Predibase

EBOOK

The Definitive Guide to Serving Open Source Models

Contents

Introduction **3**

Key Consideration 1: Reliability at Scale – Dynamic Resource Management and Autoscaling **4**

Autoscaling Done Right: Managing Spiky Traffic Without Over-Provisioning 6

Minimizing cold starts: from 14 minutes to 1 minutes 7

Seamless GPU Access Across Clouds: Predibase's Unified Solution 8

Key Consideration 2: Turbocharged Performance – Speeding Up SLM Inference **9**

Maximizing Model Efficiency: Why Performance Matters 9

Turbo LoRA + FP8: The Key to 4x Faster Throughput 10

Key Consideration 3: Cost-Efficiency without Compromise **14**

The Hidden Costs of Model Serving 14

Scaling Smarter with LoRA Exchange 14

Bonus Considerations: Enterprise-Grade Inference – Security, Observability, and Compliance **16**

Unlocking the Full Potential of AI with Optimized Inference Stacks **19**

Accelerate Innovation with Simplified AI Infrastructure **21**

Introduction

The Artificial Intelligence (AI) landscape has fundamentally transformed enterprise computing, creating an urgent need for efficient, scalable model deployment solutions. As real-time AI capabilities become essential for competitive advantage, organizations must master the delicate balance between performance and cost-effectiveness.

Small Language Models (SLMs) have emerged as the optimal choice for enterprise AI deployment, offering an unmatched combination of speed, efficiency, and adaptability. These models deliver faster inference times and simpler deployment compared to their larger counterparts, while maintaining exceptional performance through domain-specific fine-tuning that can be achieved without massive datasets. Our benchmarks consistently show that specialized SLMs can match or exceed the performance of larger models in targeted applications, with the added benefits of enhanced privacy and security compared to off the shelf LLMs.

However, many organizations face significant challenges when attempting to build their own inference infrastructure. The do-it-yourself approach often leads to spiraling costs, performance bottlenecks, and operational complexities that can derail AI initiatives. Common pitfalls include underestimating the resources required for AI deployment, failing to optimize systems for AI workloads, and neglecting the importance of dynamic resource management in spiky traffic context.

This ebook aims to provide crucial insights into achieving high reliability, performance, and cost-efficiency for SLM inference. By focusing on key components of a high-performance inference stack, we will guide you through the essential considerations for building a robust infrastructure that can support your AI ambitions while avoiding common mistakes. Whether you're just beginning your AI journey or looking to optimize existing infrastructure, this guide will equip you with the knowledge to make informed decisions and drive successful AI implementations in your organization.

KEY CONSIDERATION 1

Reliability at Scale – Dynamic Resource Management and Autoscaling

Common traffic patterns for customer-facing applications are characterized by spikes and dips in usage throughout the day and week. This inconsistent workload presents challenges for optimizing GPU usage. Unlike web servers, where autoscaling is a common and straightforward solution, GPU autoscaling for LLMs presents unique complexities:

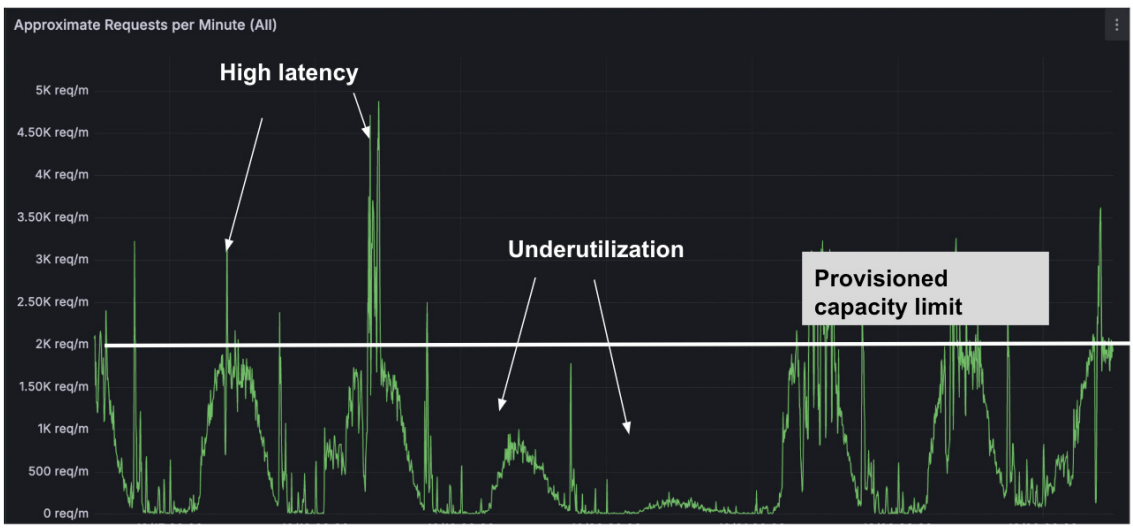
1. **LIMITED AVAILABILITY**
   Cloud providers often have restricted on-demand availability for GPUs, especially for the latest high-demand GPUs like H100s and H200s.

2. **INFLEXIBLE CONFIGURATIONS**
   Hyperscalers frequently offer rigid node setups. For instance, AWS only provides 8x GPU per node configurations for certain cards, making efficient scaling difficult.

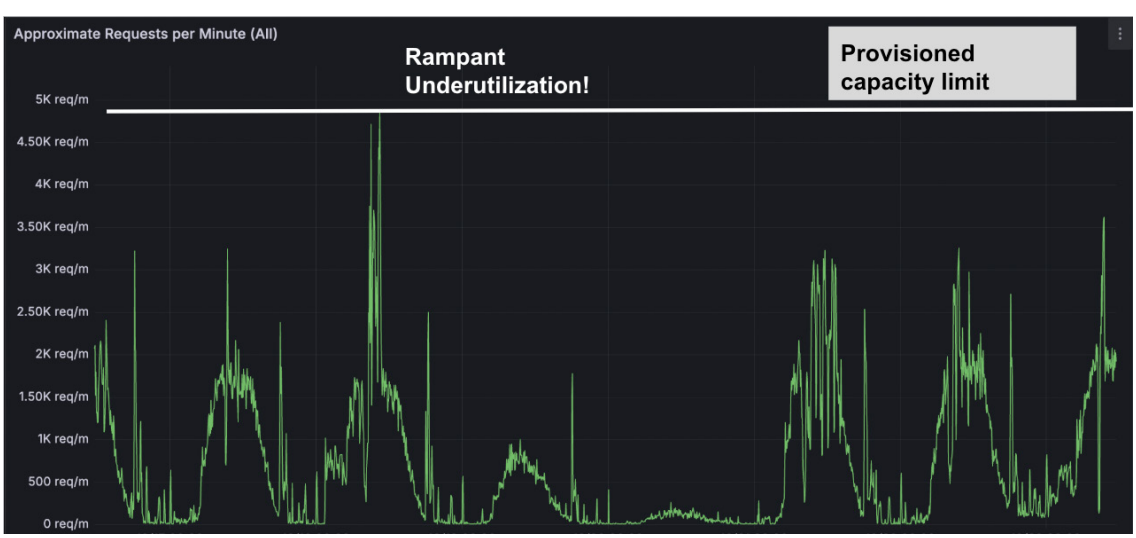
3. **COLD START ISSUES**
   Implementing autoscaling for LLMs is challenging due to the substantial data load required before operation. Model weights, often tens of gigabytes in size, need to be loaded, the GPU images (eg: docker container) themselves are very large (around 8-10 Gb) and need to be downloaded and finally the model needs to start. All these steps result in slow cold starts for new instances.

This lack of flexibility leads organizations to adopt two main provisioning strategies for GPUs:

- **Provisioning for average utilization:**
  This reduces costs but may lead to latency issues during peak times.



- **Provisioning for maximum capacity:**
  This ensures performance but results in idle GPUs during low-traffic periods. While this approach ensures constant availability, it can lead to significant waste during periods of low demand. This is a hidden cost that businesses often don't think about when they start using AI.



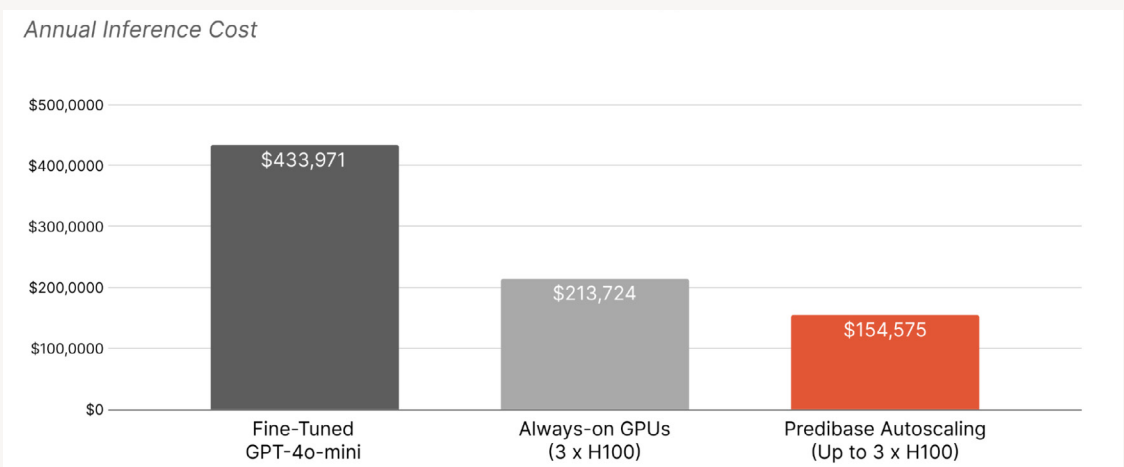
These factors combined make it challenging for organizations to implement efficient, cost-effective autoscaling solutions for LLM inference workloads.

Autoscaling Done Right: Managing Spiky Traffic Without Over-Provisioning

Predibase enables GPU autoscaling optimizations, leading to dramatic cost reductions that benefit both the company and its users. Our unified GPU autoscaling system intelligently manages resources across real-time inference workloads and batch jobs like fine-tuning, seamlessly scaling up during peak traffic and down during lulls.

During traffic spikes, the system can preempt lower-priority batch jobs to allocate GPUs for LLM replicas, ensuring rapid response to increased demand as explained in details in this webinar. This approach allows customers to benefit from additional replicas for burst usage at a very competitive price.

A compelling example illustrates the potential savings:
For an enterprise workload with specific usage patterns (12 peak hours per day, peak QPS of 50, off-peak QPS of 5, and input/output tokens of 1700/5), the cost difference is striking. A standard always-on deployment would cost over $213,000 per year. However, an autoscaling deployment reduces that to less than $155,000 annually — offering a savings of nearly $60,000 or 30%.

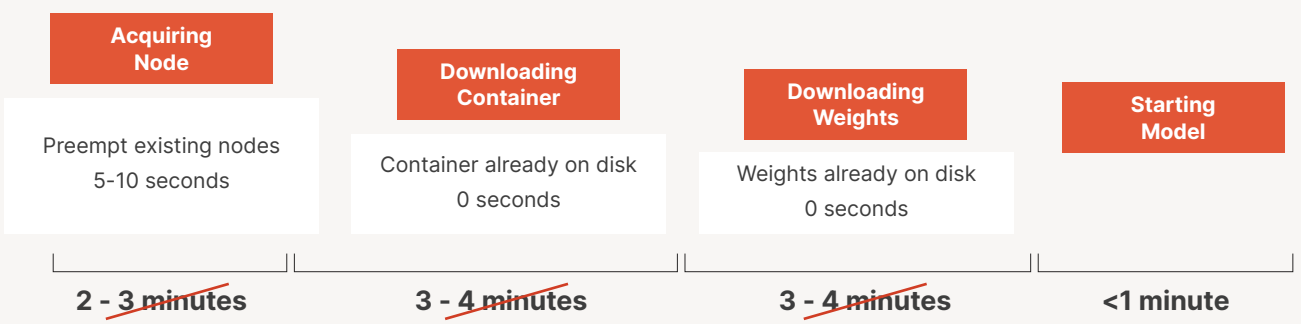


Predibase GPU Autoscaling unlocks significant cost reductions. (Assumptions: 12 peak hours per day, Peak QPS: 50, off peak QPS: 5, input/output tokens: 1500/S)

This example underscores the importance of optimized autoscaling in managing AI infrastructure costs. By dynamically adjusting GPU resources based on real-time demand, enterprises can achieve high performance without overpaying for idle infrastructure, making AI deployments far more cost-effective.

Minimizing cold starts: from 14 minutes to 1 minute

Predibase offers competitive pricing through innovative resource management and smart caching strategies. By keeping GPUs ready for immediate use and optimizing resources, we eliminate idle time and maximize efficiency. This intelligent infrastructure management allows us to provide high-performance AI capabilities at cost-effective rates, benefiting our customers. For enterprise customers, we maintain dedicated caches on their instances, storing model weights and LLM images. A cache manager ensures these resources remain available, allowing LLMs to start serving quickly when needed.



Predibase's smart caches bring down startup time from 9-12 minutes to less than 1 minute.

These optimizations have dramatically reduced the startup time for enterprise LLM replicas from 10-14 minutes to less than a minute in most cases.

Predibase's approach to autoscaling across bespoke clouds offers significant advantages over traditional hyperscalers:

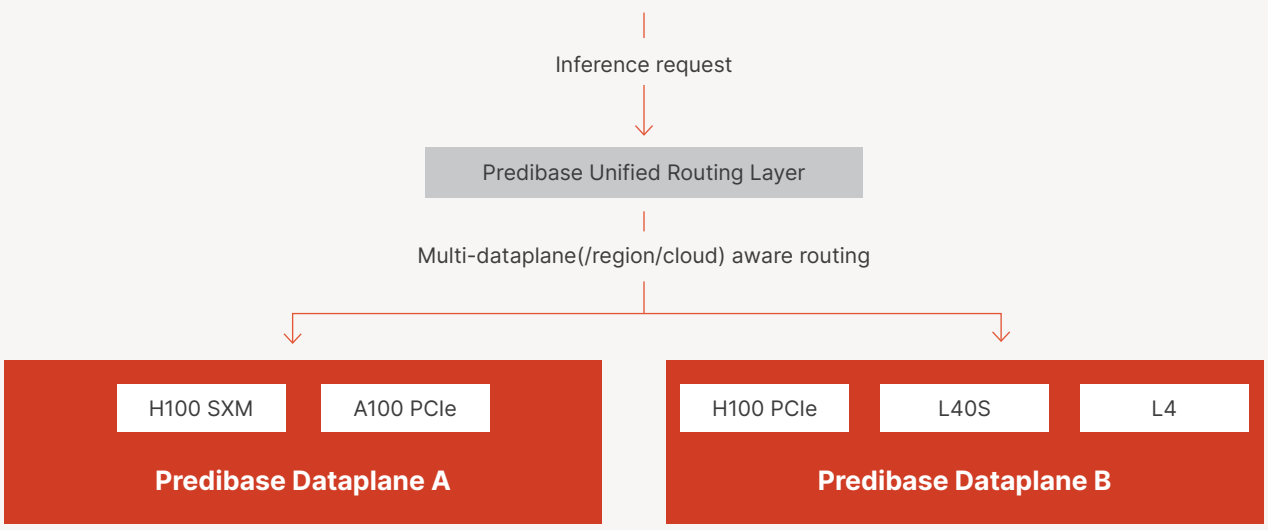
- **Faster provisioning**
  By building our own GPU autoscaler to work with AI-focused clouds that lack native autoscaling, Predibase can offer more flexible and cost-effective scaling options.
- **Reduced cold start times**
  The combination of advanced caching, preemptive scheduling, and optimized container images results in significantly faster resource provisioning compared to hyperscalers.
- **More granular control**
  Predibase's system allows for scaling at the individual GPU level, providing more precise resource allocation than the node-level scaling offered by most hyperscalers.
- **Cost-effective scaling**
  By leveraging bespoke clouds and optimizing resource allocation, Predibase can offer more cost-effective scaling solutions compared to traditional hyperscalers.

Seamless GPU Access Across Clouds: Predibase's Unified Solution

The landscape of GPU offerings is diverse and complex, featuring a wide array of options like A10Gs, A100s, and H100s, as well as various configurations such as SXM for optimized multi-GPU communication and PCIe for standard setups. However, the GPU market presents challenges, as no single cloud provider offers the full spectrum of options, and availability can be inconsistent, particularly for high-demand models.

To address these issues, Predibase has developed an innovative multi-dataplane architecture with a unified routing layer. This approach enables access to a wide range of GPU types across multiple cloud providers, offering flexible configurations including single-GPU options often unavailable from major providers. The system ensures higher availability by tapping into multiple sources and continuously integrates new GPU models as they become available in the market.

Predibase's unified routing layer seamlessly directs workloads to the most appropriate GPU resources, regardless of their cloud origin. This strategy maximizes resource availability and allows customers to benefit from the best offerings of each cloud provider without the complexity of managing multiple cloud relationships. By leveraging this multi-cloud approach, Predibase ensures that AI workloads always have access to optimal GPU resources, effectively balancing performance, cost, and availability to meet specific customer needs.



Our architecture integrates GPUs from multiple cloud providers into a single, unified compute pool. This approach ensures seamless access to diverse GPU types, enabling customers to create deployments and run training jobs efficiently, regardless of their specific requirements.

In conclusion, Predibase's unified GPU autoscaling system, coupled with its multi-cloud and multi-region deployment capabilities, offers a robust solution for managing the challenges of fluctuating demand in AI workloads. By minimizing cold starts and leveraging bespoke cloud solutions, Predibase provides a reliable, scalable, and cost-effective platform for enterprise AI deployments.

KEY CONSIDERATION 2

Turbocharged Performance — Speeding Up SLM Inference

While infrastructure and engineering optimization lay the foundation for efficient SLM deployments, Predibase's excellence in inference goes far beyond these aspects. This chapter explores three groundbreaking innovations that not only boost SLM performance but also drive down costs as a significant added benefit. By leveraging state-of-the-art LoRA, Speculative decoding and FP8 Quantization, Predibase delivers unparalleled speed, efficiency, and cost-effectiveness.

Maximizing Model Efficiency: Why Performance Matters

Latency and throughput are critical factors in AI-powered applications, directly impacting user experience and business outcomes. Low latency ensures quick responses and user satisfaction, while high throughput enables handling of multiple requests concurrently, essential for enterprise-scale workloads.

How latency impacts user experience and business outcomes

Low latency ensures that users receive quick responses from AI-powered applications, leading to a smoother and more engaging experience. For interactive applications like chatbots or interactive learning applications, even small delays can frustrate users and negatively impact their perception of the service quality.

Moreover, faster response times can directly translate to improved business metrics. For example, in e-commerce, quicker product recommendations can lead to higher conversion rates. In financial trading, ultra-low latency can provide a competitive edge by executing trades faster.

The consequences of high latency extend beyond mere inconvenience. Slow response times can disrupt the flow of user tasks and interactions, impeding productivity and hindering the achievement of goals. This frustration can tarnish the user's overall perception of the application, leading to decreased confidence and loyalty. In a competitive digital landscape, users are likely to explore alternatives that offer a smoother and faster experience, resulting in potential loss of business and decreased user retention. For businesses, the impact of latency on user experience translates directly to bottom-line results. Lower engagement rates, higher bounce rates, and reduced customer satisfaction can all stem from poor model performance, ultimately affecting revenue and market position.

Beyond mere user satisfaction, many applications, such as autonomous vehicles or fraud detection systems, require near-instantaneous responses to function effectively and safely. Ultra-low latency opens up possibilities for applications that were previously infeasible, such as real-time language translation in video calls or augmented reality experiences.

Why throughput matters for enterprise-scale workloads

While latency focuses on individual request performance, throughput measures the system's ability to handle multiple requests concurrently. For enterprise-scale workloads, high throughput is essential to meet the demands of numerous users and complex applications simultaneously. Throughput — typically measured in tokens per second or requests per second — provides insight into how efficiently a model can process large volumes of data. High throughput ensures that applications can scale to meet demand, maintain responsiveness under heavy loads, and maximize the utilization of computing resources.

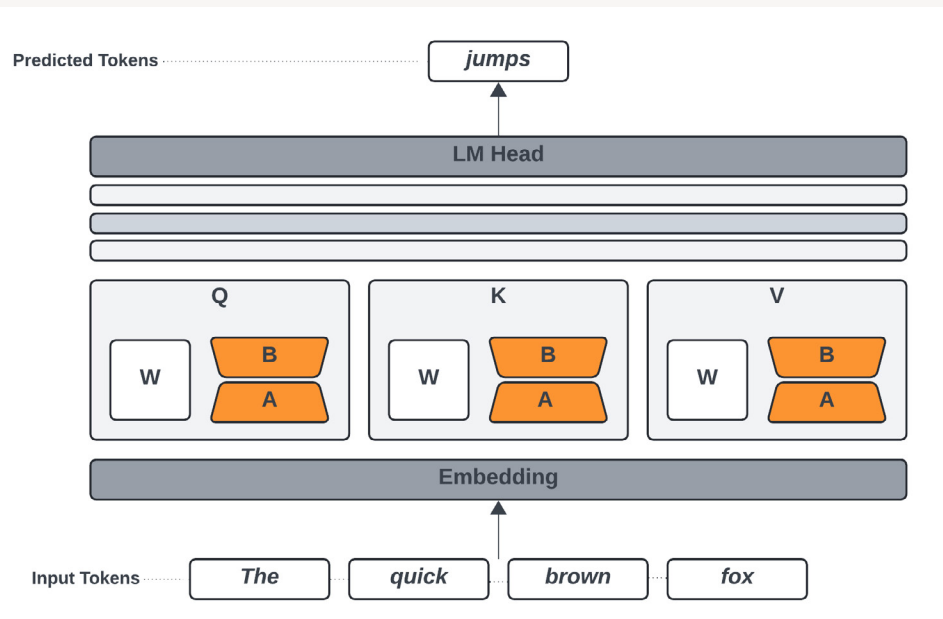
For businesses, optimizing throughput translates to improved operational efficiency and the ability to serve a larger user base without compromising performance. Low latency allows AI systems to handle more requests in a given timeframe, maximizing resource utilization and potentially reducing infrastructure costs. By reducing processing time and optimizing resource usage, low latency can lead to decreased operational expenses and improved overall cost-effectiveness of AI deployments

Turbo LoRA + FP8: The Key to 4x Faster Throughput

To address the dual challenges of latency and throughput, Predibase has developed innovative techniques that dramatically improve SLM inference performance: Turbo LoRA and FP8 quantization. When combined, these approaches can deliver up to 4x faster throughput for SLM inference.

Introduction to Low-Rank Adaptation (LoRA)

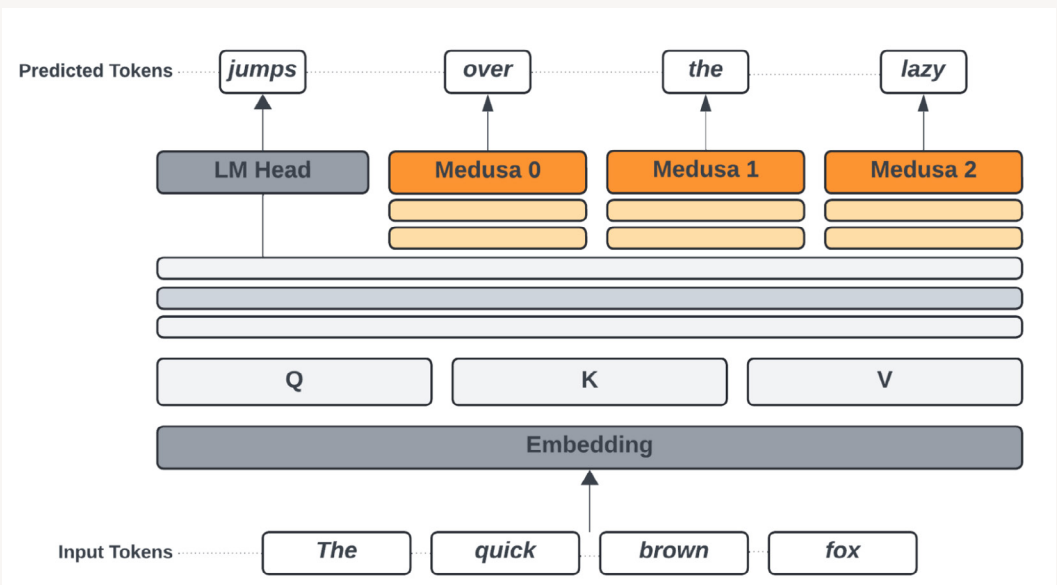
Traditional fine-tuning methods often require adjusting billions of parameters, which is computationally expensive and resource-intensive. Low-Rank Adaptation (LoRA) offers a more efficient alternative by introducing a low-rank matrix decomposition. By utilizing these low-rank updates, LoRA dramatically reduces the number of trainable parameters (typically 0.2%). This not only speeds up the fine-tuning process but also significantly decreases the memory requirements, making it possible to adapt large models on more modest hardware.



LoRA: Efficient training on specialized datasets for higher quality results

Speculative Decoding: Predicting multiple tokens in a single pass without compromising quality

Speculative decoding is a method that speeds-up next-token generation by accurately predicting several tokens ahead without the need for complex verification strategies. One popular speculative decoding method is called Medusa. However, while Medusa speeds up token generation, the additional speculators are quite large (>1Gb) and as a result they are generally fine-tuned to be quite generic to reduce the number of heavy deployed models.



Medusa: Fast token generation through speculative decoding but with heavy Medusa heads (>1Gb)

Turbo LoRA: combining speculative decoding with LoRA

Turbo LoRA represents a breakthrough in fine-tuning techniques by merging the strengths of speculative decoding and LoRA. While traditional LoRA adapters enhance model quality, they often sacrifice throughput due to the sequential nature of token generation in LLMs. Turbo LoRA overcomes this limitation through two key innovations:

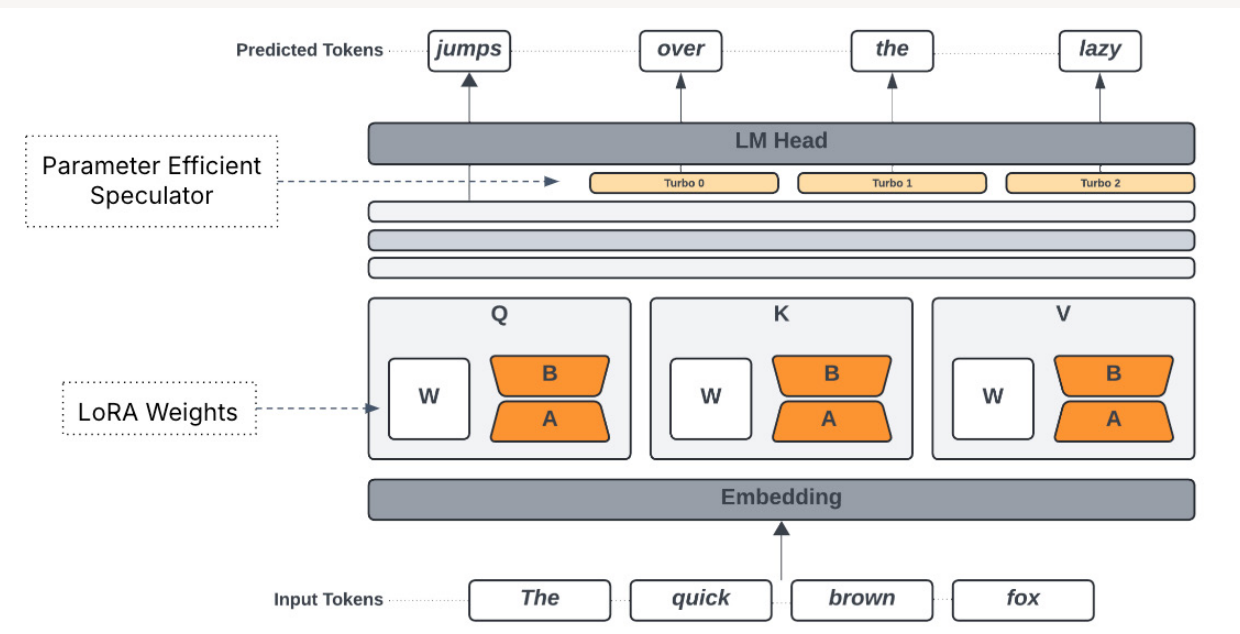
- **Joint Training**
  Turbo LoRA simultaneously trains both the LoRA adapter and a lightweight speculation adapter. This approach enables highly accurate multi-token predictions in a single pass, significantly boosting throughput. Note that the LoRA and Turbo adapters can be trained independently, allowing users to enhance existing LoRA adapters with Turbo capabilities without retraining from scratch. This flexibility enables users to train Turbo adapters while keeping the original model and LoRA weights frozen.
- **Memory Efficiency**
  The lightweight speculation adapter uses significantly less memory (tens of megabytes) compared to other methods like Medusa heads (gigabytes), allowing for more efficient resource utilization.

Thanks to this parameter efficiency, Turbo LoRA not only works well on small batch sizes but also on large batch sizes as opposed to traditional speculative decoding which is often criticized for not performing well on large batch sizes.

By combining these features, Turbo LoRA offers a powerful solution to reach both:

- **Higher throughput**
  Turbo LoRA increases text generation throughput by 2-3x compared to both standard LoRA and the base model, across different GPU types and batch sizes (see graph below).

- **Better Quality**
  Empirical results show that Turbo LoRA maintains or even improves response quality compared to standard LoRA fine-tuning, despite the speed improvements.



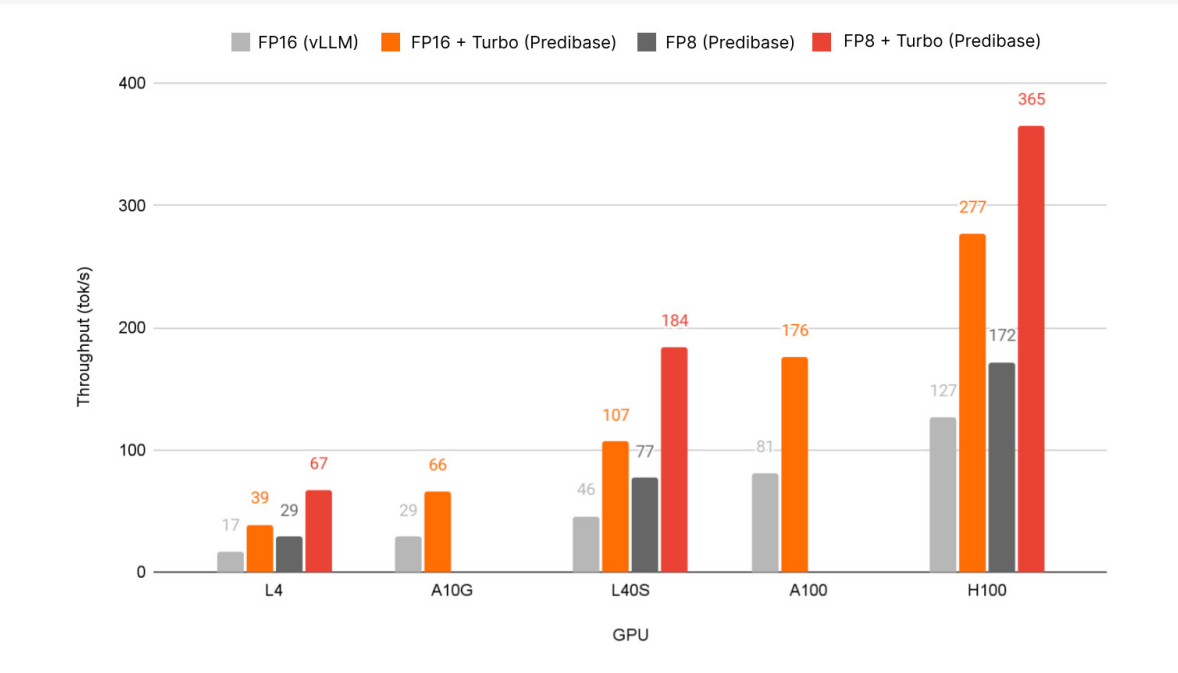
Turbo-LoRA: Parameter efficient speculation for lightweight high quality and fast token generation

FP8 Quantization: Reducing the memory footprint by 50%, boosting throughput, and minimizing costs

Complementing Turbo LoRA, FP8 quantization further boosts performance by reducing the model's memory footprint. FP8, or 8-bit floating-point representation, offers a balance between precision and efficiency. By quantizing weight activations to FP8, models can achieve:

- A 50% reduction in memory usage, allowing larger models to fit on a single GPU or enabling longer context lengths.
- Up to 2x improvement in inference throughput latency (ITL) for dense models and 1.6x for Mixture of Experts (MoE) models.
- As much as 3x throughput improvement in scenarios where memory savings enable increased batch sizes.

FP8 quantization is particularly effective for LLMs, as it halves the required memory bandwidth while nearly doubling compute speed compared to FP16. This is achieved without significant accuracy degradation, thanks to FP8's non-uniform distribution which is more robust to outliers compared to integer quantization methods.



Turbo LoRA benchmarks vs. vLLM: While the Turbo LoRA implementation more than doubled the throughput compared to vLLM, a further increase is achieved with FP8 with 2x increase on L4.

The combination of Turbo LoRA and FP8 quantization creates a powerful synergy. Turbo LoRA's ability to predict multiple tokens efficiently is further enhanced by the reduced memory footprint and increased computational speed offered by FP8 quantization. This allows for even higher throughput and lower latency, especially for enterprise-scale workloads with high concurrency requirements.

Turbo LoRA vs. vLLM: Breaking Through Bottlenecks

For many teams, vLLM is a natural starting point due to its open-source accessibility and maturity. However, as workloads scale, bottlenecks in throughput and reliability emerge, limiting its effectiveness for enterprise-scale applications. Turbo LoRA, by contrast, is purpose-built for high-speed, reliable inference, making it the preferred choice for production environments.

By leveraging these advanced techniques, businesses can dramatically improve their SLM inference performance, enabling faster response times, higher throughput, and more cost-effective deployments. Unlike vLLM, which often struggles with throughput bottlenecks and reliability issues at scale, Predibase's Turbo LoRA and inference optimizations are purpose-built for production AI.

Turbo LoRA's parameter-efficient design, combined with advanced speculative decoding and FP8 quantization, ensures 2-4x faster performance while maintaining consistent quality and stability under dynamic workloads. This not only enhances user experience but also allows companies to handle larger workloads with existing infrastructure, minimizing operational risks and maximizing the return on their AI investments.

KEY CONSIDERATION 3

Cost-Efficiency without Compromise

As enterprises increasingly adopt AI technologies, the costs associated with model serving have become a significant concern.

The Hidden Costs of Model Serving

To manage costs effectively, companies often opt for multiple specialized SLMs tailored to specific tasks, rather than relying on a single large, expensive, general-purpose AI model. As discussed in Chapter 1, implementing GPU autoscaling can significantly reduce expenses. However, another frequently overlooked cost factor is the traditional practice of deploying each model on a dedicated GPU. This approach, especially in "always-on" GPU setups, can quickly become prohibitively expensive. However, even with GPU autoscaling, dedicating entire GPUs to individual SLMs clearly results in costly resources.

Scaling Smarter with LoRA Exchange

LoRA Exchange is a cutting-edge open-source serving infrastructure designed to address the challenges of deploying multiple fine-tuned SLMs efficiently. Unlike traditional methods that require each fine-tuned model to run on dedicated GPU resources, LoRA Exchange allows organizations to serve hundreds of fine-tuned SLMs on a single GPU, drastically reducing costs.

> LoRA Exchange optimizes GPU memory usage and maintains high throughput for concurrent requests through
> several key techniques:
>
> 1. **DYNAMIC ADAPTER LOADING:** LoRA Exchange can load and unload fine-tuned adapters on demand, allowing for efficient use of GPU memory.
> 2. **TIERED WEIGHT CACHING:** This approach offloads adapter weights from GPU → CPU → disk ensuring that frequently used model weights are readily available, reducing latency.
> 3. **MULTI-ADAPTER BATCHING:** LoRA Exchange can process requests for multiple fine-tuned models in a single batch, maximizing GPU utilization.

The real-world impact of LoRA Exchange is significant. By enabling the serving of hundreds of fine-tuned SLMs from a single GPU, it dramatically reduces infrastructure costs. This capability is particularly beneficial for businesses that need to deploy various specialized models without the overhead of dedicating a GPU to each model.

Convirza's success story demonstrates how Predibase's Inference Engine and LoRA Exchange technology can enable companies to efficiently scale their AI operations, handling multiple models and variable workloads while maintaining high performance standards.

By minimizing GPU costs through efficient batching and dynamic adapter loading, LoRA Exchange enables enterprises to scale their AI operations without proportionally scaling their infrastructure costs. This innovative approach allows organizations to maintain high performance while significantly reducing the number of GPUs required, leading to substantial cost savings.

Convirza's Success with Predibase

CHALLENGE:

- Extremely variable workload with traffic spikes
- Need to scale up to double-digit A100 GPUs to maintain performance
- Requirement to serve multiple fine-tuned models efficiently
- Desire to keep average response times under 2 seconds

KEY RESULTS:

- Successfully serving 60 fine-tuned adapters concurrently
- Consistently achieving average response times under 2 seconds
- Efficiently handling high-volume, variable workloads
- Eliminated need to build and maintain complex infrastructure in-house

BENEFITS:

- **Scalability:** Able to handle traffic spikes requiring double-digit A100 GPUs
- **Efficiency:** Serving multiple models (60 adapters) on shared infrastructure
- **Performance:** Maintaining low latency (under 2 seconds) despite high load
- **Reliability:** Meeting demands of mission-critical, high-volume operations
- **Cost-effectiveness:** Avoiding the overhead of building custom infrastructure

convirza

COMPANY:
Convirza

INDUSTRY:
Call tracking and conversation analytics

SOLUTION:
Predibase Inference Engine with LoRA Exchange

> "The Predibase Inference Engine and LoRA Exchange allow us to efficiently serve 60 adapters while consistently achieving an average response time of under two seconds. Predibase provides the reliability we need for these high-volume workloads."

Giuseppe Romagnuolo
VP OF AI AT CONVIRZA

BONUS CONSIDERATIONS

Enterprise-Grade Inference – Security, Observability, and Compliance

This chapter explores the critical aspects that define an enterprise-ready platform. These include addressing the complex needs of large-scale deployments by offering comprehensive observability, enhancing security, and supporting compliance.

An effective enterprise platform integrates advanced observability tools to provide real-time insights, implements industry-standard security best practices to safeguard data and operations, and includes features designed to simplify and streamline compliance efforts. By combining these elements into a unified solution, the platform ensures scalability, reliability, and ease of deployment for enterprise-grade applications.

When evaluating a platform, there are a number of additional capabilities one should consider:

Deploying in VPC

While public cloud deployments offer flexibility and scalability for many organizations, some enterprises require tighter security controls or have specific infrastructure requirements that necessitate private deployments. For these companies, Predibase offers a powerful alternative: our inference engine can be deployed within your own Virtual Private Cloud (VPC).

This VPC option allows you to benefit from the advanced features of our inference engine while keeping your data and models within your own secure environment. This approach is ideal for organizations with stringent data privacy regulations or those in highly regulated industries.

Additionally, VPC deployments can be an attractive option for companies with existing cloud spend commitments. By leveraging Predibase's inference engine within your own infrastructure, you can utilize these pre-existing investments efficiently, potentially reducing overall costs while still accessing cutting-edge AI deployment capabilities.

SOC 2 Type 2 Compliance

For enterprises handling sensitive customer data, SOC 2 Type 2 compliance is essential. This standard ensures robust controls for security, availability, confidentiality, and privacy.

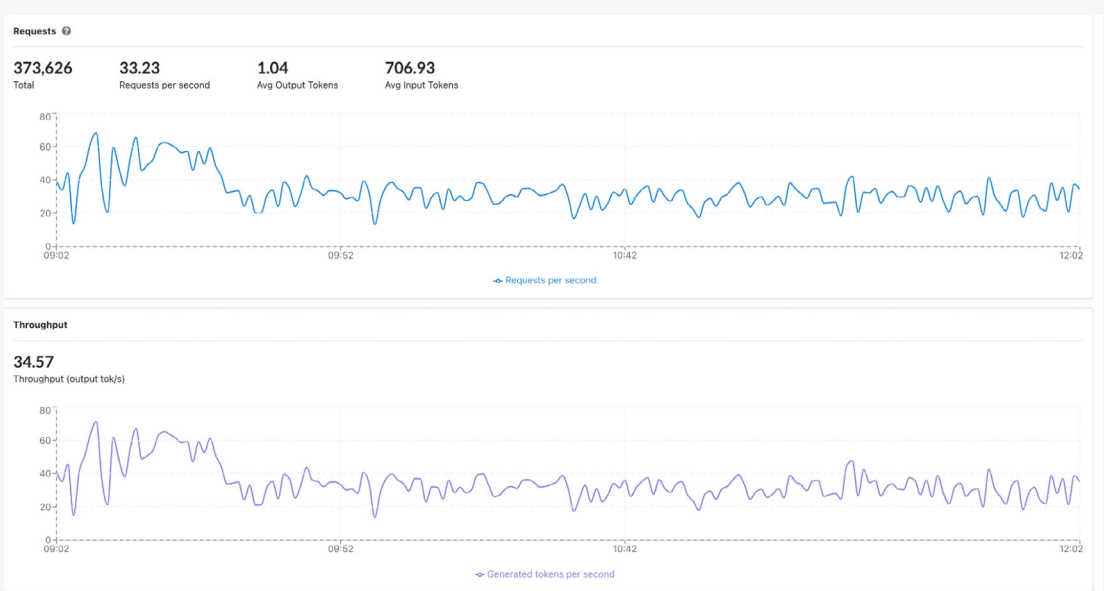
Achieving SOC 2 compliance requires:

- **Access Control:** Strict IAM policies to limit access.
- **Encryption:** Secure data in transit and at rest.
- **Monitoring:** Detailed activity logging.
- **Audits:** Regular reviews to ensure adherence.
- **Incident Response:** Plans to handle breaches effectively.

A compliant platform must prioritize security and provide tools like model export capabilities to maintain control over intellectual property while meeting these rigorous standards.

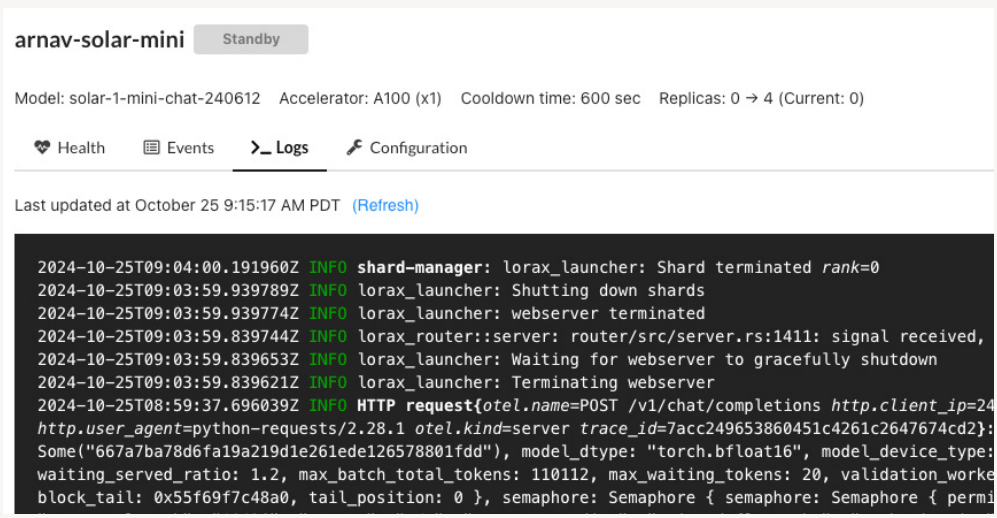
Observability and Real-Time Insights

For enterprise-scale AI deployments, observability and real-time insights are essential to ensure performance, reliability, and cost-efficiency. Monitoring key metrics like **requests**, **throughput, latency**, and **GPU utilization** allows organizations to optimize resource use, identify bottlenecks, and meet performance SLAs.



In-app metric graphs to monitor key indicators like request volume or throughput

An enterprise-grade platform must provide tools for real-time monitoring and analytics, including event logs, metric graphs, and deployment health dashboards. These features empower businesses to maintain smooth, high-performance operations while minimizing costs and ensuring infrastructure scalability.



In-app logs

Rolling Updates and Blue-Green Deployments

Rolling updates enable smooth and reliable updates to LLM deployments without downtime. By gradually replacing old replicas with new ones after passing health checks, this process ensures continuous availability, with automatic rollbacks in case of issues.

Continuous Health Checks

Continuous health checks monitor the health of all LLM replicas to maintain system performance and uptime. Failed replicas are automatically isolated, restarted, and reintegrated, ensuring minimal downtime and responsive deployments.

Resilient Request Handling

Network-level retries automatically recover from transient network failures at the load balancer level. Carefully configured to avoid overload and cascading failures, this mechanism ensures system stability while handling temporary connectivity issues.

Multi-Region High Availability

Enterprise inference platforms should include multi-region deployment architecture to ensure uninterrupted s ervice during regional outages. In the event of a regional disruption, traffic is automatically rerouted to functioning regions and additional GPUs are dynamically scaled to handle increased demand. By combining redundancy with autoscaling, it enables enterprises to maintain high-performance services and meet high-uptime SLAs, even during unexpected disruptions.

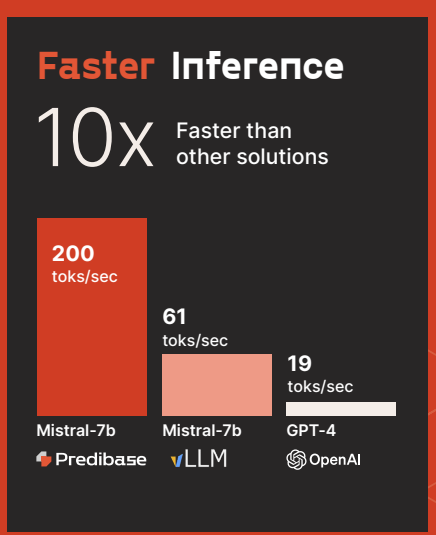
Unlocking the Full Potential of AI with Optimized Inference Stacks

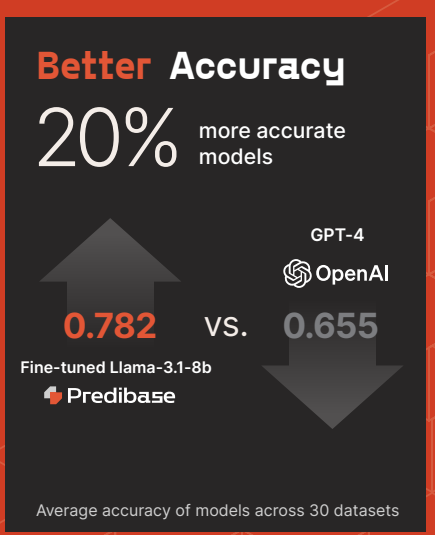
Choosing the Right Infrastructure Partner Matters

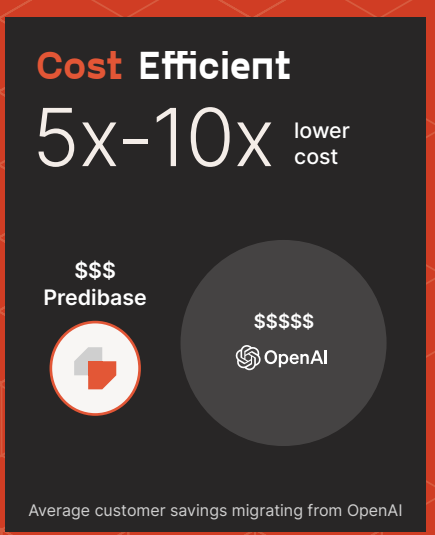
As enterprises continue to integrate AI into their core operations, the importance of a robust, efficient, and scalable inference infrastructure cannot be overstated. The Predibase Inference Engine exemplifies how a purpose-built solution can address the complex challenges of deploying fine-tuned SLMs at scale.

Predibase

Outperform GPT-4 with Small Models







TRUSTED BY LEADING BRANDS

checkr · nubank · CLEARWATER ANALYTICS · MarshMcLennan · Forethought · WWF

Optimized Performance for Mission-Critical AI

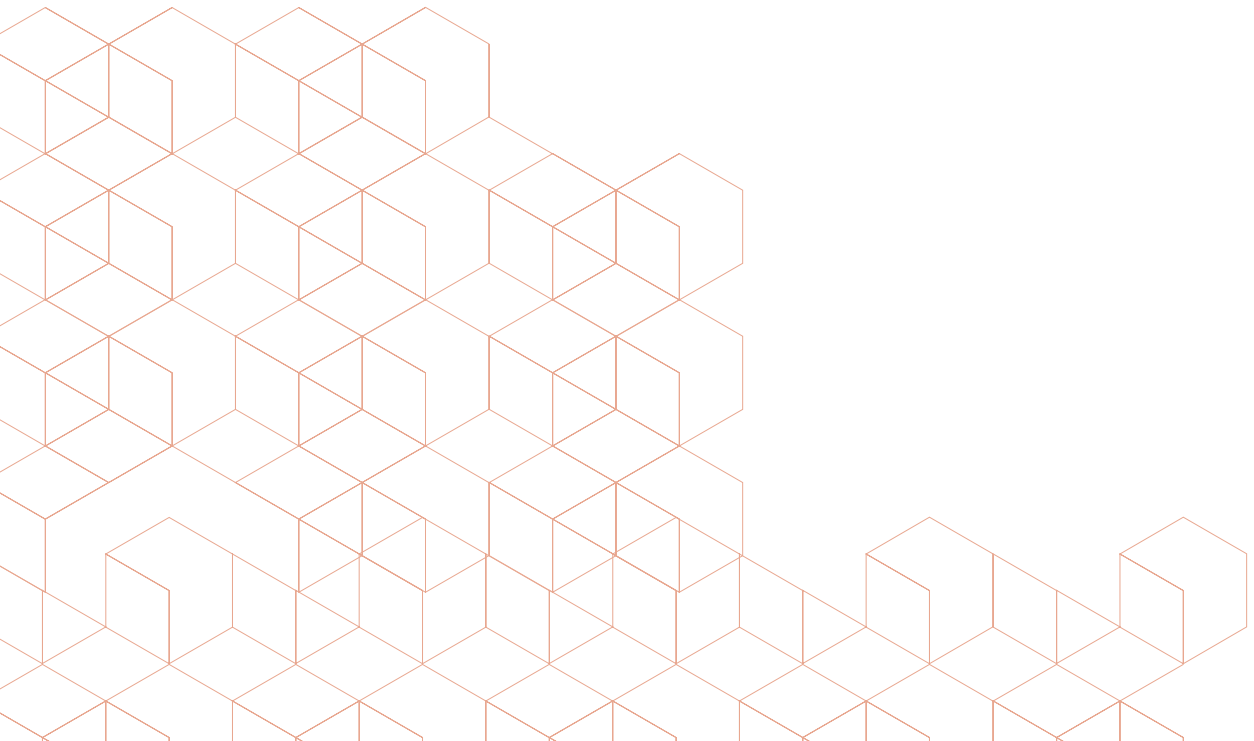
The Predibase Inference Engine is purpose-built to deliver exceptional performance for enterprises deploying fine-tuned small language models. With innovations like Turbo LoRA and LoRA Exchange, it achieves throughput improvements of 3-4x over traditional serving methods, significantly reducing infrastructure costs. The platform's ability to handle hundreds of requests per second with consistently low latency ensures seamless operation for mission-critical AI applications, empowering organizations to maximize the impact of their AI investments.

Cost-Effective Scaling Without Compromising Performance

Predibase leverages a highly sophisticated GPU autoscaling system to optimize infrastructure costs, dynamically adjusting resources to match demand. This approach ensures that enterprises can scale their AI operations efficiently, maintaining high performance at competitive rates while avoiding unnecessary expenses.

Purpose-Built for Fine-Tuned SLM Excellence

The Predibase Inference Engine is optimized for serving fine-tuned small language models (SLMs), delivering unmatched precision and performance. Tailored for fine-tuned LoRAs, the platform ensures seamless deployment of domain-specialized models. Across 30 datasets, Predibase-powered models outperform GPT-4 by 20% in accuracy, making it the superior choice for businesses seeking high-quality, tailored AI solutions.



Predibase

Accelerate Innovation with Simplified AI Infrastructure

In today's fast-paced AI landscape, enterprises need ready-to-deploy solutions that can keep pace with their innovation. The Predibase Inference Engine offers a comprehensive package that allows teams to focus on developing and refining their AI models rather than grappling with the complexities of infrastructure management. By providing a scalable, secure, and efficient platform for serving fine-tuned SLMs, Predibase enables organizations to accelerate their AI initiatives and drive tangible business value.

Don't let infrastructure limitations hold back your AI ambitions.

Take the first step towards optimizing your inference stack by signing up for your **free trial** or **scheduling a demo with an SLM expert** to see firsthand how our solutions can transform your AI operations.

Predibase