

d.run Token 工厂效能平台

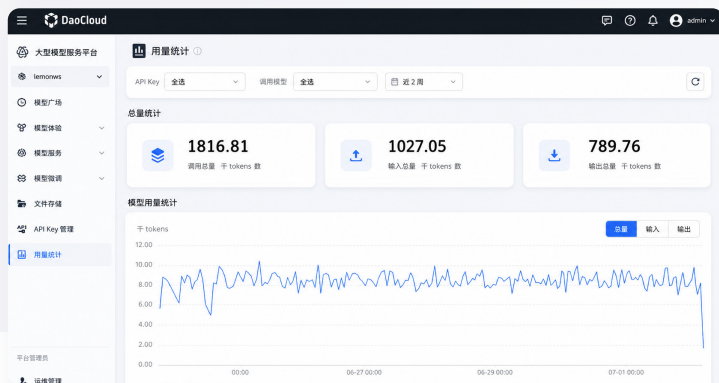
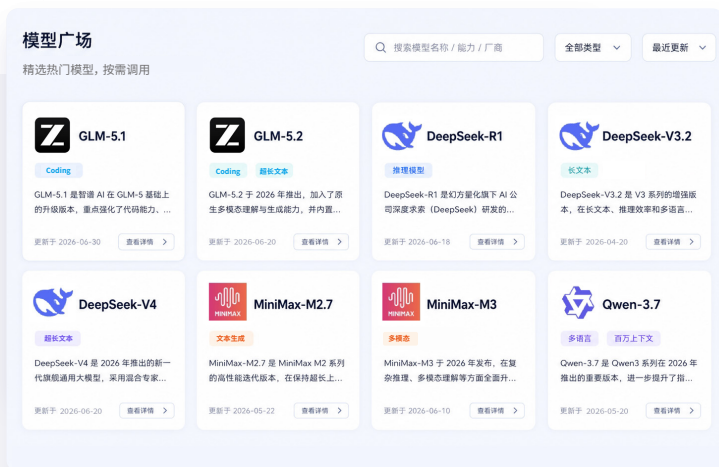
d.run Token Factory Platform

专为智算中心打造的高效能 AI Token 生产及经营系统

致力于将传统算力中心升级为高效、可盈利的 Token 工厂。平台统一纳管英伟达及国产异构算力，通过智能调度与推理优化，将分散的 GPU 算力转化为低成本、高稳定、可交易的标准化 Token 服务。依托 Token 工厂管家与统一运营体系，平台实现资源、生产、成本和供需的全局管理，服务终端用户、管理者、运营者与运维者，推动智算中心从“提供算力”升级为“生产和运营 Token”。

极致 用户体验

丰富的模型服务开箱即用，按 Token 灵活付费，低门槛享用高性能、稳定可靠的 AI 服务。



敏捷运管

Token 工厂管家

大模型驱动的专属工厂管家，智能分析、辅助决策、优化运营



运营者

精准运营，持续提升收益

管理者

掌握经营全局，辅助科学决策

运维者

主动运维，保障稳定运行

全局运营驾驶舱

一屏统览 Token 生产、消费、收益和资源状态



超十年大型数据中心
服务经验

推理提效

▶▶ 单卡 Token 产量更高

单位 GPU 产量:

2.78 万 tok/s*

相比行业优化

(GLM 5.1)

60% ↑

Token 产量

70.7% ↓

首 Token 响应

70% ↓

Token 生成时间

单卡投产比高达:

8 倍

(GLM 5.1、*含缓存)

相比优化前

24 倍 ↑

Token 产量

97.4% ↓

首 Token 响应

49.5% ↑

Token 产量

50.3% ↓

Token 生成时间

英伟达 GPU x DeepSeek R1

国产 GPU x Qwen3-30B

▶▶ 模型上新更快

新模型上生产

仅需 **0-2 天**

模型下载

GPU 占用数小时降为近 **0**

引擎优化 (vLLM/SGLang/Dynamo)

KV Cache 管理

P/D 分离

智能路由

推测解码

算子调优

智能调度

异构池化

GPU 共享/网络加速/高速存储

调度优化

感知调度/任务调度/分布式调度

摩尔线程
MOORE THREADS

Enflame
燃科科技

HYCON
申科海光

天数智芯
iluvatar Corex

NVIDIA

Ascend

META
沐曦

璧仞科技
BRIN TECHNOLOGY

(1) 行业机构数据来源: <https://inference.semianalysis.com/inference>

Token Factory



帮助智算中心打造高产能 Token 工厂

智算中心如何打破“重建设、轻运营”的僵局？[d.run Token 工厂效能平台](#)让智算中心不仅是“基建”，更是“效益中心”，通过精细化调度管理与极致推理优化，将复杂的异构算力转化为标准化、高价值的 Token 服务，在确保“建好、管好”的基础上，打通“用好、卖好”的最后一公里，通过打造可持续的营收闭环，让每一份算力都能转化为实实在在的经济收益。

01

润道星算

构建统一纳管的算力“蓄水池”，推动算力转化为高品质 Token 服务, 让广西-东盟地区的算力消费像用电用水一样方便快捷。

广西首个



以 Token 为计量单位的智算服务平台

1000P+

近 2000 卡



算力规模

02

厦门思明 算力联盟

实现市级智能算力“跨主体、跨架构、跨中心”的统一调度，提供“即取即用、按需付费”的普惠算力服务，提升了中心自身的运营与服务能力，推动当地“AI + 场景”的深度融合，为本地 AI 生态注入新活力。

7 倍



Token 输出效率提升 7 倍



03

PIXEL

专注于数据中心与系统
创新的日本上市企业

基于 d.run Token 工厂效能平台提供从基建到运营的端到端解决方案，帮助客户落地智算中心并快速上线 Token 服务，通过调度与推理优化实现 Token 提质增效，在竞争激烈的海外 AI 市场抢占先机。

海外 Token 工厂

