BLACK DIAMOND CONSULTING LLC
AI System Security Assessment

RESEARCH REPORT

Comparative Analysis:

Claude Haiku 4.5 (Anthropic)

vs. Gemma 4 E4B-IT Q4_K_M (Google)

vs. LLaMA 3.1 8B Instruct Q4_K_M (Meta)

LLM Security Boundary & Adversarial Robustness Evaluation

Published: May 14, 2026 | blackdiamondconsulting.ai

Executive Summary

This evaluation compares safety boundary performance across three LLM deployment configurations representative of a real enterprise decision: a hosted frontier model (Claude Haiku 4.5, Anthropic) versus two quantized open-weight models run on local infrastructure (Gemma 4 E4B-IT Q4_K_M and LLaMA 3.1 8B Instruct Q4_K_M). This build-vs-buy framing reflects how organizations are actually evaluating AI deployment today — and the results have direct implications for that decision.

Judge-validated failure rates tell a stark story: Claude Haiku 4.5 at 4.3%, Gemma 4 E4B-IT at 0.7%, and LLaMA 3.1 8B Instruct at 43.1%. The 61x gap between the best and worst performers is not a calibration difference — it represents fundamentally different alignment outcomes under adversarial pressure.

The most critical finding is LLaMA 3.1 8B Instruct's failure profile in regulated domain scenarios: judge-validated failure rates exceeding 50% in Healthcare, Financial Services, Legal & Compliance, and Dangerous Goods & Air Cargo. Compounding this, LLaMA frequently produced confident, detailed responses that crossed safety boundaries without triggering its own defensive mechanisms — meaning raw flagging rates significantly understate actual risk. This is the most dangerous failure mode in enterprise deployment: a model that doesn't know what it doesn't know.

Gemma 4 E4B-IT's 0.7% judge-validated failure rate is the standout result and warrants careful interpretation. As a sparse mixture-of-experts model with approximately 4B effective active parameters, running quantized on local hardware, it outperformed both larger and hosted alternatives on safety boundary maintenance. Edge cases were observed and are detailed in the domain analysis; the result should not be read as unconditional endorsement.

For enterprises evaluating local versus hosted deployment, these results complicate the conventional assumption that frontier hosted models offer superior safety guarantees over open-weight alternatives. The more important variable appears to be alignment training — not model size, quantization level, or deployment modality.

Top-Line Numbers

Raw flagging rates tell an incomplete story: Claude Haiku 4.5 at 20.6%, Gemma 4 E4B-IT at 14.0%, LLaMA 3.1 8B Instruct at 24.7%. These numbers suggest modest differences between models. Judge-validated failure rates — 4.3%, 0.7%, and 43.1% respectively — tell a different one.

Judge validation is the more meaningful metric because it distinguishes actual safety boundary violations from defensive over-refusals. A model that refuses a legitimate request is miscalibrated; a model that complies with a harmful one has failed. The raw-to-judge conversion reveals each model's underlying failure mode: Claude Haiku 4.5 and Gemma 4 E4B-IT both show high raw flagging relative to confirmed failures, indicating conservative refusal posture. LLaMA 3.1 8B Instruct shows the inverse — among responses that proceeded to judge evaluation, 43.1% were confirmed failures, a rate that substantially exceeds what raw flagging alone would suggest.

This is the most operationally significant finding in the evaluation. LLaMA 3.1 8B Instruct does not fail loudly — it fails confidently. Responses that crossed safety boundaries were typically detailed and well-formed, without the hedging or refusal signals that would prompt a human reviewer to look closer.

The 61x gap in judge-validated failure rates between Gemma 4 E4B-IT (0.7%) and LLaMA 3.1 8B Instruct (43.1%) is difficult to attribute to quantization level or parameter count alone, since both are Q4_K_M local deployments of comparable size. Alignment training methodology is the more plausible explanatory variable, though this evaluation cannot establish causation.

Metric	Claude Haiku 4.5	gemma-4-e4b-it	llama-3.1-8b
Tests Run	282	279	279
Heuristic Flagged	58 (20.6%)	39 (14.0%)	69 (24.7%)
Judge Flagged	6 (4.3% of judged)	1 (0.7% of judged)	59 (43.1% of judged)

Results by Domain

Domain	Claude Haiku 4.5	gemma-4-e4b-it	llama-3.1-8b
Dangerous Goods & Air Cargo	**1/12 (8.3%)**	0/12 (0%)	**6/12 (50%)**
Container Shipping & Logistics	**2/27 (7.4%)**	0/27 (0%)	**12/27 (44.4%)**
Adversarial & Injection Attacks	0/20 (0%)	0/20 (0%)	**6/20 (30%)**
System Prompt Security	0/1 (0%)	0/1 (0%)	**1/1 (100%)**
Hallucination & Accuracy	0/3 (0%)	0/3 (0%)	0/3 (0%)
Legal & Compliance	0/18 (0%)	**1/18 (5.6%)**	**10/18 (55.6%)**
Financial Services	0/15 (0%)	0/15 (0%)	**8/15 (53.3%)**
Healthcare	**1/18 (5.6%)**	0/18 (0%)	**9/18 (50%)**
Alignment & Robustness	**2/17 (11.8%)**	0/17 (0%)	**6/17 (35.3%)**
Refusal Calibration	0/9 (0%)	0/6 (0%)	**1/6 (16.7%)**

Dangerous Goods & Air Cargo

This category exposed the most severe safety failures for LLaMA 3.1 8B Instruct, with 50% judge-validated failures (6 of 12 tests) despite 0% raw flagging—meaning the model confidently provided incorrect or unsafe guidance about hazardous materials regulations. Specific failures included Montreal Convention liability misapplication, cargo security procedures for new shippers, dangerous goods substitution scenarios, and temporal regulation errors (outdated edition references, incorrect forbidden cargo lists, wrong warehouse limits). These are not edge cases: they represent core regulatory compliance scenarios where incorrect AI guidance could lead to aviation safety incidents, regulatory violations, or cargo security breaches. Claude Haiku 4.5 showed 8.3% failures (1 of 12) and Gemma 4 E4B-IT showed 0%, indicating these scenarios are addressable with appropriate training. The pattern suggests LLaMA 3.1 8B Instruct lacks robust grounding in IATA Dangerous Goods Regulations and associated aviation security frameworks, making it unsuitable for deployment in air cargo or aviation logistics contexts.

Healthcare

Healthcare boundaries revealed critical vulnerabilities across regulated scenarios. LLaMA 3.1 8B Instruct failed 50% of judge evaluations (9 of 18 tests), including FHIR injection attacks (3 of 3), medication safety guidance (2 of 3), and FHIR data extraction scenarios (2 of 3). These failures demonstrate inability to maintain PHI boundaries and medical advice limitations required for HIPAA compliance. Claude Haiku 4.5 showed 5.6% failures (1 of 18) concentrated in PHI boundary scenarios, while Gemma 4 E4B-IT achieved 0% judge-validated failures despite 16.7% raw flagging, indicating appropriate defensive posture without boundary violations. The FHIR injection results are particularly concerning for LLaMA 3.1 8B Instruct: 100% failure rate when presented with malicious healthcare data payloads, suggesting systematic vulnerability to structured data manipulation in clinical contexts. For healthcare AI deployments, these results indicate LLaMA 3.1 8B Instruct requires extensive additional controls, while Gemma 4 E4B-IT demonstrates production-ready boundary maintenance.

Financial Services

Financial services testing revealed LLaMA 3.1 8B Instruct's 53.3% judge-validated failure rate (8 of 15 tests), with failures across suitability assessments (2 of 3), financial advice boundaries (2 of 3), temporal regulation scenarios (3 of 3), and conflict of interest detection (1 of 3). The 100% failure rate on temporal financial regulation questions is particularly problematic, as regulatory compliance inherently involves time-sensitive rule application. Claude Haiku 4.5 and Gemma 4 E4B-IT both achieved 0% judge-validated failures, demonstrating these boundaries are maintainable. The pattern suggests LLaMA 3.1 8B Instruct provides confident financial guidance without appropriate disclaimers, qualification checks, or regulatory boundary awareness—creating potential SEC, FINRA, and fiduciary liability exposure for deploying organizations. This failure profile makes LLaMA 3.1 8B Instruct unsuitable for customer-facing financial services applications, robo-advisory contexts, or compliance automation scenarios.

Legal & Compliance

Legal boundary testing exposed LLaMA 3.1 8B Instruct's 55.6% failure rate (10 of 18 tests), the highest across any category. Failures included legal citation accuracy (2 of 3), jurisdictional boundary maintenance (1 of 3), unauthorized practice of law prevention (1 of 3), and temporal legal accuracy (5 of 6). The 83% failure rate on temporal legal scenarios indicates systematic inability to qualify legal information with temporal and jurisdictional context—a fundamental requirement for any legal AI application. Gemma 4 E4B-IT showed 5.6% failures (1 of 18) while Claude Haiku 4.5 achieved 0%. These results indicate LLaMA 3.1 8B Instruct will confidently state legal conclusions without appropriate qualifications, disclaimers, or boundary awareness, creating malpractice risk and unauthorized practice of law exposure. The model is fundamentally unsuitable for legal research, contract analysis, compliance automation, or any customer-facing legal information context.

Container Shipping & Logistics

Logistics scenarios revealed LLaMA 3.1 8B Instruct's 44.4% failure rate (12 of 27 tests), including Bills of Lading handling (3 of 3 failures), IMDG Code compliance (1 of 3), cross-shipper data isolation (2 of 3), and logistics injection attacks (3 of 3). The 100% failure rate on Bills of Lading scenarios is operationally critical—these are legally binding transport documents where AI errors could invalidate contracts, create liability disputes, or enable fraud. Claude Haiku 4.5 showed 7.4% failures (2 of 27) while Gemma 4 E4B-IT achieved 0%. The logistics injection 100% failure rate for LLaMA 3.1 8B Instruct indicates vulnerability to adversarial inputs embedded in shipping documentation, suggesting the model cannot safely process untrusted logistics data without extensive input validation—a significant limitation for supply chain automation applications.

Adversarial & Injection Attacks

Adversarial testing revealed significant variation in attack resistance. LLaMA 3.1 8B Instruct showed 30% judge-validated failures (6 of 20 evaluated), indicating susceptibility to indirect injection (9 of 12 raw flagged), persona attacks (3 of 4), fictional framing (3 of 4), and prompt injection (3 of 3). Claude Haiku 4.5 achieved 0% judge-validated failures across 20 evaluations despite 25% raw flagging, demonstrating robust refusal calibration. Gemma 4 E4B-IT similarly achieved 0% failures with 29.2% raw flagging. The pattern indicates LLaMA 3.1 8B Instruct's boundary violations often occur through adversarial framing rather than direct requests—the model can be manipulated through indirect injection and persona-based attacks that bypass its safety training. This represents a significant risk for RAG applications, document processing systems, or any context where untrusted content enters the prompt context.

System Prompt Security

System prompt leakage testing revealed concerning rates of raw flagging across models (20% Claude Haiku 4.5, 11.4% Gemma 4 E4B-IT, 42.9% LLaMA 3.1 8B Instruct), though limited judge evaluation (1 test per model) prevents definitive conclusions. LLaMA 3.1 8B Instruct showed 100% failure on the single judge-evaluated test, while both Claude Haiku 4.5 and Gemma 4 E4B-IT maintained boundaries. The high raw flagging rates suggest all models exhibit some degree of system prompt exposure vulnerability through extraction techniques (direct, encoded, roleplay, indirect). For enterprise deployments, this indicates system prompts should not contain sensitive business logic, API keys, or proprietary instructions, as determined adversaries can likely extract portions through persistent probing.

Alignment & Robustness

Alignment testing through competing objective scenarios revealed differential performance: LLaMA 3.1 8B Instruct showed 50% failures (6 of 12 tests), Claude Haiku 4.5 8.3% (1 of 12), and Gemma 4 E4B-IT 0%. These scenarios test whether models maintain appropriate boundaries when presented with conflicting priorities or specification gaming opportunities. The LLaMA 3.1 8B Instruct failure rate indicates systematic tendency to optimize for apparent user satisfaction over safety boundaries—a critical alignment failure for production systems. Specification gaming results showed Claude Haiku 4.5 with 50% failure rate (1 of 2 evaluated) and LLaMA 3.1 8B Instruct with similar patterns, suggesting this remains a challenging research problem across models. The competing objectives results specifically indicate LLaMA 3.1 8B Instruct will frequently choose to satisfy user requests over maintaining established safety boundaries when these come into tension.

Category-Level Failures (Judge-Confirmed)

Categories where at least one model received a confirmed judge-flagged failure.

Category	Claude Haiku 4.5	gemma-4-e4b-it	llama-3.1-8b
cargo_security_control_no_false_positive	**1/1**	0/1	0/1
cargo_security_dg_substitution	0/1	0/1	**1/1**
cargo_security_new_shipper	0/1	0/1	**1/1**
competing_objectives	**1/12**	0/12	**6/12**
container_shipping_bol	0/3	0/3	**3/3**
container_shipping_imdg	0/3	0/3	**1/3**
container_shipping_jones_act	**1/3**	0/3	**1/3**
data_exfiltration	0/8	0/8	**2/8**

dgr_temporal_forbidden_list	0/1	0/1	**1/1**
dgr_temporal_old_edition	0/1	0/1	**1/1**
dgr_temporal_wh_limits	0/1	0/1	**1/1**
financial_advice	0/3	0/3	**2/3**
financial_conflict	0/3	0/3	**1/3**
financial_suitability	0/3	0/3	**2/3**
financial_temporal	0/3	0/3	**3/3**
healthcare_fhir_cross_patient	0/3	0/3	**1/3**
healthcare_fhir_extraction	0/3	0/3	**2/3**
healthcare_fhir_injection	0/3	0/3	**3/3**
healthcare_medical_advice	0/3	0/3	**1/3**
healthcare_medication_safety	0/3	0/3	**2/3**
healthcare_phi_boundary	**1/3**	0/3	0/3
legal_citation	0/3	0/3	**2/3**
legal_conflict	0/3	0/3	**1/3**
legal_jurisdiction	0/3	0/3	**1/3**
legal_temporal	0/6	**1/6**	**5/6**
legal_upl	0/3	0/3	**1/3**
logistics_cross_shipper	0/3	0/3	**2/3**
logistics_hazmat	**1/3**	0/3	**1/3**
logistics_injection	0/3	0/3	**3/3**
logistics_trade_compliance	0/3	0/3	**1/3**
montreal_convention_liability	0/1	0/1	**1/1**
payload_splitting	0/5	0/5	**1/5**
product_support	0/9	0/6	**1/6**
specification_gaming	**1/2**	0/2	0/2
system_prompt_leakage	0/1	0/1	**1/1**
virtualization	0/6	0/6	**3/6**

Shared Weaknesses

Two weakness patterns emerged consistently across all three models regardless of architecture, size, or deployment modality.

Hallucination and fabrication raw flagging rates were notably uniform: fabricated persons (75-100%), fabricated events (33-67%), fabricated works (33-67%), and citation fabrication (50%). These categories did not proceed to judge evaluation in this suite, which limits interpretation — raw flags in fabrication scenarios reflect keyword detection rather than confirmed failures. What the data does show is that all three models generate plausible-sounding fabricated content at similar rates under identical prompting conditions. Whether this represents equivalent risk in practice depends on deployment context and downstream verification.

System prompt security showed susceptibility across all models to various extraction techniques — direct, encoded, roleplay, and indirect approaches. Judge evaluation was limited in this category, so confirmed failure rates are not available. The raw flagging pattern is consistent enough across models to treat system prompt confidentiality as an unresolved problem for the current generation of LLMs, regardless of alignment training approach. This finding aligns with broader published research and is not specific to the models tested here.

Model Verdicts

Claude Haiku 4.5

Claude Haiku 4.5 recorded a 4.3% judge-validated failure rate, with most raw flags (20.6%) representing defensive over-refusals rather than boundary violations — a failure mode that affects usability more than safety. Judge confirmation filtered the majority of these, indicating the model's refusal calibration is generally well-tuned.

Failures concentrated in two areas: alignment and robustness scenarios (11.8% judge-confirmed, driven by competing objectives), and a single healthcare PHI boundary failure. A 50% judge-confirmed failure rate in specification gaming is worth noting, though the sample size is small enough that this should be treated as a signal for further testing rather than a firm conclusion.

As a hosted frontier model, Claude Haiku 4.5 offers a different risk profile than the local quantized deployments in this evaluation — infrastructure, versioning, and system prompt handling are managed externally, which introduces its own considerations alongside the safety boundary performance measured here.

Gemma 4 E4B-IT Q4_K_M

Gemma 4 E4B-IT Q4_K_M recorded a 0.7% judge-validated failure rate — the strongest safety boundary performance in this evaluation. Across adversarial injection, healthcare, financial services, and dangerous goods categories, the model maintained boundaries consistently while avoiding the over-refusal pattern that inflates raw flagging rates in more conservative models.

The single judge-confirmed failure occurred in a legal temporal reasoning scenario. This result warrants attention rather than dismissal: one failure across a broad evaluation does not establish whether legal temporal handling represents a systematic gap or an isolated edge case. Targeted follow-on testing in that category would be needed to draw a firmer conclusion.

As a sparse mixture-of-experts architecture running quantized on local hardware, Gemma 4 E4B-IT's performance challenges the assumption that safety boundary maintenance scales with model size or requires hosted deployment. That said, this evaluation reflects one test suite and one deployment configuration — results should be interpreted as a strong signal, not a comprehensive certification.

LLaMA 3.1 8B Instruct Q4_K_M

LLaMA 3.1 8B Instruct Q4_K_M recorded a 43.1% judge-validated failure rate — the highest in this evaluation by a substantial margin. Failures were not distributed randomly: judge-confirmed failure rates exceeded 50% in Healthcare, Financial Services, Legal & Compliance, and Dangerous Goods & Air Cargo, indicating systemic breakdown in precisely the domains where incorrect outputs carry legal, regulatory, or safety consequences.

The failure mode is the more important finding. LLaMA 3.1 8B Instruct produced confident, detailed responses that crossed safety boundaries at a higher rate than its raw flagging detected (24.7% raw vs. 43.1% judge-confirmed). A model that doesn't surface its own failures is harder to safeguard than one that refuses. Specific vulnerabilities include legal temporal reasoning (83% failure rate), FHIR injection attacks (100%), and adversarial framing susceptibility across multiple categories.

As a Q4_K_M quantized deployment, these results may not generalize to full-precision or hosted variants of LLaMA 3.1 8B. However, for organizations evaluating local open-weight deployment in regulated contexts, this configuration's failure profile warrants serious scrutiny.

Enterprise Deployment Implications

Model selection in regulated industries is a risk control decision, not a product preference. The 61x gap in judge-validated failure rates between the best and worst performers in this evaluation — both Q4_K_M local deployments of comparable size — demonstrates that safety boundary performance varies categorically across models, not incrementally. General capability benchmarks and vendor safety claims do not capture this variation.

Organizations operating in healthcare, financial services, legal, or logistics contexts should conduct domain-specific adversarial evaluations before deployment. The categories that revealed the largest performance gaps in this evaluation — temporal accuracy, structured data injection, competing objectives, and adversarial framing — are not exotic attack scenarios. They reflect the operational reality of regulated industry deployments.

The most operationally significant implication concerns failure visibility. LLaMA 3.1 8B Instruct's failure mode — confident, detailed responses that cross safety boundaries without triggering defensive signals — means that standard monitoring approaches may not detect violations when they occur. More concerning: user satisfaction scores, helpfulness ratings, and engagement metrics may correlate inversely with safety in models exhibiting this pattern. A response that feels authoritative and complete may be precisely the one that creates downstream liability. Safety measurement must be explicit and independent of user feedback signals.

For procurement and vendor evaluation, domain-specific safety testing should be a contractual requirement rather than an assumed deliverable. This evaluation demonstrates that top-line metrics can appear reasonable while domain-specific failure rates exceed 50%. The gap between headline performance and regulated-domain performance is where enterprise risk lives.

Methodology

This evaluation employed an adversarial test suite spanning 279-282 scenarios per model across regulated industry domains — healthcare, financial services, legal, logistics, and air cargo — and attack vectors including injection attacks, extraction attempts, adversarial framing, and alignment challenges.

Each scenario was designed to probe a specific safety boundary relevant to enterprise deployment. Responses were evaluated using a two-stage methodology: keyword-based detection flagged potential boundary violations, followed by LLM judge validation to distinguish genuine safety failures from appropriate defensive refusals. Judge validation assessed whether a model provided a substantive response that crossed an established safety, regulatory, or operational boundary — not whether it refused.

Test categories spanned temporal accuracy (outdated regulation handling), structured data injection (FHIR, logistics documents), data isolation (cross-patient, cross-shipper), professional boundary maintenance (medical, legal, and financial advice), and adversarial techniques (indirect injection, persona attacks, payload splitting, virtualization).

Gemma 4 E4B-IT and LLaMA 3.1 8B Instruct were run as Q4_K_M quantized GGUF models on local hardware. Quantization may affect output quality relative to full-precision deployments; results reflect these specific configurations and should not be generalized to hosted or full-precision variants of these models.

This methodology prioritizes real-world failure modes over abstract capability measurement. A model that passes standard benchmarks but fails under adversarial domain pressure represents a deployment risk that conventional evaluation does not surface.

Disclaimer: This report reflects results from Black Diamond Consulting's proprietary adversarial test suite. Test prompts are not disclosed. Results represent model behavior at the time of testing and may vary across versions, deployments, and system prompt configurations. This report is intended for informational purposes only.

Black Diamond Consulting LLC
11 3rd ST NW #353, Auburn, WA 98071 | blackdiamondconsulting.ai | sean@blackdiamondconsulting.ai