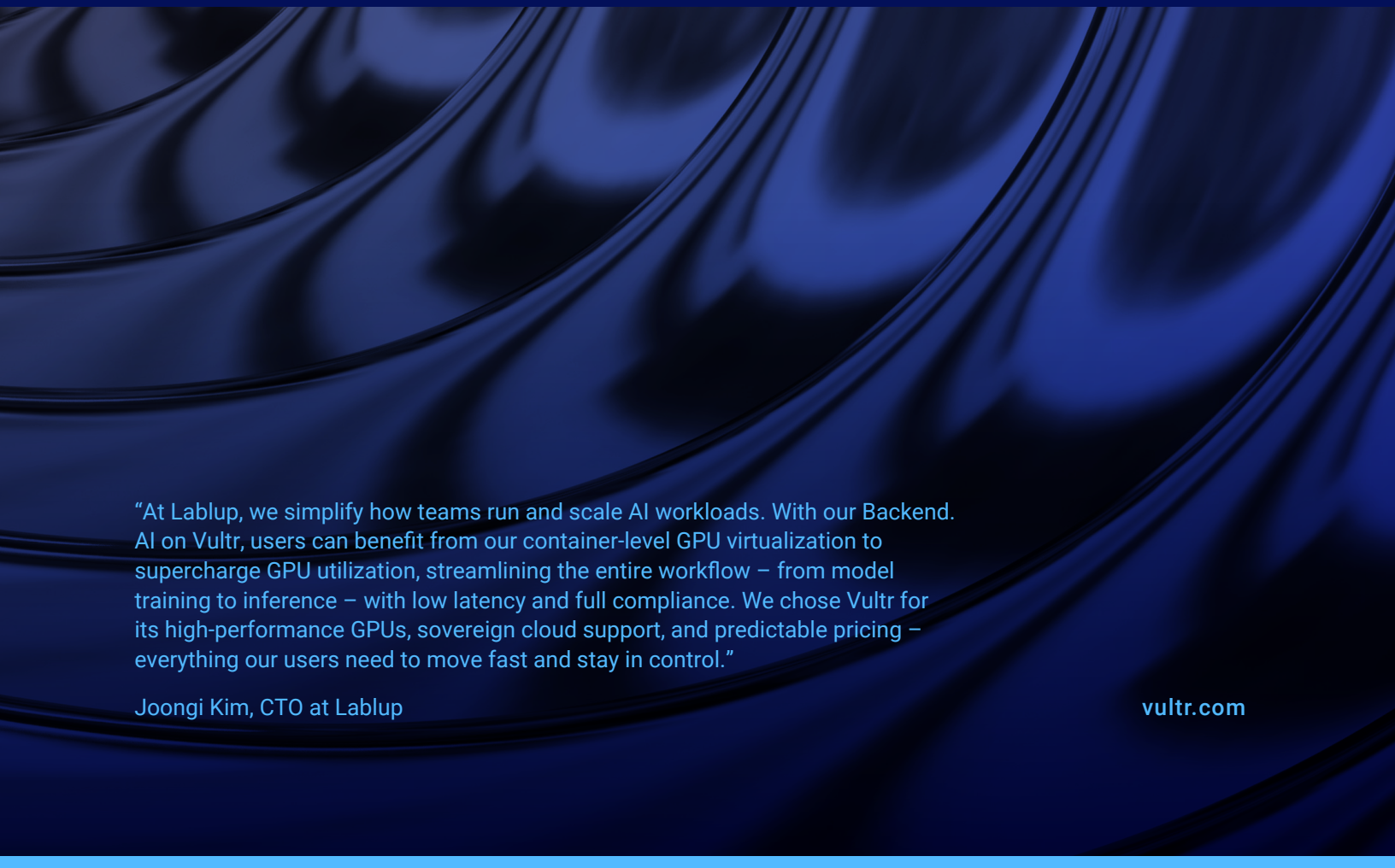
VULTR

Vultr + Lablup

Supercharge Legal Documentation LLMs with NVIDIA HGX B200

Fine-tune and deploy legal AI faster, with zero DevOps pain.



"At Lablup, we simplify how teams run and scale AI workloads. With our Backend. AI on Vultr, users can benefit from our container-level GPU virtualization to supercharge GPU utilization, streamlining the entire workflow – from model training to inference – with low latency and full compliance. We chose Vultr for its high-performance GPUs, sovereign cloud support, and predictable pricing – everything our users need to move fast and stay in control."

Joongi Kim, CTO at Lablup

vultr.com

VULTR lablup

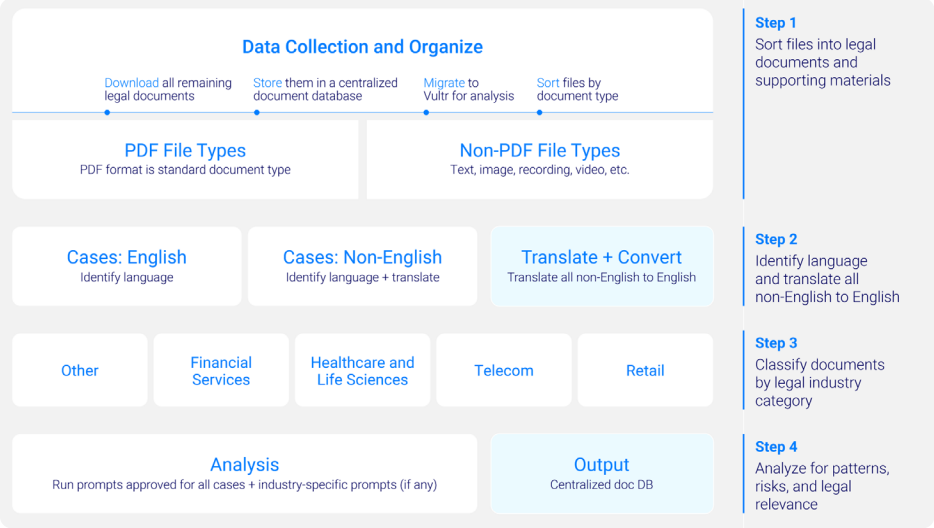
Use Case

End-to-End LLM Performance, Accelerated by NVIDIA HGX B200

Legal AI is here, but infrastructure lags

Legal tech is rapidly transforming thanks to generative AI. According to Gartner, 89% of legal and compliance leaders say identifying GenAI opportunities is a top priority in 2025.[1] From contract review and clause comparison to regulatory triage and court judgment analysis, large language models (LLMs) are helping legal teams streamline repetitive work and focus on high-value tasks.

However, document-heavy workflows demand scalable infrastructure – something many legal teams lack the time, budget, or technical expertise to build themselves. The chart below shows how GenAI reveals patterns in legal and compliance data.



Challenges facing legal tech teams

Many legal teams encounter roadblocks when scaling AI. Their local compute infrastructure is often overloaded or outdated. Managing AI workloads across teams is complex, and deploying inference APIs across regions requires time and expertise they may not have. Hiring DevOps to manage it all adds cost, and staying compliant with jurisdiction-specific data laws complicates things.

- Too much time spent managing infrastructure
- Not enough GPU availability when it's needed
- Compliance concerns slow down AI adoption

Most legal teams aren't set up to run AI at scale. Without fast, flexible, and compliant infrastructure, even the best GenAI ideas can stall before they start.

Industry
AI Software

Use Case
Fine-tuning and global inference for open-source LLMs on legal documents

Key Technologies

- NVIDIA HGX B200
- Lablup Backend.AI
- Vultr Cloud GPU
- PyTorch
- NGC containers

Customer Spotlight
Lablup

Legal document AI across industries

Financial services: Classify loan documents and investment contracts for risk.

Healthcare: Scan patient consent forms for legal gaps or compliance issues

Public sector: Analyze policy drafts and translate legal language for stakeholders

Media and entertainment: Manage rights and licensing clauses in production agreements

[1]Gartner "Generative AI Use Case Spotlight for Legal and Compliance Teams"

Vultr and Lablup: Simplified AI infrastructure

Vultr and Lablup remove the friction from building and scaling AI workflows. Vultr offers high-performance global infrastructure with access to NVIDIA HGX B200 GPUs and predictable pricing, delivering cost-effective compute without vendor lock-in. Its AI-native, composable architecture adapts to any stage of the AI lifecycle, from fine-tuning to global inference.

Built-in compliance, including ISO 27001, SOC 2 Type II, and GDPR, ensures deployments meet strict data requirements.

Backend.AI adds intelligent orchestration, automates job scheduling, optimizes GPU usage with fractional scaling, and simplifies model training and inference: no DevOps needed. Together, they deliver a fast, secure, and efficient platform for GenAI in legal and compliance use cases.

Legal AI needs infrastructure that can keep up

In this use case, a legal-tech company is training a domain-specific LLM to identify risks in commercial contracts. The model will serve global enterprise clients and must meet performance and compliance expectations.

Challenge

A legal-tech company is building a domain-specific LLM to analyze commercial contracts and flag compliance risks. Their GPU servers are fully booked, and adding hardware would take weeks. The team must process thousands of legal documents for fine-tuning, ensure inference is available globally for enterprise clients, and meet data residency requirements in Europe and North America without hiring DevOps or rebuilding infrastructure.

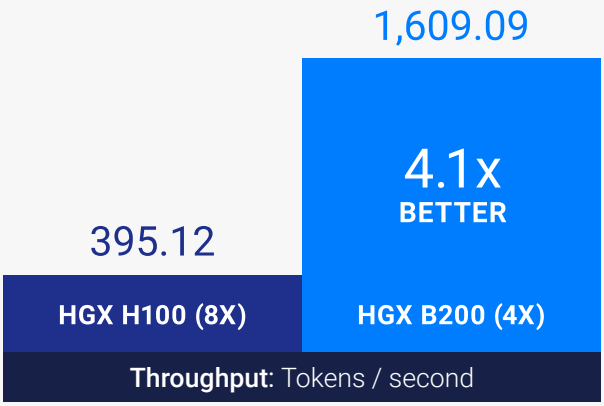
Solution

The team uses Vultr Cloud GPU, accelerated by NVIDIA HGX B200, and runs Backend.AI for orchestration. They fine-tune their LLM with PyTorch across multiple GPUs, schedule jobs overnight, and release resources when idle. Fractional GPU scaling improves efficiency. Once trained, they deploy inference APIs across Vultr regions with Backend.AI, which handles GPU-aware routing, ensuring fast, compliant access for users worldwide.

How the benchmark was conducted

The team benchmarked the Qwen3-235B-A22B model to evaluate performance on a representative instruction-following task: selecting 100 questions based on 20 example inputs. Tests were conducted using vLLM version 0.8.5 with CUDA 12.8 and NVIDIA Driver 570. The comparison involved two environments: 4× NVIDIA HGX B200 GPUs on Vultr Cloud GPU and 8× HGX H100 GPUs in the company's on-prem cluster.

Benchmark: Legal document LLM on Vultr using NVIDIA HGX B200



Metrics	8 × HGX H100	4 × HGX B200
Throughput Tokens/s	395.12	1 ,609.09
Cost effectiveness	1.3 x unit cost	½ the GPU count
Overall efficiency		~6x

Results and operational impact

Inference deployment: Inference APIs ran on NVIDIA NIM and Backend.AI across 32 Vultr regions, delivering ~42 ms latency with GPU-aware routing and full compliance.

Performance and efficiency: The Vultr setup with 4× HGX B200 GPUs hit 1,609 tokens/sec – 4.1× faster than 8× H100s on-prem, with ~6× better efficiency due to higher utilization and smarter orchestration.

Provisioning and Utilization: Cluster spin-up took under 5 minutes, with Backend. AI driving >95% GPU utilization through fractional scaling and container-level control.

Business Impact: The team fine-tuned and deployed LLMs faster – with no DevOps overhead. Global rollout met all compliance needs, accelerated model iterations, and freed engineers to focus on product, not infrastructure.

Additional resources

AI-Driven Content Processing →
Sovereign Cloud Market Landscape →

Contact our sales team, or visit vultr.com to learn more.