Patronus Insights

The road to sub-cent AI Security inference

Ausgabe 3

In dieser Ausgabe

00 Inhalt

AI Security wird zum Inference-Problem

Mit Agenten, Tools und langen Kontexten wächst nicht nur die Zahl der Security-Entscheidungen. Auch der Umfang der zu prüfenden Daten steigt. Die entscheidende Frage lautet deshalb nicht nur, wie schnell ein Security-Modell läuft, sondern wie oft der teuerste Modellpfad überhaupt notwendig ist.

Patronus ARK wurde ursprünglich unter einer harten Vorgabe entwickelt: Echtzeit-AI-Security sollte auf Commodity-CPUs funktionieren. Diese Einschränkung hat die Architektur stärker geprägt als jede einzelne Quantisierung.

Statt jeden 256-Token-Chunk durch den teuersten Encoder zu schicken, arbeitet ARK als Kaskade. Günstige strukturelle und semantische Stufen treffen frühe Entscheidungen. Nur dort, wo Unsicherheit, Risiko oder Widerspruch verbleiben, wird ein Chunk an L3 weitergereicht.

Der entscheidende Punkt: Das ist kein beliebiger Pre-Filter. Die Kaskade wurde gegen den vollständigen L3-Pfad trainiert und evaluiert. Über mehr als 150.000 Trainingsbeispiele, rund 50.000 Validation-Beispiele, unabhängige Holdouts und OOD-Benchmarks erreicht die selektive Architektur ungefähr dasselbe F1- und False-Positive-Verhalten wie 100 % L3-Inference.

Am gewählten Operating Point werden nur rund 25 % der Chunks an L3 promoted. Damit reduziert ARK die teure Inference nicht, indem weniger geprüft wird, sondern indem Deep Inference gezielt dort eingesetzt wird, wo sie die Entscheidung verändert.

Was zunächst als CPU-Optimierung entstand, wird auf GPU-Hardware noch interessanter: Die CPU übernimmt High-Volume-Routing, die GPU nur den schwierigen Teil des Entscheidungsraums.

Promoted zu L3

~25%

bei ~gleicher F1/FPR

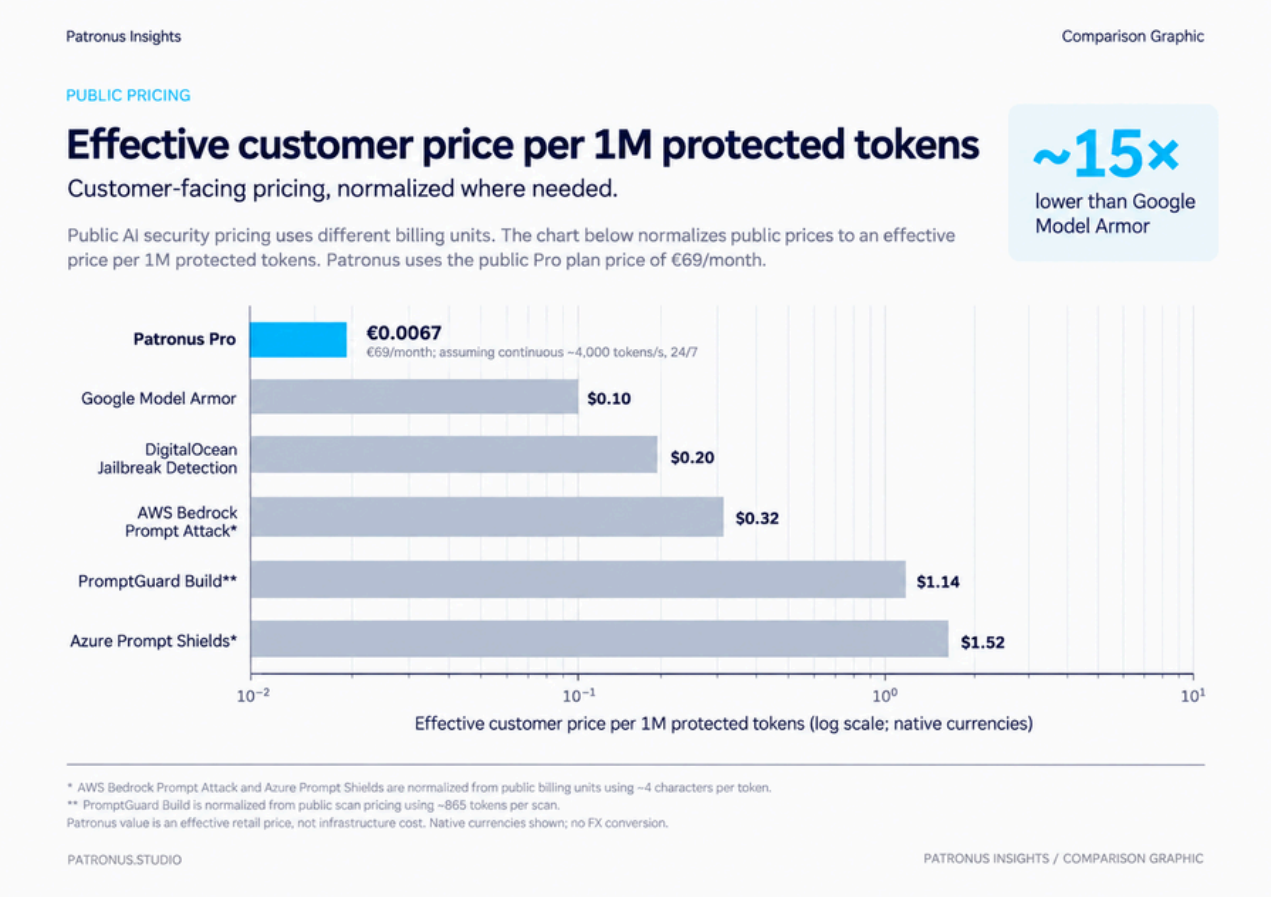
Train + Validation

>200k

plus Holdouts & OOD

Die aktuelle ARK-Kostenkurve in Zahlen

Die folgenden Werte verbinden gemessenen Durchsatz mit dem öffentlich geplanten Patronus-Pro-Preis.



Measured stack: CPU L2 peak ~963k input tok/s · ~25 % promotion · L3 demand ~241k tok/s · Full NVIDIA L4 ~387k L3 tok/s · 2.40 ms single chunk · 10.59 ms batch 16 · ~1,511 chunks/s.

Public pricing: Patronus Pro ~€0.0067 / 1M protected tokens at ~4,000 tok/s continuous usage · Google Model Armor $0.10 / 1M · DigitalOcean $0.20 / 1M. AWS and scan-based services are normalized in the methodology.

~25 % Deep Inference erhalten die Qualität von 100 % Deep Inference

DAS KERNERGEBNIS

GPU-Auslastung zu senken ist trivial, wenn die Erkennungsqualität schlechter werden darf. Genau darum geht es hier nicht.

VALIDIERTER OPERATING POINT

ARK wurde gegen den vollständigen L3-Pfad trainiert und evaluiert. Über mehr als 150.000 Trainingsbeispiele, rund 50.000 Validation-Beispiele, unabhängige Holdouts und OOD-Benchmarks erreicht die Kaskade ungefähr dasselbe F1- und False-Positive-Verhalten wie 100 % L3-Inference.

ÖKONOMISCHE FOLGE

Am gewählten Operating Point werden nur rund 25 % der Chunks promoted. Anders gesagt: ~25 % Deep Inference reichen in unserem evaluierten Setup aus, um ungefähr die Qualität von 100 % Deep Inference zu erhalten. Alles danach ist eine ökonomische Konsequenz.

Der günstige Pfad routet fast eine Million Input-Tokens pro Sekunde

MESSERGEBNIS

Auf dem aktuellen 12-Cube-/36-Worker-CPU-Setup erreicht der optimierte L2-Pfad bei 288 parallelen Clients rund 4.839 Requests/s und ~963k Input-Tokens/s. Die Client-Latenz liegt bei 51,16 ms p50 und 92,63 ms p95.

L2 peak

~963k tok/s

SÄTTIGUNGSPUNKT

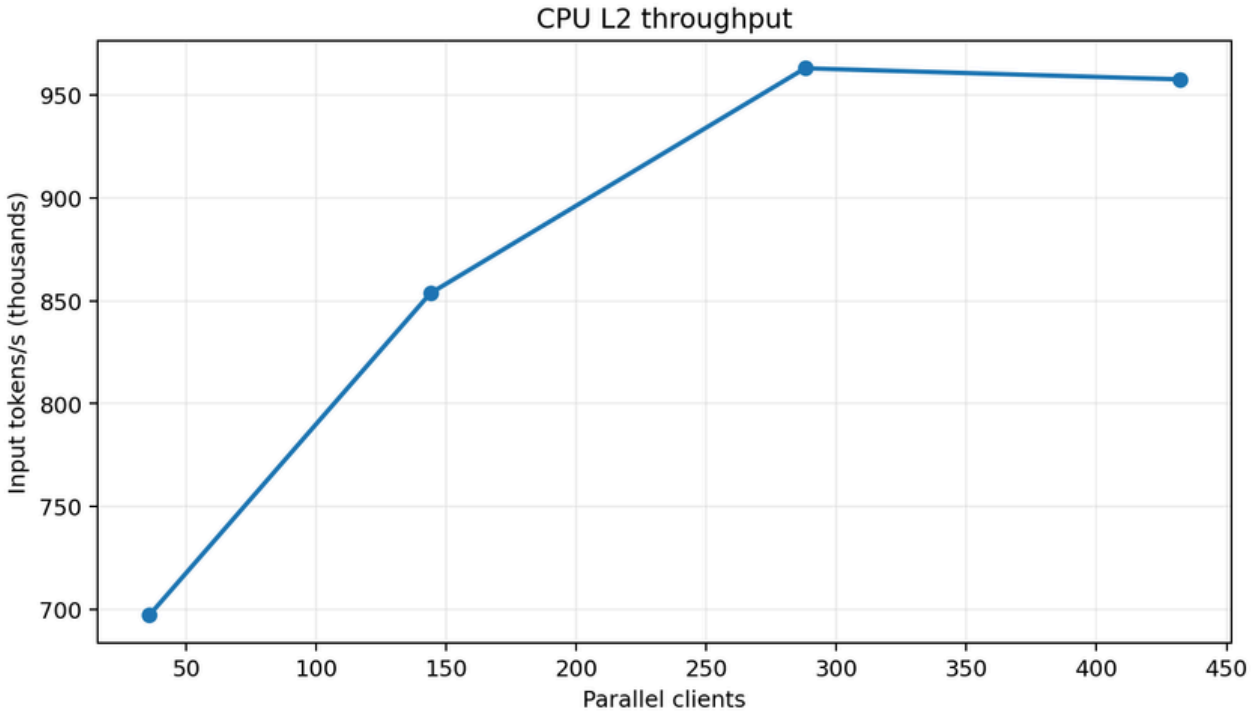
Bei 432 Clients bleibt der Durchsatz mit rund 958k Tokens/s nahezu gleich, während p95 auf etwa 131 ms steigt. Der Cluster liegt damit bereits nahe an seinem aktuellen L2-Sättigungspunkt.

Client p95

92.63 ms

WAS DAS FÜR L3 BEDEUTET

Bei ~25 % Promotion werden aus ~963k Input-Tokens/s nur rund ~241k L3-Tokens/s. Die CPU-Flotte kann also fast eine Million Tokens pro Sekunde vorsortieren, ohne dass der GPU-Pfad dasselbe Volumen verarbeiten muss.



Eine volle L4 kann den aktuellen promoted L3-Workload aufnehmen

WOLF DEFENDER SMALL

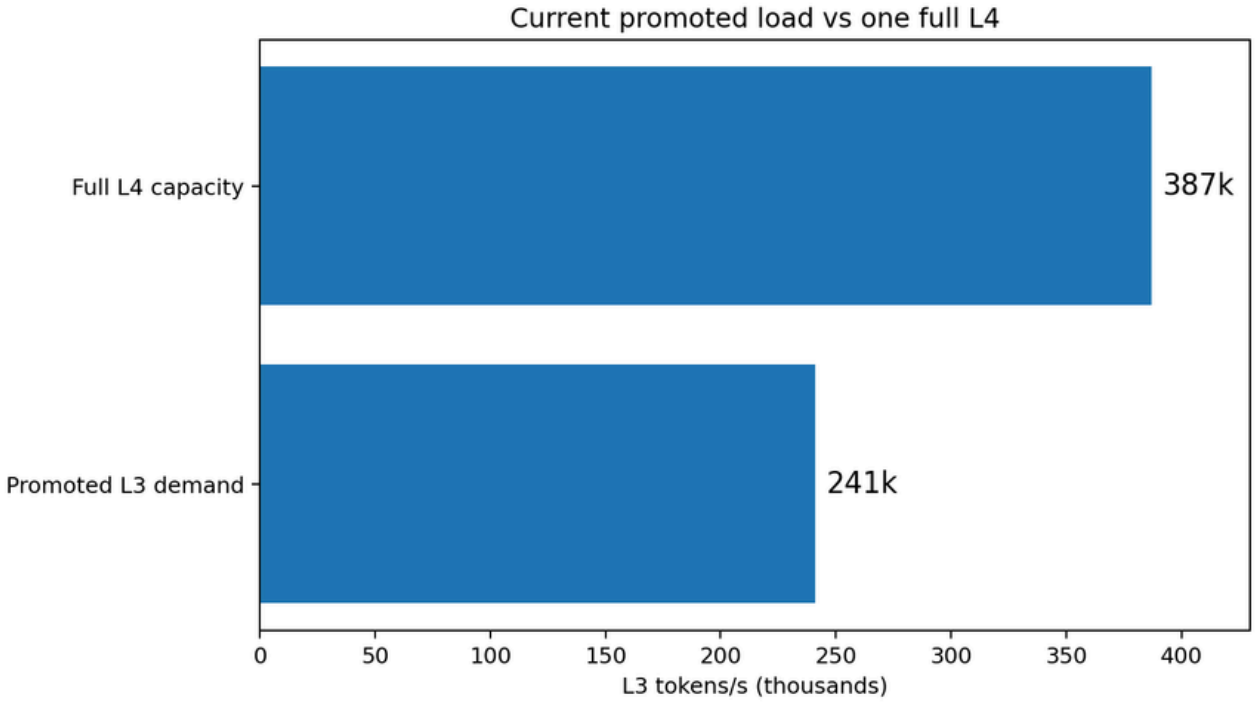
Mit Wolf Defender Small, TensorRT und 256-Token-Chunks messen wir auf einer vollen NVIDIA L4 2,40 ms für einen einzelnen Chunk und 10,59 ms für Batch 16. Batch 16 erreicht rund 1.511 Chunks/s, also ungefähr 387k L3-Tokens/s.

KAPAZITÄTSVERGLEICH

Der aktuelle CPU-Peak erzeugt rund 241k promoted L3-Tokens/s. Eine einzelne volle L4 kann diesen Workload mit deutlichem Headroom verarbeiten.

WARUM DAS RELEVANT IST

Die GPU skaliert mit dem promoted Anteil, nicht mit dem gesamten geschützten Traffic. Das ist die zentrale Hardware-Folge von Selective Inference.



CPU-first muss nicht CPU-only bedeuten

DER ARCHITEKTUREFFEKT

Die CPU-Constraints haben ARK dazu gezwungen, teure Inference selektiv einzusetzen. Sobald GPU-Beschleunigung hinzukommt, wird dieselbe Architektur dadurch effizienter: CPUs übernehmen High-Volume-Routing, GPUs bleiben für dichte Deep Inference reserviert.

TENSORRT ALS ZWEITER HEBEL

Auf einer 1/4 L4 sank Batch 8 nach der TensorRT-Optimierung von rund 45,8 ms auf 18,0 ms. Selective Inference reduziert, wie oft L3 läuft; TensorRT erhöht, wie dicht die verbleibende L3-Arbeit ausgeführt werden kann.

DAS ERGEBNIS

Die schnellste teure Inference ist die, die nie ausgeführt werden musste. CPU-Constraints haben die selektive Architektur hervorgebracht; GPU-Beschleunigung macht die verbleibende teure Berechnung sehr dicht.

Der Weg zu ~€0,0067 pro Million geschützter Tokens

PATRONUS PRO

Patronus Pro ist mit €69 pro Monat und ungefähr 4.000 Tokens/s sustained throughput geplant. Bei kontinuierlichen ~4.000 Tokens/s, 24/7, sind das 10,368 Milliarden geschützte Tokens pro 30-Tage-Monat oder rund €0,0067 pro 1M geschützte Tokens.

ÖFFENTLICHE REFERENZEN

Google Model Armor veröffentlicht $0,10 pro 1M Tokens. DigitalOcean Jailbreak Detection veröffentlicht $0,20 pro 1M Tokens. Andere Anbieter rechnen nach Text Units oder Scans ab und benötigen deshalb explizite Normalisierungsannahmen.

DER VERGLEICH

Bei kontinuierlichen ~4.000 Tokens/s liegt Patronus Pro ungefähr 15× unter Googles veröffentlichtem Preis von $0,10 pro 1M Tokens, vor Währungsumrechnung. Verglichen werden Kundenpreise, nicht interne Infrastrukturkosten der Wettbewerber.

08 SCALING

Skaliere die Stufe, die tatsächlich sättigt

Ein heterogener Stack verändert, wie Kapazität hinzugefügt wird. Free- und Light-Traffic können auf dem CPU-first Pfad bleiben: L1 und L2 absorbieren den Großteil der Requests, nur promoted Chunks nutzen den gemeinsamen GPU-Pool. Pro-Traffic kann priorisiert oder direkt auf GPU-Kapazität geroutet werden, wenn niedrigere Latenz oder höherer garantierter Durchsatz wichtiger sind.

Entscheidend ist die unabhängige Skalierung. Wird der CPU-Ingress zum Bottleneck, kommen CPU-Worker hinzu. Übersteigt der promoted L3-Bedarf die GPU-Kapazität, kommen GPUs hinzu. Verbessert sich die Promotion-Entscheidung, steigt die unterstützte Gesamtlast ohne Änderung der GPU-Flotte.

Die Promotion-Rate wird damit zu einer operativen Effizienzkennzahl. Eine Reduktion von 25 % auf 20 % bei gleicher Erkennungsqualität bedeutet rund 20 % weniger Deep-Inference-Bedarf für denselben geschützten Traffic.

VIER KONSEQUENZEN

1. **CPU als Basiskapazität. High-Volume-Routing bleibt auf Commodity-Hardware.**
2. **GPU als Shared Accelerator. Nur der schwierige semantische Anteil erreicht Wolf Defender Small.**
3. **Unabhängiges Scaling. CPU- und GPU-Pools folgen unterschiedlichen Bottlenecks.**
4. **Pricing folgt dem Pfad. Günstige Tiers können kaskadiert bleiben; Premium-Traffic kann priorisierte GPU-Kapazität erhalten.**

09 METHODIK & CAVEATS

Was die Zahlen zeigen – und was nicht

Die Benchmark-Zahlen in dieser Ausgabe sind gemessene Patronus-Werte. Der Preisvergleich verwendet öffentliche Kundenpreise und normalisiert nur dort, wo ein Anbieter nicht direkt pro Token abrechnet.

BENCHMARK-BASIS

CPU: 12 Cube L / 36 Worker, L2-only, 256-Token-Chunks. GPU: Wolf Defender Small mit TensorRT auf NVIDIA L4. Promotion: ~25 % am validierten Operating Point.

ERKENNUNGSQUALITÄT

Die ~25-%-Promotion ist kein frei gewähltes Kostenlimit. Die Kaskade wurde auf >150k Trainingsbeispielen, ~50k Validation-Beispielen sowie unabhängigen Holdouts und OOD-Benchmarks gegen die 100-%-L3-Baseline evaluiert. Am gewählten Operating Point bleiben F1 und False-Positive-Verhalten ungefähr auf Full-L3-Niveau.

PREIS-NORMALISIERUNG

Patronus Pro: €69/Monat bei angenommenen kontinuierlichen ~4.000 Tokens/s, 24/7. Google und DigitalOcean veröffentlichen direkte Tokenpreise. AWS und scan-basierte Anbieter benötigen zusätzliche Annahmen wie Zeichen pro Token oder Tokens pro Scan.

SCHLUSSFOLGERUNG

Wenn ungefähr ein Viertel des Traffics Deep Inference benötigt, ohne die gemessene Erkennungsqualität wesentlich zu verändern, sinkt die marginale GPU-Arbeit pro geschütztem Token bereits vor weiteren Hardware-Optimierungen.

10 QUELLEN

Patronus Protect — Wolf Defender Prompt Injection Modell
https://huggingface.co/patronus-studio/wolf-defender-prompt-injection

Prompt Injection Cost — öffentlicher AI-Security-Preisvergleich
https://promptinjectioncost.com/

Google Cloud — Model Armor Pricing
https://cloud.google.com/security-command-center/pricing

DigitalOcean — Inference / Guardrail Pricing
https://docs.digitalocean.com/products/inference/details/pricing/

AWS — Bedrock Guardrails Pricing
https://aws.amazon.com/bedrock/pricing/

PromptGuard — öffentliches Pricing
https://www.promptguard.co/pricing