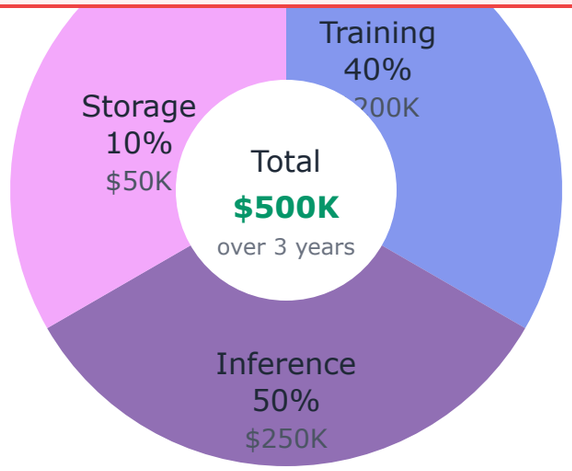
AI System Cost Breakdown and Optimization

Typical 3-Year Cost Breakdown

Critical Insight

Inference represents 50% of total costs.
Optimization here provides the largest ROI.



Cost Optimization Strategies

Training Cost Reduction

Spot Instances:	75% reduction	$1,224 → $306
Mixed Precision:	40%	20 → 12 GPU-hrs
Gradient Accum:	50%	$2.50 → $1.20/hr

Combined: 60-80% savings
$200K → $40K-$80K training cost

Inference Cost Reduction (Largest Impact)

Quantization:	67% reduction	$15K → $5K/mo
Pruning:	40%	30-40% faster
Distillation:	60%	97% performance
Caching:	10-50%	Based on hit rate
Batching:	5-20× throughput	+10-50ms latency

Combined: 70-90% savings
$250K → $25K-$75K inference cost

Cost Structure by Scale

Startup Scale

Volume:
100,000 predictions/day

Approach:
- Managed services (SageMaker, Vertex AI)
- Optimize for speed and flexibility

$0.005 per prediction
$15K/month

Optimization Transition Point

Volume: 1-10M predictions/day
Investment: $200K-$500K in optimization
Annual savings: $500K-$5M

ROI: 1-3 months
Optimization pays for itself quickly at scale

Enterprise Scale

Volume:
10,000,000 predictions/day (100× startup)

Approach:
- Custom infrastructure with optimization
- Quantization + pruning + caching
- Reserved instances for baseline capacity
- Spot instances for peak load
- Optimize for efficiency at scale

$0.001 per prediction
$300K/month

vs. Managed Services at Startup Pricing:
Saves $1.2M/month ($14.4M/year)