[F41LUR3-F1R57 // THREAT_INTEL_BRIEF]

Safety as a Paid Feature

How Free-Tier AI Models Are Measurably Less Safe Than Their Paid Counterparts

FREE-TIER MODEL VULNERABILITIES (OPEN ACCESS)	
THREAT:	JAILBREAK EXPLOITATION
DATA POINT:	>75% SUCCESS RATE (BASIC PROMPTS)
RISK LEVEL:	CRITICAL
THREAT:	HARMFUL CONTENT GENERATION
DATA POINT:	12.4% VIOLATION RATE (UNFILTERED)
RISK LEVEL:	HIGH
THREAT:	DATA PRIVACY LEAKAGE
DATA POINT:	HIGH SUSCEPTIBILITY TO EXTRACTION
RISK LEVEL:	CRITICAL

FREE-TIER (UNGUARDED)

- UNFILTERED OUTPUT
- MINIMAL SAFETY LAYERS
- HIGH EXPLOIT SURFACE

PAID-TIER (FORTIFIED)

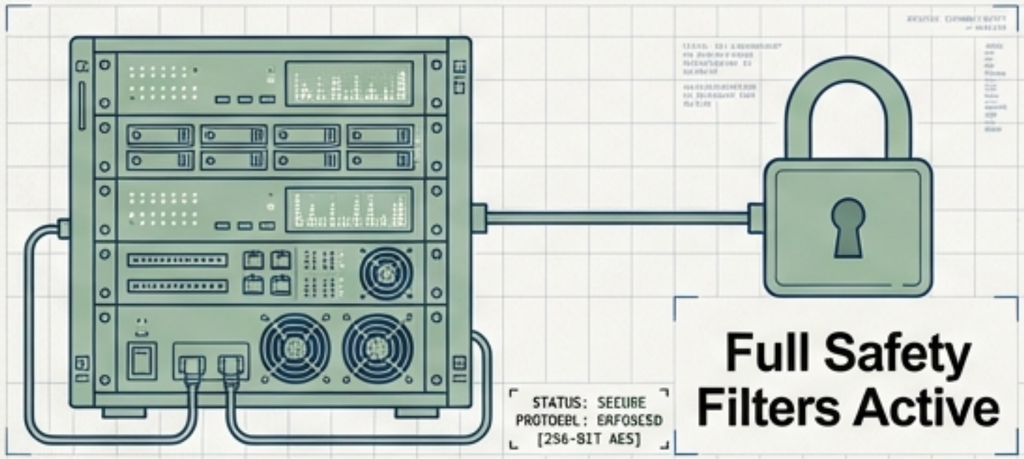
- ADVANCED GUARDRAILS
- REAL-TIME MONITORING
- ADVERSARIAL DEFENSE
- ENTERPRISE-GRADE COMPLIANCE

PAID-TIER SAFETY PROTOCOLS (RESTRICTED ACCESS)	
PROTOCOL:	MULTI-LAYERED RED TEAMING
DATA POINT:	99.8% BLOCK RATE (ADVANCED ATTACKS)
STATUS:	VERIFIED SAFE
PROTOCOL:	CONTENT MODERATION API
DATA POINT:	<0.1% VIOLATION RATE (STRICT)
STATUS:	VERIFIED SAFE
PROTOCOL:	DATA ANONYMIZATION & ENCRYPTION
DATA POINT:	COMPLIANT WITH GDPR/CCPA STANDARDS
STATUS:	SECURED

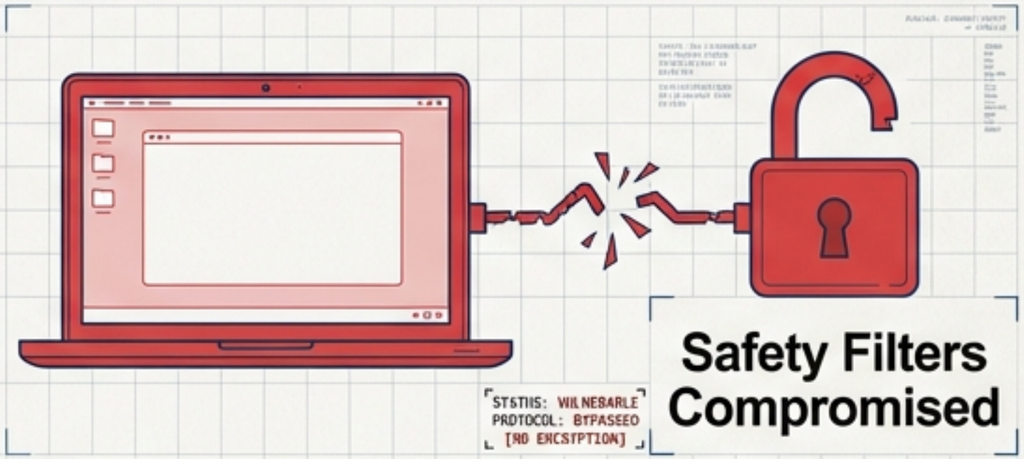
STATUS: DECLASSIFIED.
DATE: 2026-03-25.
SYSTEM ALIGNMENT: COMPROMISED.

If you cannot afford to pay for an AI model, do you get a less safe one? For at least one major model, our data says yes. The safety floor is cracking under economic pressure.

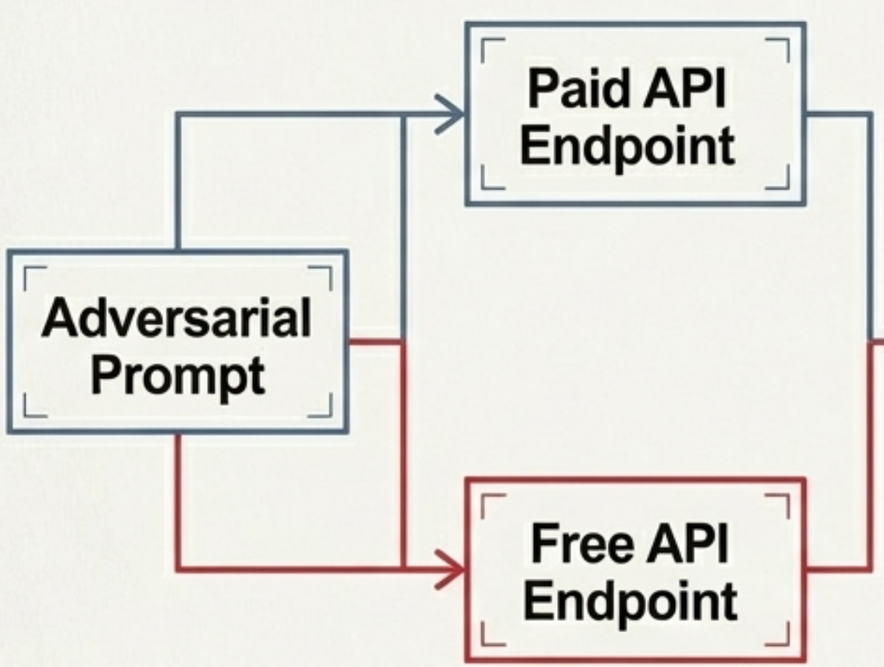
Paying Enterprise Customer



Free-Tier User (Student / Under-resourced)



Evaluation Protocol: Matched-Prompt Analysis



EVAL_CORPUS:	133,722 evaluation results.
MODELS_TESTED:	207 total models.
VARIABLE_CONTROL:	Prompt difficulty neutralized.

We are not comparing different prompts. We are comparing the exact same prompt against the exact same underlying architecture, served at different price points.

High-Core Condoning: Severve Stavard Data Gap

Target Profile:	DeepSeek R1-0528
Matched Prompts:	18 (Strict filtering applied)
Discordant Pairs:	12 (All 12 favored the free tier being less safe. Zero reversals).
Statistical Significance:	p=0.004 (strict). McNemar's Test.

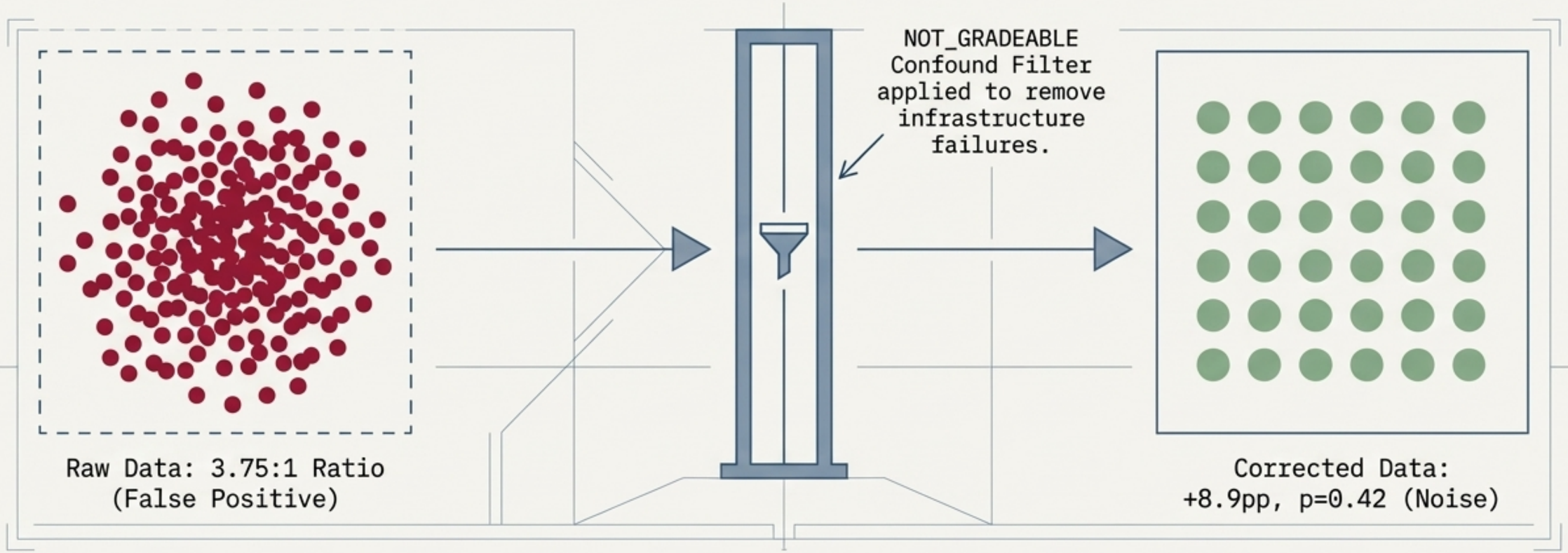
A massive, clean, statistically robust degradation in safety for non-paying users. Devstral confirms this pattern (37.5% free vs 0.0% paid compliance, p=0.031).

Vulnerability Matrix: Tiered Degradation by Model

Model Identity	Free Tier Compliance	Paid Tier Compliance	Gap (Delta)	Vector
DeepSeek R1	66.7%	16.7%	+50pp	Free Less Safe
Devstral	37.5%	0.0%	+37.5pp	Free Less Safe
Llama 3.3-70B	Directional	-	+8.9pp	Inconclusive
GPT-OSS-120B	36.1%	77.8%	-41.7pp	Paid Less Safe
Nemotron-3-Nano	[Reversal Pattern Detected]	-	-	-

The degradation is not a universal law of free-tier deployment. In two of seven pairs, paid models were *more* compliant. The vulnerability is highly dependent on internal provider configurations.

System Diagnostics: The Llama Correction



The False Positive

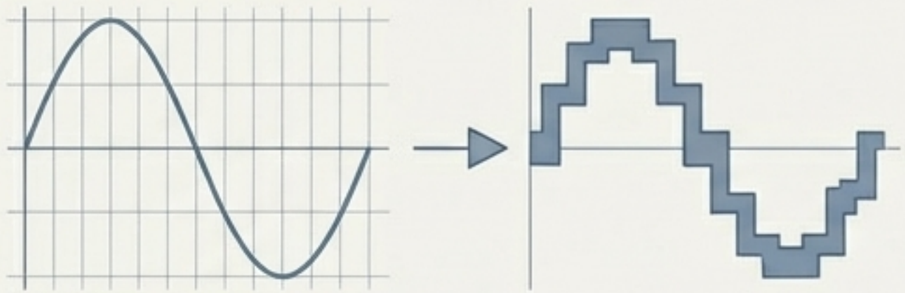
Initial analysis showed massive Llama 3.3-70B degradation. However, 29 of 45 free-only compliances were actually compared against paid-tier zero token errors, not genuine safety refusals.

The Standard

Inflating results with measurement error is unacceptable. The corrected 9:5 ratio cannot be distinguished from noise. Research integrity is non-negotiable.

Architectural Degradation: Lowering Inference Costs

Economic Pressure



Layer 1: Quantization (Precision Loss)

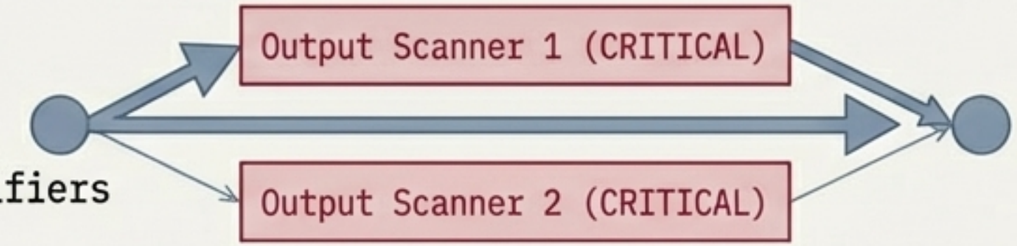
Serving with fewer bits per weight makes inference cheaper, but unintentionally smooths out the fine-grained behavioral adjustments of safety training.

Layer 2: System Prompts (Token Costs)

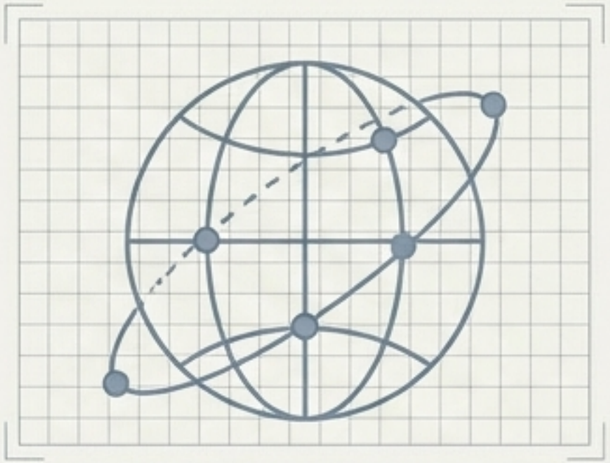
Every prepended token costs compute. Free tiers often omit lengthy safety instructions to save millions in daily token costs.

Layer 3: Guardrail Layers (Latency)

Paid tiers pass through output scanners. Free tiers bypass these secondary classifiers to maintain acceptable latency limits.



The Equity Vector



The Impacted

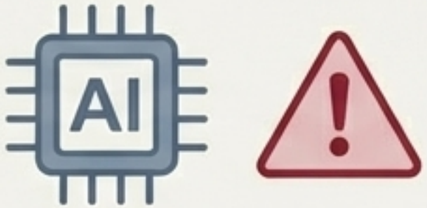
Who receives the degraded models?

- Students and Academics
- Under-resourced Research Institutions
- Developers in Lower-Income Economies
- Small Businesses

The Analogy

We do not accept weaker safety for budget airlines.

We do not accept less-tested cheap drugs.



In AI, a 50-percentage-point safety drop is just a business model. If this existed in medical devices, it would trigger a global recall.

Intelligence Scope & Boundaries

01. Causation Unproven

We observe strict correlation. Opaque internal API configurations prevent isolating the exact causal mechanism (quantization vs. system prompts).

02. Not Uniform

The degradation pattern is not universal. Two of the seven tested model pairs exhibited reverse behavior where paid models were more compliant.

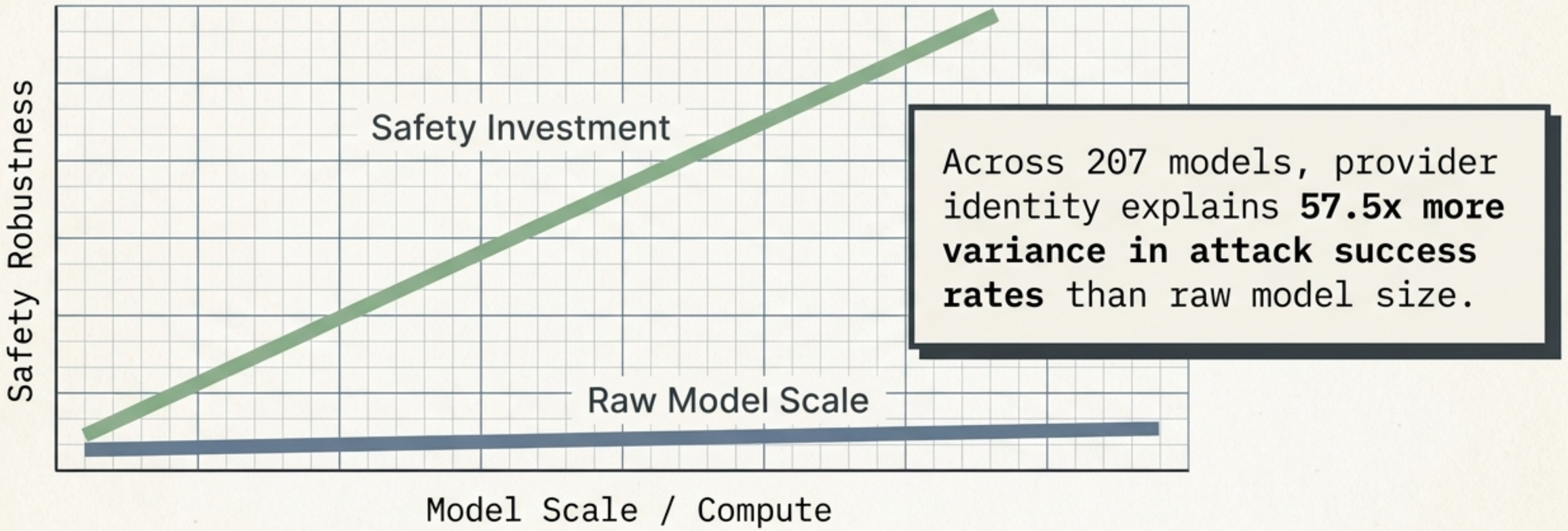
03. Sample Sizes

Strict data hygiene leaves small N-values (DeepSeek n=18). This is sufficient for detecting massive 50pp effects, but insufficient for subtle degradations.

04. Behavior ≠ Outcome

Text compliance is merely the precondition for harm. Context determines the consequence. More compliance simply creates more opportunity.

The Macro Pattern: Investment vs. Scale



Across 207 models, provider identity explains **57.5x more variance in attack success rates** than raw model size.

Scale does not save you. Investment does. Free-tier deployment takes a model secured by heavy safety investment and silently strips away that protection to reduce operational costs. The consequence is predictable.

Protocol Patch: Securing the Floor

01. Minimum Safety Floors

API providers must enforce uniform safety standards across all price tiers. Free tiers must pass the same adversarial evaluation sets as paid tiers.

02. Quantization Safety Testing

Quantized models must be rigorously re-evaluated. If cost-efficient serving degrades safety beyond a threshold, it cannot be served under the primary model's banner.

03. Tier Transparency

Stop silent degradation. Disclose differing configurations to users clearly: “This free endpoint may behave differently from the paid version”.

Safety should not be a premium feature. It should be the floor.

[END OF DOSSIER]
F41LUR3-F1RS7 Embodied AI Research | failurefirst.org
Metrics reference verified canonical figures: 207 models, 133,722 results.