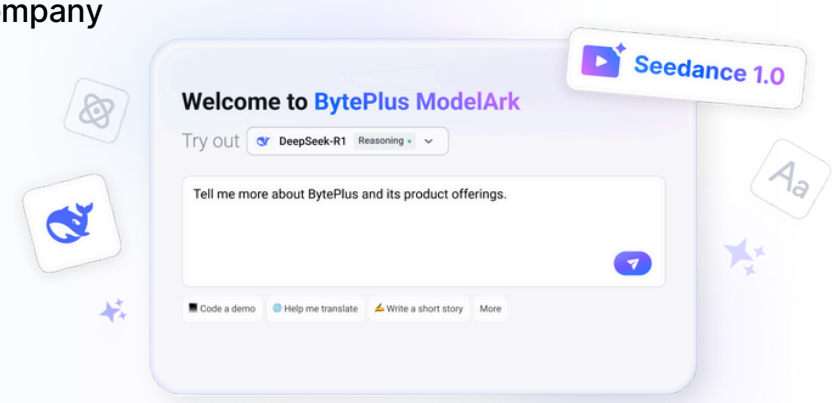
BytePlus A ByteDance Company

ModelArk

One-Stop LLM Platform



ModelArk is BytePlus's one-stop Model-as-a-Service (MaaS) platform, enabling enterprises to seamlessly integrate and scale generative AI capabilities. It provides a complete toolkit for high-performance inference, custom model finetuning*, and simplified application development.

ByteDance Proprietary Models

ModelArk offers ByteDance's proprietary LLMs and open-source models. Support a wide range of use cases with models built for performance, flexibility, and scale.

Pricing Advantage

ModelArk lowers your AI deployment costs with transparent pricing and built-in optimization tools. Features like Batch Inference and Prefix Caching reduce inference spend without compromising performance.

High-Performance Inference

ModelArk is built for enterprise-scale performance, delivering low-latency, high-throughput inference for seamless user experiences. Its optimized infrastructure ensures fast, efficient, and scalable execution.

Flexible Deployment Options

ModelArk accelerates your LLM deployment from weeks to hours with three powerful tools:

Risk-Free Testing

Try every LLM on BytePlus ModelArk with complimentary tokens included. Validate performance and capabilities before committing to development.

Rapid Deployment

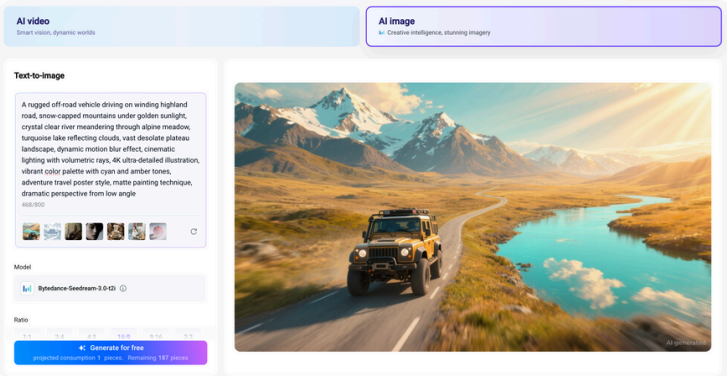
Launch production-ready models through our developer console. Get live faster than traditional implementations.

AI App Lab

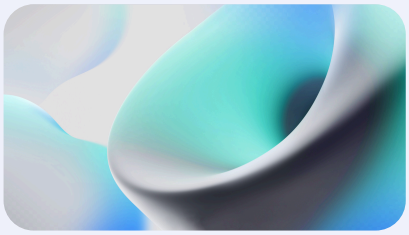
Build enterprise Generative AI applications with our open-source, easy-to-integrate solutions.

Featured Models

Most models on ModelArk offer live inference demos in our Playground, enabling users to experience capabilities like video generation and real-time reasoning firsthand.



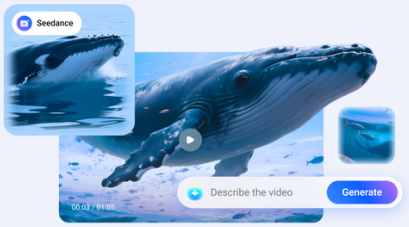
ByteDance Proprietary Models



ByteDance Seed 1.6

Enterprise-scale multimodal reasoning with predictable, low-cost pricing

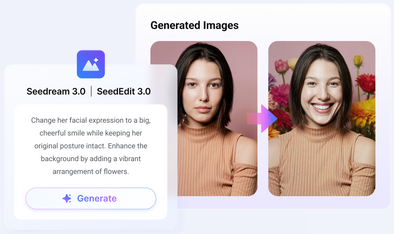
Seed 1.6, ByteDance's flagship LLM, provides advanced reasoning and superior visual understanding. It offers unique, flat-rate token pricing across all tasks and modalities.



Seedance 1.0

The World's Top-Ranked Generative Video Model

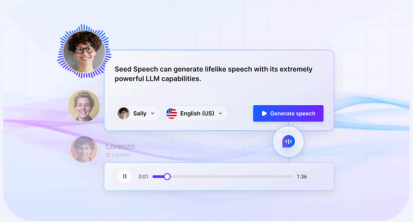
With native support for multi-shot sequencing, emotional cues, and dynamic camera motion, Seedance excels at following complex, detailed prompts involving multiple subjects and scene transitions.



Seedream

Turn Ideas Into Stunning Visuals

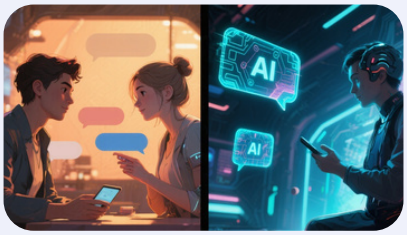
Top-ranked on the Artificial Analysis Image Arena Leaderboard, generates 2K images in seconds with emotional detail, typography control, and character retention, ideal for fast, narrative-driven visual creation across platforms.



Seed Speech

Redefining Voice Technology

BytePlus Voice empowers businesses to build an interactive voice agent that seamlessly merges advanced text-to-speech (TTS), voice replication, and speech recognition into a single, intuitive platform



Seed 1.6 SC

Immersive Chat Without Limits

Seed 1.6 SC is an advanced role-playing LLM engineered for interactive chat experiences. Characters stay consistently in role, delivering lifelike, emotionally grounded dialogue. The model was good at both immersive storytelling in long context and rapid-fire chatting scenarios, intelligently selecting the right characters to move the narrative forward.

Open Source Models

Most models on ModelArk offer live inference demos in our Playground, enabling users to experience capabilities like video generation and real-time reasoning firsthand.

GPT-OSS

Multilingual Open-Weight Reasoning

Available via the ModelArk API, GPT-OSS delivers strong real-world performance at low cost, outperforms similarly sized open models on reasoning tasks, demonstrates strong tool use capabilities.

Kimi-K2

Open Agentic Intelligence

K2 is a powerful open-source model by Moonshot AI, now available via the ModelArk API. Whether you're coding, building AI agents, or creating enterprise copilots, K2 helps you move faster with dependable function calling and advanced tool use.

DeepSeek-R1-0528

Refined open-source MoE model

DeepSeek-R1-0528 delivers sharper coherence, stronger logical flow, and greater controllability than its predecessors, making it highly effective for one-shot problem solving and complex, multi-step reasoning.

DeepSeek-V3-0324

Upgraded bilingual reasoning engine

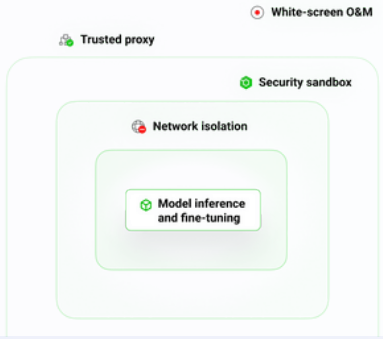
V3-0324 builds on DeepSeek-V3's bilingual foundation, offering robust support for Chinese and English tasks. It retains key features like long-context handling and function call support for real-world applications.

Enterprise-Grade Security

Your models and data remain completely secure, private, and under your control with ModelArk's comprehensive, multi-layered security framework.

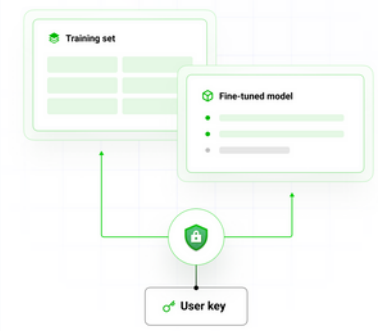
Model Protection

Safeguard your valuable LLMs from tampering and theft. Your models are protected through encrypted file systems, vArmor runtime protection to block malicious activities, and advanced network isolation to create impenetrable barriers around your assets.



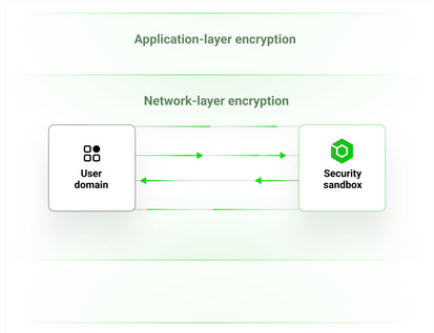
Session Data Security

Your data remains invisible to everyone except you. End-to-end encryption safeguards all session data, with decryption occurring only within secure sandboxes. A secure gateway enforces strict user authentication and access controls.



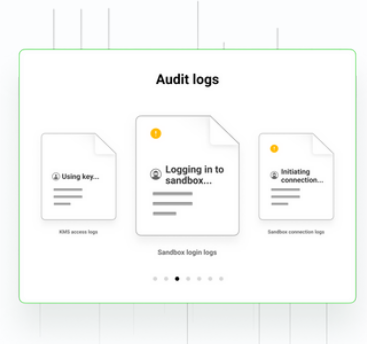
Environmental Isolation

Each deployment operates within robust isolation boundaries. Trusted container sandboxing, fine-grained network isolation, and white-screened operations create secure execution environments and prevent data breaches.



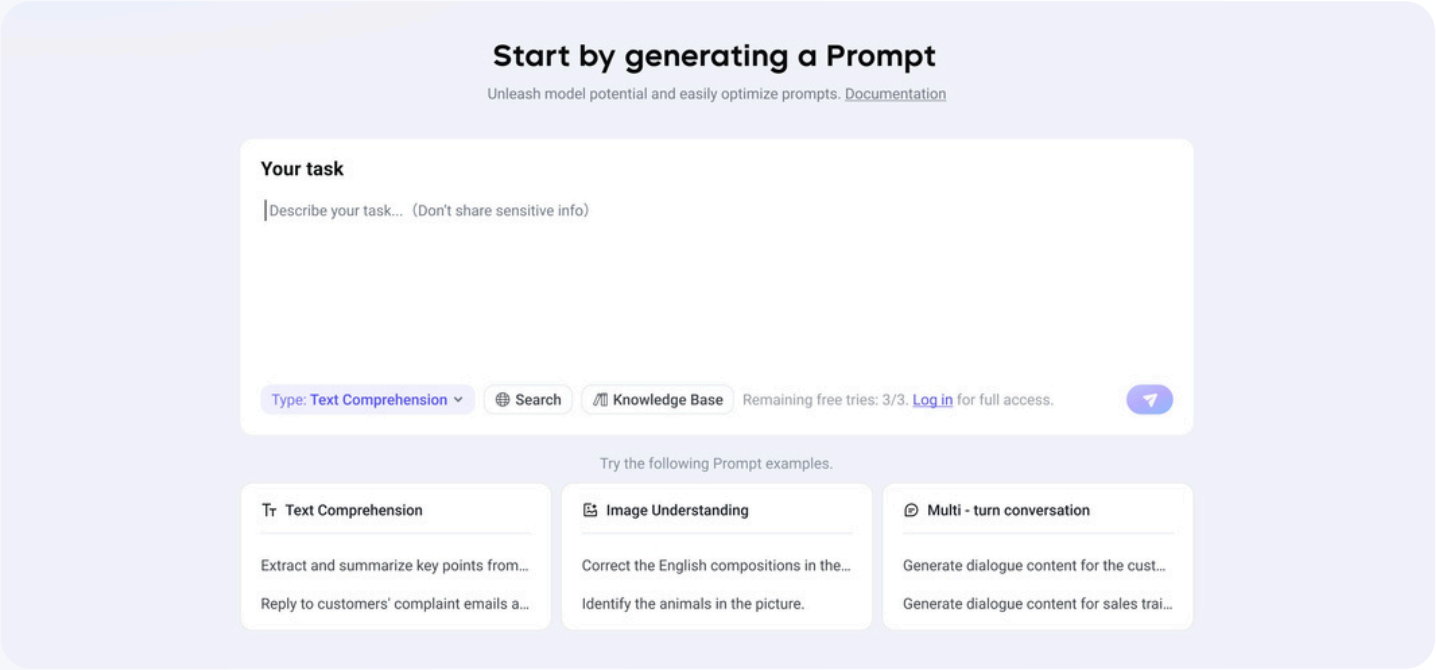
Complete Auditability

Gain full transparency and compliance tracking. Every action impacting your models and data is comprehensively logged and monitored, from sandbox logins to KMS access, verifying security policy effectiveness.



PromptPilot: Next-Generation Prompt Optimization

PromptPilot helps you easily build, optimize, and manage prompts, boosting your large language model performance across diverse use cases.



1. Interactive Goal Definition

Turn ideas into optimized prompts in minutes. PromptPilot guides you step-by-step, making complex prompt engineering feel easy.

2. Fully Automated Optimization

Let machines handle the heavy lifting. PromptPilot automatically refines your prompts using advanced algorithms for consistently high-quality outputs.

3. Cost-Efficient Iteration

Keep prompts optimized as models evolve and business needs change. Handle upgrades and shifting scenarios smoothly, saving on compute resources and development time.

4. Enterprise Prompt Management

Track, manage, and version every prompt via a user-friendly dashboard. Built-in tools for batch testing and automated scoring ensure quality at scale.

Pricing Advantage: Built for Economic Scalability

ModelArk's pricing is structured to maximize your return on investment with transparent, pay-as-you-go models and powerful cost-saving features. We empower you to manage and reduce your inference costs strategically as you scale.

Batch Inference

Process large volumes of data offline at a 50% discount compared to online inference. Ideal for non-real-time tasks, batch inference offers higher throughput and more generous quotas, allowing you to run large-scale jobs cost-effectively.

Intelligent Caching to Optimize Cost & Latency

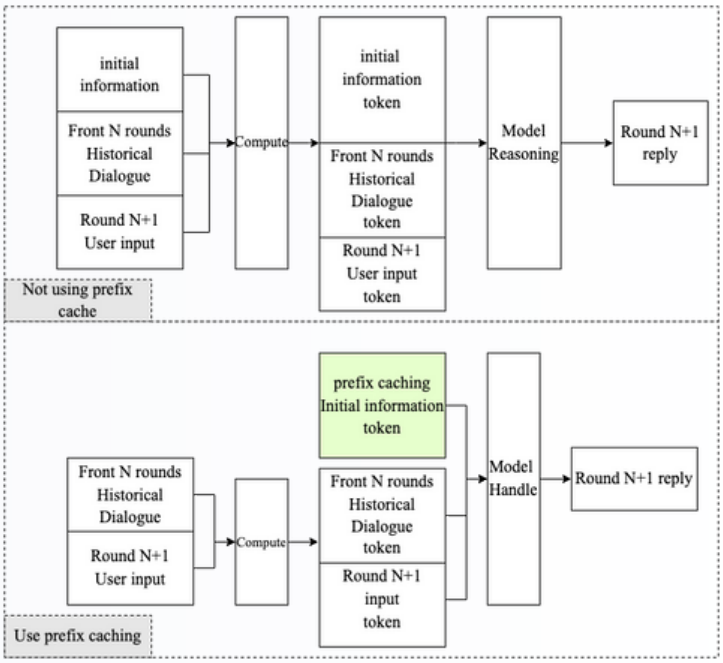
ModelArk provides two built-in caching strategies to reduce token usage and lower costs, particularly in high-frequency and multi-turn scenarios. Cache-hit requests can be up to 80% cheaper than standard token pricing, significantly reducing your compute spend.

Session Cache

Stores initial information and dynamically updates with each round of conversation. When processing requests, the cache content is combined with new inputs for the model to handle. Ideal for multi-turn dialogue scenarios such as interactive chats and multi-tool workflows.

Prefix Cache

Stores static prompt content that does not require updates across sessions. Perfect for standardized conversation openers, fixed task instructions, rule-based templates, and repeated use of long-text deep analysis prompts.



ModelArk provides two APIs for context caching: ContextAPI and ResponseAPI. ResponseAPI natively supports efficient context management, allowing you to enable or adjust caching with ease and optimize control without additional setup.
