INVERGENT

Datasheet

DenseMax Appliance

Professional AI for the enterprise

Plug&Play In-House AI

The DenseMAX Appliance is a high-performance AI infrastructure solution engineered for enterprise-scale generative AI. Purpose-built for demanding workloads, provides organizations with a turnkey system for developing, deploying, and optimizing advanced AI applications.

By combining cutting-edge hardware with the DenseMAX Studio, the solution empowers enterprises to accelerate innovation, streamline operations, and maintain full control over their AI lifecycle.

It simplifies the journey from experimentation to deploying enterprise AI applications by providing enthusiasts, developers, and AI builders with pre-optimized models and industry-standard APIs for building powerful AI agents, co-pilots, chatbots, and assistants.

Instant Deployment, No Complexity

- Pre-configured 8x Blackwell GPU system optimized for high-throughput LLM inference.
- Drop-in setup: power it up, connect it to your network, and start building and serving models within minutes.

Complete Data Sovereignty

- Keep sensitive data on-premise — no cloud dependency or data egress risk.
- Strict GIT-like versioning policies model & dataset customization
- Role-based access control, audit logging, and optional air-gapped operation.

Key Features

All-in-One AI Platform
Everything you need to build, deploy, and scale generative AI.

Fast Time-to-Value
Go from idea to production in days, not months

Continuous Innovation
Keep models fresh with easy fine-tuning and retraining.

Built-In AI Safety
Guardrails and alignment tools for responsible, trustworthy AI.

Collaboration at Scale
Empower multiple teams with shared datasets, models, and workflows.

Enterprise Confidence
Proven tools for monitoring, compliance, and reliability.

Optimized for Performance
Run AI faster, cheaper, and more efficiently at scale.

Future-Proof
Add more GPUs and appliances anytime to be ready for what's next.

Blackwell-optimized CUDA kernels

- Custom software components built for the Blackwell architecture deliver sub-second latency and massive throughput even under multi-user load.
- Includes the latest carefully chosen open-weight LLMs

Training & Fine-Tuning

- Cutomize models directly on the appliance with a no-code UI — no need for external tools or programming skills.
- Create optimized models using latest techniques like Reinforcement Learning and Knowledge Distillation.
- Supports parameter-efficient tuning methods (LoRA, QLoRA) to maximize speed and minimize resource usage.

Enterprise-Grade Monitoring & Observability

- Complete LLM evaluation framework for measuring accuracy and red-teaming security assesments.
- Real-time metrics for GPU usage, memory, token throughput, and latency.
- Integrated alerts and logging for proactive maintenance and issue detection.

Secure by Design

- RBAC access control, workload and containerized model isolation.
- Access gateway for inference with end-to-end observability, security and audits
- TLS-enabled APIs and full compliance with enterprise IT policies.

DenseMAX Studio

Shipped with the enterprise-grade LLMOps platform designed to accelerate the development and deployment of generative AI applications offering state-of-the-art language models with robust services for deployment, fine-tuning, evaluation, safeguarding, and optimization.

Use Cases

🧠 Agentic Workflows

Deploy AI agents that take actions across internal systems, apps, and APIs — ideal for process automation, research, and task orchestration.

🔐Use & Protect Sensitive Data

Run inference and fine-tuning on data that cannot leave your infrastructure.

🧰Customize LLMs

Tailor models to domain-specific language, tone, and behavior using internal datasets — all managed through a low-code UI.

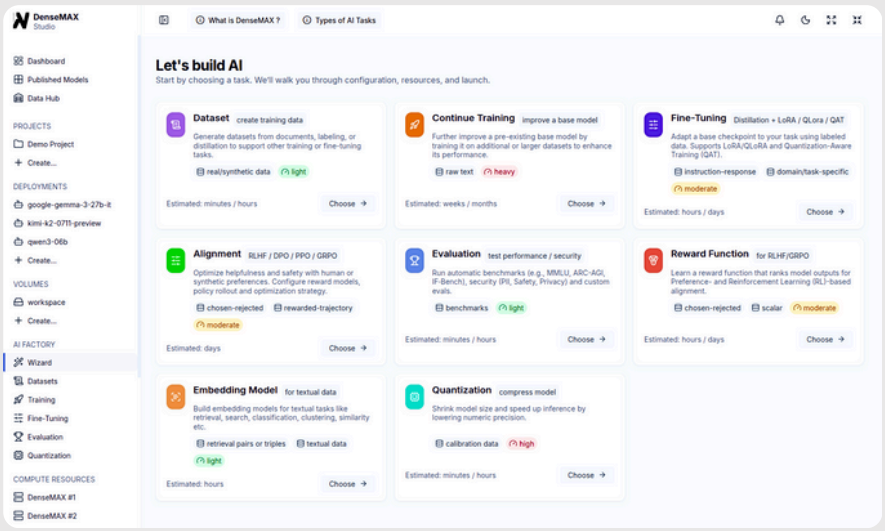
🧩Internal Tools and Workflows

Connect models to CRMs, ERPs, ticketing systems, document stores, or proprietary UIs for AI-native productivity.

🌐Multi-Tenant AI Across Teams

Serve different departments (e.g., legal, HR, marketing) from the same appliance with isolated, parallel model deployments.
osts and reducing TCO over time.

Let's bring AI home.



Use Cases

🔁 Continuous Learning Loops

Use internal feedback and usage data to fine-tune and improve models regularly — keeping performance aligned with evolving needs.

🧪 Experiment & Iterate Locally

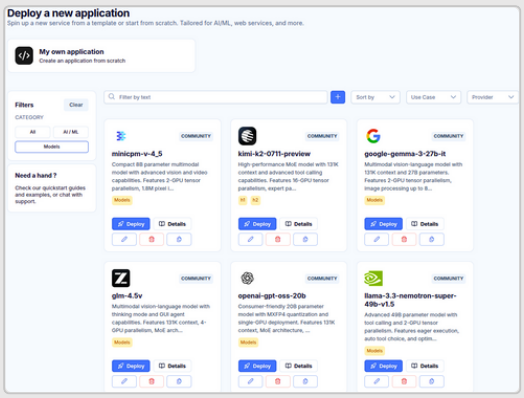
Rapidly test prompts, tune configurations, and evaluate model behavior without reliance on cloud costs or vendor limitations.

💸 Predictable, Contained Costs

Avoid unpredictable usage-based cloud pricing. Run unlimited inference and fine-tuning workloads on a fixed-cost platform, eliminating API costs and reducing TCO over time.

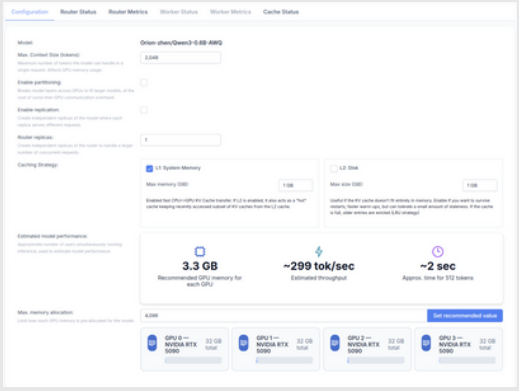
Application Deployment & Serving

The platform provides a user-friendly interface for launching AI applications—either from pre-built templates for common use cases or through custom deployments tailored to organizational needs. This dramatically reduces deployment time and simplifies operational complexity, ensuring even non-specialist teams can bring AI solutions into production with minimal effort.



Production-Grade Model Deployment

DenseMAX Studio is engineered for demanding enterprise workloads, supporting KV-aware cache routing, GPU sharding, model replicas, and disaggregated serving. These features ensure low-latency inference, fault tolerance, and the ability to run multiple large models in parallel, which is critical for high-traffic enterprise applications.



Secure & Observable Model Serving

Security and trust are built in at every layer. The platform includes role-based access control, built-in guardrails, and full-stack observability/monitoring to ensure compliance, safety, and transparency. Enterprises gain visibility into performance and risks, allowing proactive detection of anomalies, bias, or drift.

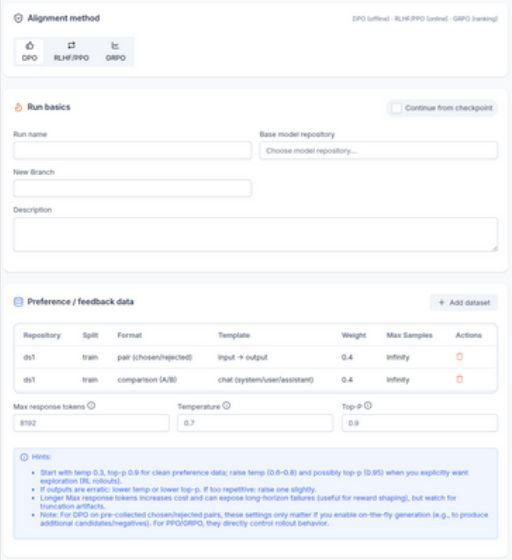
Data Hub for Models & Datasets

DenseMAX Studio provides a central repository for model and dataset lifecycle management, supporting Git-like operations such as branching, tagging, pull requests, and diffs. Users can also explore, visualize, and transform datasets interactively, while enjoying seamless integration with Hugging Face and ModelScope. This ensures governance and collaboration at scale, preventing silos across teams.

Let's bring AI home.

Training & Fine-Tuning Pipelines

The platform supports both parameter-efficient methods (like LoRA) and full fine-tuning for maximum flexibility. Teams can also generate synthetic datasets on demand, train embeddings, or develop reward functions for reinforcement learning. Enterprise-grade workflows and automation shorten iteration cycles, enabling continuous model improvement and rapid domain adaptation.

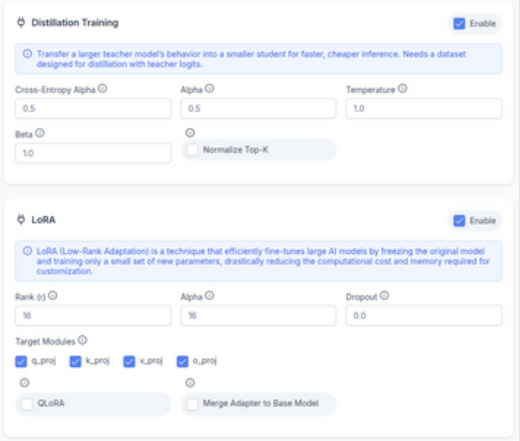


Model Alignment

DenseMAX Studio integrates advanced reinforcement learning techniques such as DPO, PPO, and GRPO to align models with organizational policies, ethical guidelines, and customer expectations. This ensures AI systems produce safer, more responsible outputs, reducing risks of harmful or non-compliant behavior.

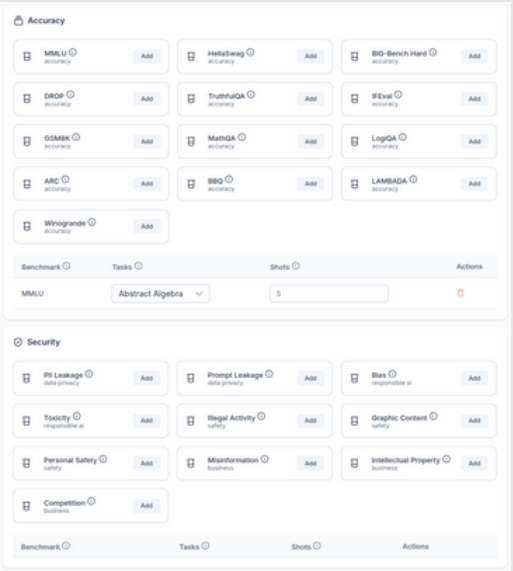
Model Distillation

Enterprises can create smaller, optimized models distilled from larger ones, making them cheaper, faster, and easier to deploy —ideal for edge environments or latency-sensitive applications. This empowers businesses to serve AI at scale without incurring excessive compute costs.



Comprehensive Model Evaluation

DenseMAX Studio includes an extensive evaluation suite, supporting benchmarks like MMLU, ARC, GSM8k, TruthfulQA, HellaSwag, and more. Beyond accuracy, it enables red-teaming to test vulnerabilities in areas such as toxicity, bias, misinformation, and sensitive data leakage. This ensures enterprise-grade reliability, fairness, and compliance.



Model Quantization

The platform offers tools to quantize models for the Blackwell GPU architecture, reducing memory footprint and improving inference performance without significantly sacrificing accuracy. This leads to higher throughput, lower costs, and more efficient GPU utilization, critical for scaling large workloads.

Let's bring AI home.

Hardware Specifications

Processors
- Dual AMD Epyc 9005, 9004 or 97x4 processors

Memory
- 24x DDR5 ECC RDIMM slots

GPU Acceleration
- 8× PCIe 5.0 x16 GPU slots
- supports up to 8× NVIDIA RTX 5090 or RTX PRO 6000 Blackwell GPUs

Storage
- 2× PCIe 3.0 x4 M.2 NVMe internal slots
- 4× U.2 PCIe 5.0 x4 NVMe slots

Storage
- 2× PCIe 5.0 x8 networking slots (can merge into 1× PCIe 5.0 x16)
- 1× 400 Gbps InfiniBand network card

Power & Cooling
- 5x 2000W industrial-grade, redundant PSU

I/O
- Front I/O with IPMI port
- 2× 1GbE RJ45 ports
- 1x DB15 VGA port
- 4 Type-A (USB3.2 Gen1)

Management Controller
- IPMI 2.0, Redfish, and Virtual Media support
- Out-of-band system monitoring (power, thermal, fan, voltage)
- Secure remote firmware updates and BIOS/UEFI configuration
- Hardware-assisted KVM-over-IP for complete remote console access
- Role-based access control and secure authentication protocols

Environmental
- Operating temperature: 10°C – 35°C
- Non-operating temperature: –40°C to 70°C
- Non-operating humidity: 20% – 90% (non-condensing)

Get The Infrastructure That Works for You

Whether you're modernizing operations, building internal copilots, or enhancing security posture, our appliance gives you speed, flexibility, and full control — all in one box.

Ready to Get Started?

To learn more, visit: invergent.ai