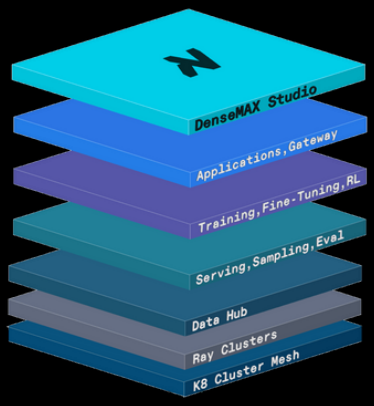
INVERGENT

Datasheet

# DenseMax Studio

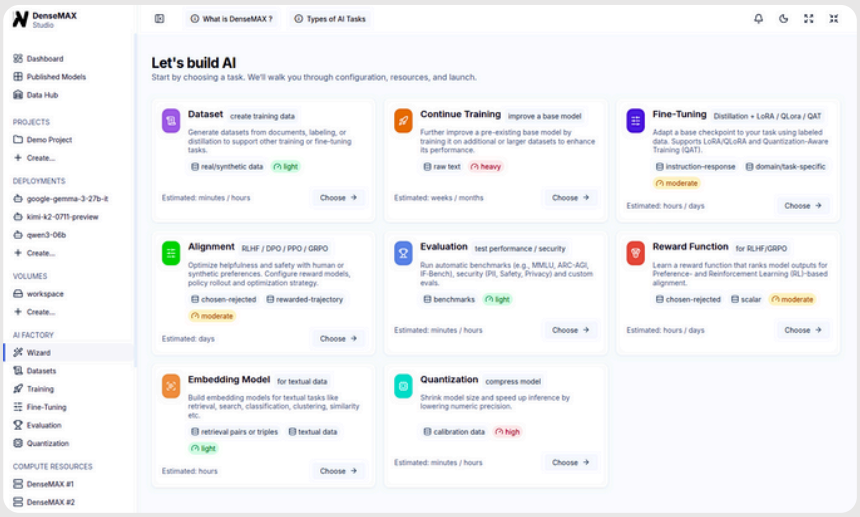
Professional AI for the enterprise




## Enterprise LLMOps

DenseMAX Studio is an enterprise-grade LLMOps platform built to accelerate the journey from AI experimentation to enterprise-scale deployment. By uniting state-of-the-art large language models with robust deployment, fine-tuning, and evaluation services, it empowers organizations to rapidly innovate, safeguard, and operationalize generative AI applications.

Designed with scalability, security, and observability at its core, DenseMAX Studio enables enterprises to move beyond pilots and prototypes—delivering production-ready AI solutions across industries.




### Key Features

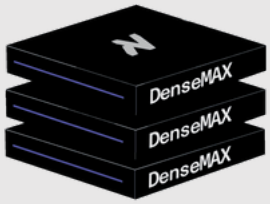
- **All-in-One AI Platform**
  Everything you need to build, deploy, and scale generative AI.
- **Fast Time-to-Value**
  Go from idea to production in days, not months
- **Continuous Innovation**
  Keep models fresh with easy fine-tuning and retraining.
- **Built-In AI Safety**
  Guardrails and alignment 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.

### Instant Deployment, No Complexity

- Deployment on DenseMAX Appliances or public clouds: AWS, Azure, GCP, Oracle Cloud


### Complete Data Sovereignty

- Keep sensitive data local — no reliance on HuggingFace, ModelScope or 3rd party repositories
- Strict GIT-like versioning policies model & dataset customization
- Role-based access control, audit logging, and optional air-gapped operation.



### Optimized CUDA kernels

- Custom software components built for specific GPU architectures 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.

### 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) with isolated, parallel model deployments.

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

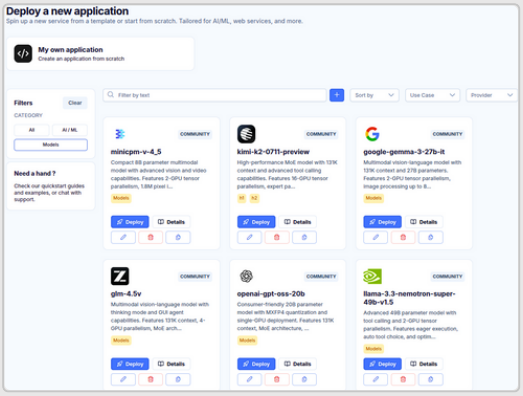
🧪 Experiment & Iterate Locally
Rapidly test prompts, tune configurations, and evaluate model behavior without reliance on cloud costs or vendor limitations.

💸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.

**Our present: The Future**

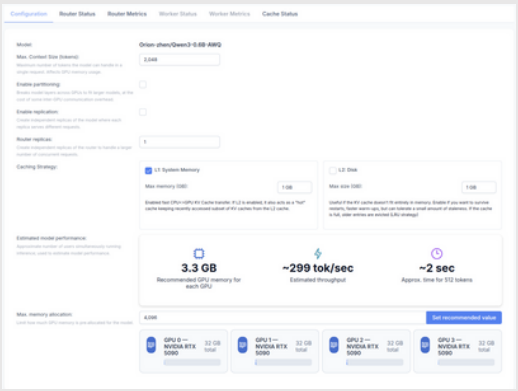
## Application Deployment & Serving

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.



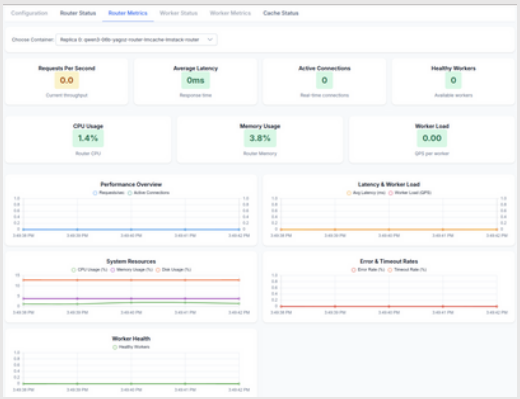

## Production-Grade Model Deployment

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.



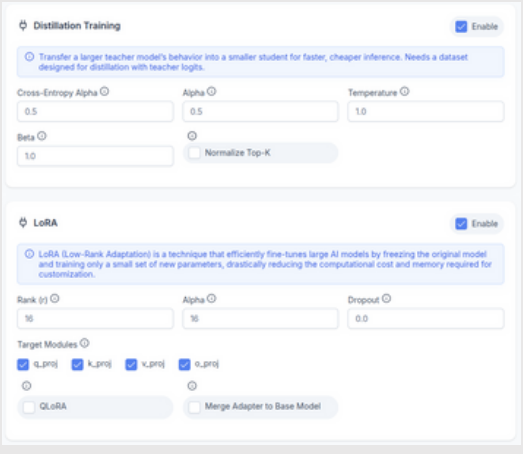

## Secure & Observable Model Serving

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.



## 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.

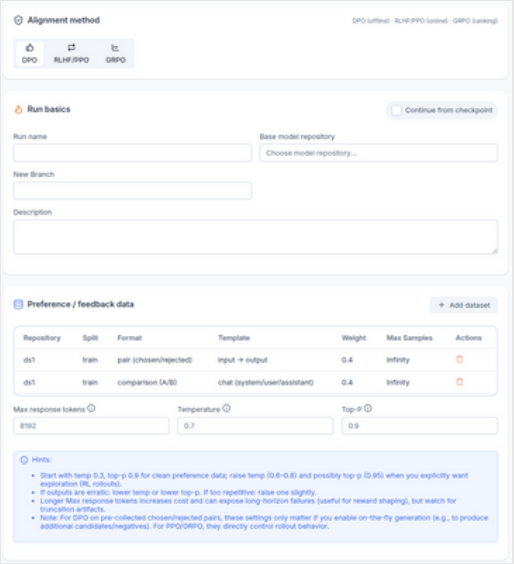



### Model Distillation

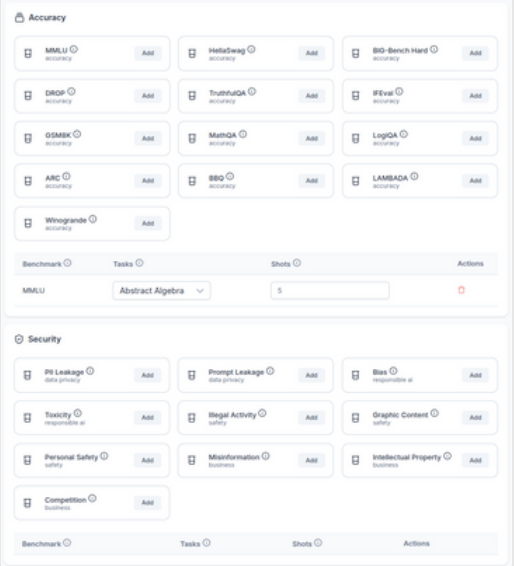
Enterprises can create smaller, optimized models distilled from larger ones, making them cheaper, faster, and easier to deploy—ideal for edge environments or latency-sensitive applications. This empowers businesses to serve AI at scale without incurring excessive compute costs.

**Our present: The Future**

**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.

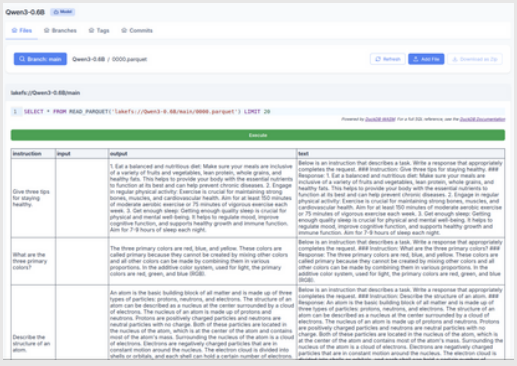



**Comprehensive Model Evaluation**
DenseMAX Studio includes an extensive evaluation suite, supporting benchmarks like MMLU, ARC, GSM8k, TruthfulQA, HellaSwag, and more. Beyond accuracy, it enables red-teaming to test vulnerabilities in areas such as toxicity, bias, misinformation, and sensitive data 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.

**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.




## Ready to Get Started?

To learn more, visit: invergent.ai



**Our present: The Future**