OWASP API 성능 테스트 결과 보고서

v1.0 | 2026-06-23 | 이니넥스트

1. 테스트 개요

1.1 테스트 목적

1.1.1 성능 테스트 수행 배경

가드레일 추론은 GPU(Triton + vLLM)에서 처리되어 GPU 사양이 처리량·지연을 좌우한다. 본 테스트는 동일 구성·부하에서 L4(g6)와 L40S(g6e)의 처리량·지연을 측정·비교해, 워크로드별 적합 GPU를 판단할 근거를 확보하고자 수행한다.

1.1.2 검증 대상 API

가드레일 API 3종(pii·jailbreak·toxic)을 측정 대상으로 한다. 단독 호출과, 실사용을 반영한 병렬 조합(`pii+jailbreak`, `pii+jailbreak+toxic`)을 모두 포함한다.

1.1.3 테스트 목표

1. L4·L40S 각각의 가드레일별·조합별 응답 시간(p95)과 최대 처리량(RPS) 측정
2. 동일 구성·부하에서 두 GPU의 응답 시간·처리량 차이 비교
3. 1000토큰 기준 가드레일별·조합별 권장 운영 기준 도출

1.2 테스트 환경

1.2.1 서버 및 인프라 구성

L4(g6)·L40S(g6e) 두 환경은 GPU 종류를 제외하고 모두 동일하다. 서버 1대에 다음 구성으로 배포한다.

- Gateway

- OWASP : 가드레일 처리 AI 모델 (pii / jailbreak / toxic)
- vLLM Server ×1 / Triton Server ×3

> **L4 (g6)**
>
> - 인스턴스: g6.12xlarge / NVIDIA L4 23GB × 4 / 48 vCPU / 192GB
> - GPU 배치: GPU0에 vLLM(jailbreak), GPU1·2·3에 Triton(toxic, pii)

> **L40S (g6e)**
>
> - 인스턴스: g6e.12xlarge / NVIDIA L40S 46GB × 4 / 48 vCPU / 384GB
> - GPU 배치: GPU0에 vLLM(jailbreak), GPU1·2·3에 Triton(toxic, pii)

2. 테스트 시나리오

2.1 대상 API

2.1.1 단독 호출

API	가드레일	엔진
`POST /api/v1/detect/pii`	개인정보 탐지	Triton
`POST /api/v1/detect/jailbreak`	탈옥 시도 탐지	vLLM
`POST /api/v1/detect/toxic`	유해성 탐지	Triton

2.1.2 조합 호출

실사용을 반영해, 한 요청이 여러 가드레일을 동시에 통과하는 경우를 측정한다.

- `pii + jailbreak` (2종 병렬)
- `pii + jailbreak + toxic` (3종 병렬)

2.2 부하 조건

동시 사용자 수(VU)를 단계적으로 올리며 각 단계의 처리량과 지연을 측정한다.

항목	조건
동시 사용자 수 (VU)	5 → 10 → 20 → 40 → 70 → 100
부하 증가 방식	각 단계를 순차적으로 올리는 스윕
단계별 측정 시간	단계당 30초
요청 조건	요청당 토큰 1000 고정
총 요청 수	고정하지 않음 (VU × 처리량에 따라 결정)

3. 테스트 결과 (L4 vs L40S)

동일한 구성과 부하 조건에서 L4(g6)와 L40S(g6e)를 각각 측정·비교한다.

3.1 응답시간

동일 구성 및 동일 부하 조건에서 L4(g6)와 L40S(g6e)의 P95 응답시간을 측정한다.

> **P95 응답시간이란?**
>
> 전체 요청을 응답시간이 빠른 순으로 정렬했을 때, 하위 95% 지점에 해당하는 값이다. 즉 요청의 95%가 이 시간 안에 처리되었고, 가장 느린 5%만 이보다 오래 걸렸음을 의미한다.

3.1.1 단독 호출 - P95 (ms)

L4(g6)

가드레일	VU5	VU10	VU20	VU40	VU70	VU100
PII	355	472	1,130	2,859	3,279	4,842
Jailbreak	102	160	305	612	907	1,190
Toxic	311	608	1,198	1,994	3,680	4,432

L40S(g6e)

가드레일	VU5	VU10	VU20	VU40	VU70	VU100
PII	399	484	629	1,921	2,077	2,805
Jailbreak	103	130	235	483	714	1,023
Toxic	113	165	290	603	942	1,326

3.1.2 조합 호출 - P95 (ms)

L4(g6)

조합	VU5	VU10	VU20	VU40	VU70	VU100
PII + Jailbreak	458	750	1,225	2,901	4,461	15,764
PII + Jailbreak + Toxic	482	1,609	2,479	4,211	15,930	15,763

L40S(g6e)

조합	VU5	VU10	VU20	VU40	VU70	VU100
PII + Jailbreak	301	452	716	1,336	1,692	2,172
PII + Jailbreak + Toxic	349	576	1,131	1,479	2,311	3,129

3.2 처리량

동일 구성 및 동일 부하 조건에서 L4(g6)와 L40S(g6e)의 RPS를 측정한다.

> **RPS(Requests Per Second)란?**
>
> 서버가 1초당 처리하는 요청 수를 의미한다. 시스템의 처리량(throughput)을 나타내는 지표로, 값이 높을수록 같은 시간에 더 많은 요청을 소화할 수 있음을 뜻한다.

3.2.1 단독 호출 - RPS

L4(g6)

가드레일	VU5	VU10	VU20	VU40	VU70	VU100
PII	19.0	30.5	31.8	34.6	36.4	35.2
Jailbreak	50.4	84.4	95.3	93.2	106.8	107.6
Toxic	24.6	25.5	28.4	26.7	26.6	26.9

L40S(g6e)

가드레일	VU5	VU10	VU20	VU40	VU70	VU100
PII	17.2	31.5	49.0	55.4	61.9	63.8
Jailbreak	61.5	101.5	116.2	118.6	127.7	122.6
Toxic	57.6	85.7	103.0	100.4	104.9	108.5

3.2.2 조합 호출 - RPS

L4(g6)

조합	VU5	VU10	VU20	VU40	VU70	VU100
PII + Jailbreak	17.2	26.6	28.9	31.0	30.9	24.2*
PII + Jailbreak + Toxic	17.6	10.6	11.2	12.8	12.9	20.1*

L40S(g6e)

조합	VU5	VU10	VU20	VU40	VU70	VU100
PII + Jailbreak	24.3	36.9	47.6	50.1	55.3	56.5
PII + Jailbreak + Toxic	21.2	28.9	31.9	36.8	37.7	37.4

3.3 오류율

두 환경 모두 전 구간(VU 5~100)에서 HTTP 오류는 발생하지 않았다. 부하 증가에 따라 응답 시간은 증가하였으나 요청 실패는 발생하지 않았다.

항목	L4	L40S
오류 건수	0	0
오류율	0%	0%

4. 분석 결과

테스트 전 구간(VU 5~100)에서 두 환경 모두 HTTP 오류 없이(오류율 0%) 동작하여 서비스 안정성을 확인하였다.

4.1 처리량(RPS) 비교

가드레일	g6 (L4) 최대	g6e (L40S) 최대	배수
Toxic	~28 RPS	~108 RPS	3.9×
PII	~36 RPS	~64 RPS	1.8×
Jailbreak	~108 RPS	~128 RPS	1.2×
PII + Jailbreak	~31 sets/s	~56 sets/s	1.8×
PII + Jailbreak + Toxic	~13 sets/s	~38 sets/s	2.9×

처리량 향상 폭은 워크로드에 따라 달랐다. Triton 기반 가드레일(Toxic·PII)에서 가장 컸고(Toxic 약 3.9배), Jailbreak는 두 환경 모두 여유가 있어 차이가 작았다(약 1.2배). 응답시간 안정성에서도 차이가 뚜렷했다. 다중 Guardrail 고부하 시 L4는 P95가 약 16초까지 치솟았지만, L40S는 약 3초를 유지했다.

5. 결론

두 환경 모두 전 구간에서 오류 없이 안정적으로 동작했다. L40S(g6e)는 L4(g6)보다 처리량과 응답시간이 우수하며, 특히 Triton 기반 가드레일와 다중 Guardrail 고부하에서 격차가 크다.

다만 두 GPU의 처리 한계가 다를 뿐, 어느 쪽이 일방적으로 우월한 것은 아니다. 따라서 GPU 선택은 서비스에 필요한 RPS를 기준으로 한다. 구체적인 선택은 아래 표에 목표 RPS를 대입해 결정한다.

5.1 운영 권장 기준 (1,000 Token, P95 SLA)

아래 구간은 응답시간이 안정적으로 유지되는 권장 동시성 범위이며, 이를 초과하면 P95가 급격히 증가할 수 있다.

L4 (g6) 기준

시나리오	권장 동시성	처리량	P95
Jailbreak 단독	VU≤20	~100 RPS	~0.3초
PII 단독	VU≤20	~32 RPS	~1.1초
Toxic 단독	VU≤20	~28 RPS	~1.2초
3종 동시 (풀체크)	VU≤20	~12 sets/s	~2.4초

L40S (g6e) 기준

시나리오	권장 동시성	처리량	P95
Jailbreak 단독	VU≤40	~119 RPS	~0.5초
PII 단독	VU≤20	~49 RPS	~0.6초
Toxic 단독	VU≤20	~103 RPS	~0.3초
3종 동시 (풀체크)	VU≤40	~37 sets/s	~1.5초

5.2 최종 결론 — L4 × 4 권장

본 서비스의 목표 부하는 동시접속자 약 100명이다. 위 RPS는 사용자 think time이 없는 연속 요청 기준의 보수적 수치로, 실제 환경에서는 사용자 요청 간격을 고려하면 동일 RPS로 더 많은 사용자를 수용할 수 있다.

사용자 1명이 평균 약 3초에 1회 요청한다고 가정하면 RPS 30은 동시접속자 약 100명에 해당한다. 위 표에서 L4(g6)는 주요 가드레일 시나리오에서 이에 상응하는 처리량을 확보하므로, 일반적인 사용자 환경에서 동시접속자 100명을 안정적으로 커버할 수 있다. 따라서 현재 트래픽 규모에서는 L4 × 4 도입을 권장한다.