Savings Report

Sample report — AI-heavy startup

Generated August 5, 2026

Estimated monthly savings

$6,373.56

All figures are estimates based on on-demand us-east-1 pricing.

AI & ML — subtotal $4,874.87/mo

$1,561.47/mo — Previous-generation GPU endpoint ep-vision-legacy (1x ml.p3.2xlarge -> ml.g5.2xlarge): Redeploy the model on ml.g5.2xlarge (new endpoint or blue/green production-variant swap). Validate latency/throughput parity on the new instance type before cutting traffic over.

$1,027.84/mo — Idle endpoint ep-recsys-realtime (1x ml.g5.xlarge, 0 invocations): Delete the endpoint (model artifacts remain in S3 and it can be recreated in minutes). If it is needed intermittently, move to a serverless or asynchronous inference endpoint that scales to zero.

$967.04/mo — Sparse-traffic endpoint ep-nightly-batch (1x ml.g5.2xlarge, 79% zero-traffic days): Convert to an asynchronous inference endpoint (queues requests, scales instances to zero between bursts) or a scheduled batch transform if scoring is periodic. Keep real-time only if interactive latency is genuinely required.

$642.17/mo — Bedrock on-demand spend above Provisioned Throughput breakeven (111111111111): Evaluate Bedrock Provisioned Throughput (1-month or 6-month model units) for the steady baseline; keep on-demand for burst overflow. Verify per-model utilization in the Bedrock console before committing.

$378.00/mo — Frontier-model Bedrock spend to re-evaluate (222233334444): Evaluate routing a share of frontier-model calls to a cheaper tier: benchmark quality on your real prompts, then route by task complexity (classification/extraction rarely needs a frontier model). Start with the highest-volume prompt families; a quality-evaluation harness makes this safe to do incrementally.

$298.35/mo — On-demand SageMaker training that fits managed Spot (111111111111/us-east-1): Enable managed Spot training (EnableManagedSpotTraining) with S3 checkpointing for interruption tolerance. Typical savings run 60-70% vs on-demand; keep on-demand only for jobs that cannot checkpoint or must meet a hard deadline.

All findings

Est. $/mo	Finding	Category	Effort	Recommendation
$1,561.47	Previous-generation GPU endpoint ep-vision-legacy (1x ml.p3.2xlarge -> ml.g5.2xlarge)	ai_ml	Migration	Redeploy the model on ml.g5.2xlarge (new endpoint or blue/green production-variant swap). Validate latency/throughput parity on the new instance type before cutting traffic over.
$1,027.84	Idle endpoint ep-recsys-realtime (1x ml.g5.xlarge, 0 invocations)	ai_ml	Config change	Delete the endpoint (model artifacts remain in S3 and it can be recreated in minutes). If it is needed intermittently, move to a serverless or asynchronous inference endpoint that scales to zero.

Est. $/mo	Finding	Category	Effort	Recommendation
$967.04	Sparse-traffic endpoint ep-nightly-batch (1x ml.g5.2xlarge, 79% zero-traffic days)	ai_ml	Migration	Convert to an asynchronous inference endpoint (queues requests, scales instances to zero between bursts) or a scheduled batch transform if scoring is periodic. Keep real-time only if interactive latency is genuinely required.
$642.17	Bedrock on-demand spend above Provisioned Throughput breakeven (111111111111)	ai_ml	Migration	Evaluate Bedrock Provisioned Throughput (1-month or 6-month model units) for the steady baseline; keep on-demand for burst overflow. Verify per-model utilization in the Bedrock console before committing.
$534.30	Uncovered EC2 on-demand spend (111111111111)	commitments	Config change	Purchase a 1-year no-upfront Compute Savings Plan sized to the steady baseline (start at ~70% of the uncovered spend after actioning the idle/rightsizing findings above, then ratchet up). Compute SPs follow the workload across instance families, sizes, and regions, so they stay safe through migrations.
$378.00	Frontier-model Bedrock spend to re-evaluate (222233334444)	ai_ml	Rearchitecture	Evaluate routing a share of frontier-model calls to a cheaper tier: benchmark quality on your real prompts, then route by task complexity (classification/extraction rarely needs a frontier model). Start with the highest-volume prompt families; a quality-evaluation harness makes this safe to do incrementally.
$298.35	On-demand SageMaker training that fits managed Spot (111111111111/us-east-1)	ai_ml	Migration	Enable managed Spot training (EnableManagedSpotTraining) with S3 checkpointing for interruption tolerance. Typical savings run 60-70% vs on-demand; keep on-demand only for jobs that cannot checkpoint or must meet a hard deadline.
$247.38	NAT gateway data-processing spend (111111111111)	network	Migration	Add gateway VPC endpoints for S3 and DynamoDB (free) and interface endpoints for other high-volume AWS services so that traffic stops paying the NAT per-GB toll. Check VPC Flow Logs for the top talkers before choosing endpoints.

Est. $/mo	Finding	Category	Effort	Recommendation
$155.00	Inter-AZ data transfer spend (111111111111)	network	Rearchitecture	Identify the chattiest cross-AZ flows (VPC Flow Logs / Cost Explorer usage-type drilldown), then co-locate tight client-server pairs in one AZ and enable topology/AZ-aware routing where the stack supports it. Keep multi-AZ for real failover paths -- this targets accidental cross-AZ chatter.
$146.00	Idle EC2 instance i-idle-oldgen (m4.xlarge, avg CPU 0.5%)	compute	Config change	Stop the instance (or terminate it if the workload is retired). If it is needed on a schedule, use an instance scheduler so it only bills during working hours.
$140.16	Idle EC2 instance i-idle-batch01 (m5.xlarge, avg CPU 1.5%)	compute	Config change	Stop the instance (or terminate it if the workload is retired). If it is needed on a schedule, use an instance scheduler so it only bills during working hours.
$124.83	Idle RDS instance rds-analytics-idle (db.m5.large, 0 connections)	database	Config change	Stop the instance (RDS restarts stopped instances after 7 days -- schedule the stop weekly, or snapshot and delete if the database is retired). Savings shown are instance hours only; storage keeps billing while stopped.
$42.34	Previous-generation EC2 instance i-oldgen-web01 (c4.2xlarge -> c5.2xlarge)	compute	Migration	Migrate to c5.2xlarge (same architecture, drop-in for most workloads). If the stack can run on ARM, the equivalent Graviton family is cheaper still -- worth testing after the x86 move.
$27.60	S3 bucket logs-archive-raw has no lifecycle policy (4,000 GB Standard)	storage	Config change	Add a lifecycle rule transitioning objects to S3 Intelligent-Tiering (safe default -- no retrieval fees, per-object monitoring only) or to Standard-IA after 30 days for known-cold data. Expire incomplete multipart uploads while you're in there.
$25.00	Stale snapshot snap-old500 (500 GB, 400 days old)	storage	Config change	Delete after confirming no AMI or restore dependency references this snapshot. Consider a snapshot lifecycle policy (Amazon Data Lifecycle Manager) so old snapshots are pruned automatically going forward.

Est. $/mo	Finding	Category	Effort	Recommendation
$16.43	Idle Application Load Balancer alb-idle-legacy (0 requests)	network	Config change	Delete the load balancer after confirming no DNS records point at it. If it fronts an autoscaling group kept for failover, document that as an intentional exception rather than leaving it running unexplained.
$12.00	Log group /aws/lambda/etl-daily never expires (800 GB stored)	observability	Config change	Set a retention policy (30-90 days for application logs; export to S3 with a lifecycle rule first if compliance needs longer). Existing bytes past the new retention age out automatically.
$10.00	gp2 volume vol-0a1 (500 GB) -> gp3	storage	Config change	Migrate to gp3 via ModifyVolume (online, no downtime). gp3 baseline is 3000 IOPS / 125 MBps; provision extra only if the workload exceeded gp2 burst performance.
$10.00	Unattached gp2 volume vol-0c3 (100 GB)	storage	Config change	Snapshot for safety, then delete. If it must be retained, a snapshot costs a fraction of a provisioned volume.
$4.00	gp2 volume vol-0b2 (200 GB) -> gp3	storage	Config change	Migrate to gp3 via ModifyVolume (online, no downtime). gp3 baseline is 3000 IOPS / 125 MBps; provision extra only if the workload exceeded gp2 burst performance.
$3.65	Unassociated Elastic IP 203.0.113.10 (eipalloc-idle1)	network	Config change	Release the address if it is no longer needed, or associate it with an instance/ENI if it is being held deliberately. Check for DNS references to this IP first -- a released Elastic IP cannot be recovered.

Full findings detail (including resource identifiers) is available in findings.json alongside this report.