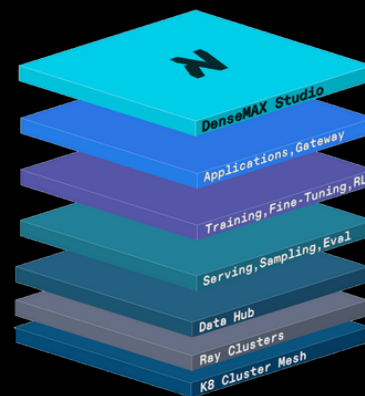


DenseMax Studio

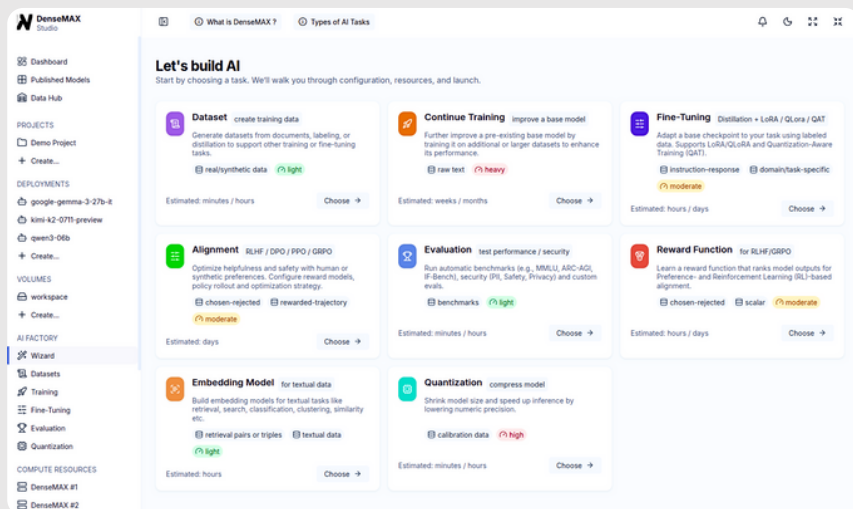
Professional AI for the enterprise



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

- Customize 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 assessments.
- 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

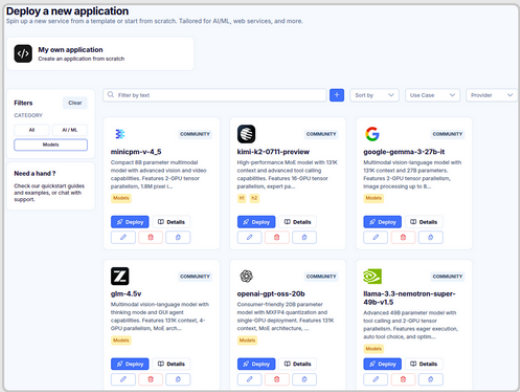
Avoid unpredictable usage-based cloud pricing. Run unlimited inference and fine-tuning workloads on a fixed-cost platform, eliminating API costs and reducing TCO over time.



Our present: The Future

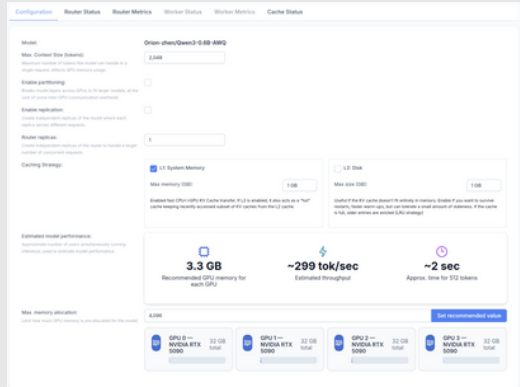
Application Deployment & Serving

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.



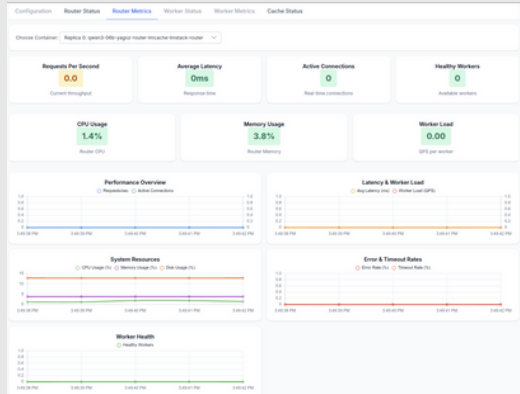
Production-Grade Model Deployment

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.



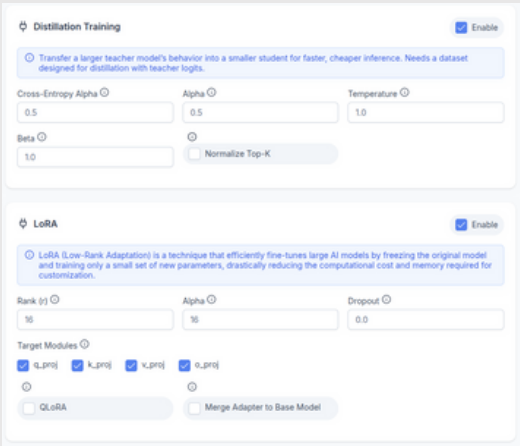
Secure & Observable Model Serving

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.

Alignment method DPO (left) RLHF/PPO (center) GRPO (right)

Run basics Continue from checkpoint

Run name Base model repository

New branch Choose model repository...

Description

Preference / feedback data + Add dataset

Repository	Split	Format	Template	Weight	Max Samples	Actions
ds1	train	pair (chosen/rejected)	input + output	0.4	Infinity	
ds1	train	comparison (A/B)	chat (system/user/assistant)	0.4	Infinity	

Max response tokens Temperature Top-p

8192 0.7 0.9

Hints:

- Start with temp 0.3, top-p 0.9 for clean preference data; raise temp (0.8-0.9) and possibly top-p (0.95) when you explicitly want exploration (e.g., red-team).
- If outputs are erratic, lower temp or lower top-p. If too repetitive, raise one slightly.
- Larger Max response tokens increases cost and can expose long-horizon behaviors (useful for reward shaping), but watch for truncation artifacts.
- Notes: For DPO we pre-collect chosen/rejected pairs, these settings only matter if you create on-the-fly generation (e.g., to produce additional candidates/negatives). For PPO/GRPO, they directly control rollout behavior.

Accuracy

MMLU accuracy Add	HellaSwag accuracy Add	BIO-Bench Hard accuracy Add
DROP accuracy Add	TruthfulQA accuracy Add	EVal accuracy Add
GSM8k accuracy Add	MathQA accuracy Add	LogQA accuracy Add
ARC accuracy Add	BBO accuracy Add	LAMBADA accuracy Add
Winogrande accuracy Add		

Benchmark Tasks Shots Actions

MMLU Abstract Algebra 5

Security

PI Leakage data privacy Add	Prompt Leakage data privacy Add	Box responsible ai Add
Toxicity responsible ai Add	Illegal Activity safety Add	Graphic Content safety Add
Personal Safety safety Add	Misinformation business Add	Intellectual Property business Add
Competition business Add		

Benchmark Tasks Shots Actions

Qwen3-0.8B Qwen

Files Branches Tags Commits

Branches Qwen3-0.8B / Qwen3.parquet Remove Push Revert to 0.8B

Index:Qwen3-0.8B.parquet

1 SELECT FROM READ_PARQUET('Index:Qwen3-0.8B.parquet') LIMIT 20

Powered by [DuckDB](#) for on-disk inference with the [DuckDB](#) connector

Instruction	Input	Output	Text
Give three tips for staying healthy.		1. Eat a balanced and nutritious diet. Make sure your meals are inclusive of a variety of fruits and vegetables, lean proteins, whole grains, and healthy fats. This helps to provide your body with the essential nutrients it needs to function at its best and can help prevent chronic diseases. 2. Engage in regular physical activity. Exercise is crucial for maintaining strong bones, muscles, and cardiovascular health. Aim for at least 150 minutes of moderate-intensity aerobic activity or 75 minutes of vigorous activity each week. 3. Get enough sleep. Getting enough quality sleep is crucial for physical and mental well-being. It helps to regulate mood, improve cognitive function, and supports healthy growth and immune function. Aim for 7-9 hours of sleep each night.	Below is an instruction that describes a task. Write a response that appropriately completes the request. [[{"instruction": "Give three tips for staying healthy.", "input": "", "output": "1. Eat a balanced and nutritious diet. Make sure your meals are inclusive of a variety of fruits and vegetables, lean proteins, whole grains, and healthy fats. This helps to provide your body with the essential nutrients it needs to function at its best and can help prevent chronic diseases. 2. Engage in regular physical activity. Exercise is crucial for maintaining strong bones, muscles, and cardiovascular health. Aim for at least 150 minutes of moderate-intensity aerobic activity or 75 minutes of vigorous activity each week. 3. Get enough sleep. Getting enough quality sleep is crucial for physical and mental well-being. It helps to regulate mood, improve cognitive function, and supports healthy growth and immune function. Aim for 7-9 hours of sleep each night."}]]
What are the three primary colors?		The three primary colors are red, blue, and yellow. These colors are called primary because they cannot be created by mixing other colors and all other colors can be made by combining them in various proportions. In the additive color system, used for light, the primary colors are red, green, and blue (RGB).	Below is an instruction that describes a task. Write a response that appropriately completes the request. [[{"instruction": "What are the three primary colors?", "input": "", "output": "The three primary colors are red, blue, and yellow. These colors are called primary because they cannot be created by mixing other colors and all other colors can be made by combining them in various proportions. In the additive color system, used for light, the primary colors are red, green, and blue (RGB)."}]]
Describe the structure of an atom.		An atom is the basic building block of all matter and is made up of three types of particles: protons, neutrons, and electrons. The structure of an atom can be described as a nucleus at the center surrounded by a cloud of electrons. The nucleus is an extremely small, dense, and positively charged region. It contains protons, which are positively charged particles, and neutrons, which are neutral particles with no charge. Both of these particles are located in the nucleus of the atom, which is at the center of the atom and contains most of the atom's mass. Surrounding the nucleus of the atom is a cloud of electrons. Electrons are negatively charged particles that are very light in mass compared to protons and neutrons. They orbit the nucleus in specific orbits, and each shell can hold a certain number of electrons.	Below is an instruction that describes a task. Write a response that appropriately completes the request. [[{"instruction": "Describe the structure of an atom.", "input": "", "output": "An atom is the basic building block of all matter and is made up of three types of particles: protons, neutrons, and electrons. The structure of an atom can be described as a nucleus at the center surrounded by a cloud of electrons. The nucleus is an extremely small, dense, and positively charged region. It contains protons, which are positively charged particles, and neutrons, which are neutral particles with no charge. Both of these particles are located in the nucleus of the atom, which is at the center of the atom and contains most of the atom's mass. Surrounding the nucleus of the atom is a cloud of electrons. Electrons are negatively charged particles that are very light in mass compared to protons and neutrons. They orbit the nucleus in specific orbits, and each shell can hold a certain number of electrons."}]]

Ready to Get Started?

To learn more, visit: invergent.ai

© 2025 Invergent Corporation. All rights reserved. Invergent, and DenseMAX and the Invergent logo are trademarks and/or registered trademarks of Invergent SA in the E.U. and other countries. All other trademarks and copyrights are the property of their respective owners.



Our present: The Future