SAFE-AGENT-L: A Legal Compliance and Governance Framework for Autonomous LLM Agents in Large-Scale Retail Systems

Vasanth Rajendran
Amazon, Seattle, USA
vasraj@amazon.com

Abstract

Large Language Models (LLMs) are rapidly evolving from generative assistants into autonomous decision engines embedded within large-scale retail ecosystems. Modern retail platforms increasingly delegate high-impact tasks to LLM agents—such as generating product descriptions, composing storefronts, enriching structured attributes, classifying restricted goods, evaluating seller submissions, determining merchandising placements, and planning dynamic category layouts. Because these decisions directly modify customer-facing surfaces and marketplace states, they immediately propagate to millions of users and thousands of businesses.

Crucially, these operations take place within one of the most heavily regulated commercial environments, involving consumer protection laws, advertising rules, product safety mandates, restricted-goods regulations, comparative pricing laws, and regional compliance frameworks. Unlike conventional search and recommendation systems, autonomous LLM agents do not merely retrieve or rank content—they create new representations, facts, claims, and regulatory-sensitive descriptors. A single hallucinated medical claim, misrepresented price comparison, or misclassified restricted product can trigger statutory violations under regimes such as the FTC Act, CPSA, FDCA, EU UCPD, EU Omnibus Directive, CMA guidelines, and state consumer-protection statutes. Because LLM errors scale linearly with system throughput, even small error rates can lead to thousands of legal violations per hour.

This paper introduces SAFE-AGENT-L, the first integrated legal-compliance–assured governance framework designed specifically for autonomous LLM agents operating in retail and e-commerce systems. SAFE-AGENT-L comprises three synergistic layers: (1) Grounded Legal Alignment, which injects statutory constraints, policy rules, schema requirements, prohibited-claim lists, and jurisdiction-aware compliance filters into model reasoning; (2) Risk-Aware Action Governance, which computes a composite risk score using model uncertainty, violation-prediction models, and rule-sensitive classifiers; and (3) Multi-Stage Compliance Guardrails, which conduct deterministic validation, enforce mandatory overrides, escalate ambiguous outputs, or fall back to legally safe templates. SAFE-AGENT-L operationalizes legal requirements directly into agent behavior and provides a rigorous mechanism for preventing, detecting, and mitigating legally non-compliant model actions.

By conceptualizing retail as a legally regulated embodied environment—where agents must navigate structured constraints, irreversible actions, and jurisdiction-specific state transitions—SAFE-AGENT-L enables safe and verifiable autonomy at global scale. We also propose measurable compliance metrics aligned with real regulatory audits and present evaluation methodologies suitable for deployment contexts requiring traceability, auditability, and high-assurance trustworthiness.

Introduction

The emergence of autonomous LLM agents marks a structural shift in how retail platforms manage content, compliance, and decision-making. Until recently, the core functions of online marketplaces—catalog ingestion, product description generation, merchandising curation, pricing presentation, detail-page construction, and product safety classification—were executed either manually or via deterministic pipelines. Today, large-scale retailers increasingly rely on LLM-driven agents to perform these operations autonomously, with minimal human supervision, and at global scale.

However, unlike previous forms of automation, LLM agents produce novel linguistic, semantic, and structured outputs. This generative autonomy represents both an opportunity and a risk. While it drastically improves speed, adaptability, and coverage, it also creates the possibility that a model will hallucinate factual attributes, misinterpret seller data, misclassify dangerous goods, or generate advertising claims that violate legal standards.

Consider several examples:

- A model incorrectly claims a dietary supplement is “FDA approved.”
- An LLM mislabels a toy as “safe for children under 3” without evidence.
- A pricing-generator agent produces false “50% off” messaging.
- A restricted-goods classifier fails to flag a pesticide unregistered for sale in the EU.
- A product-description generator adds creative—but illegal—superlatives such as “guaranteed to cure migraines.”

Each of these outputs constitutes a prosecutable violation under current regulations and can trigger multi-million-dollar enforcement actions, lawsuits, or marketplace sanctions.

Because retail operates under a dense web of legal regimes—including consumer protection law, comparative pricing standards, marketing and advertising rules, product-safety laws, chemical regulations, and age-restriction mandates—LLM autonomy cannot be treated solely as a technical problem. It is a legal-compliance problem embedded within a technical system.

SAFE-AGENT-L addresses this gap by providing a principled, operational, and enforcement-aligned framework that transforms legal requirements into computable constraints and agent behaviors. This ensures that every LLM-generated action can be validated for compliance before reaching production.

Retail as a Legally Regulated Environment

Retail content generation differs from other AI application domains because every output is subject to mandatory legal obligations. These obligations vary by jurisdiction, product category, marketplace role, and claim type. The SAFE-AGENT-L framework treats these legal constraints as part of the environment dynamics that the LLM agent must navigate.

Consumer Protection and Deceptive Practices Law

In the United States, the Federal Trade Commission (FTC) enforces prohibitions against unfair or deceptive acts under Section 5 of the FTC Act. This includes:

- unsubstantiated health or performance claims,
- inaccurate pricing comparisons,
- deceptive statements about product capabilities,
- omissions of material safety information.

The EU's Consumer Rights Directive (CRD), Unfair Commercial Practices Directive (UCPD), and Digital Services Act (DSA) impose parallel obligations. These statutes treat consumers as vulnerable parties entitled to accurate, complete, and verifiable information.

Advertising and Marketing Regulation

Many LLM-generated product descriptions qualify as advertising statements. This invokes additional regulatory frameworks:

- Lanham Act (U.S.): prohibits false or misleading advertising.
- EU Omnibus Directive: requires transparent discount disclosure.
- UK CMA Pricing Practices Guide: governs "was/now" pricing claims.

LLM-generated exaggerations or superlatives may qualify as deceptive advertising, exposing retailers to litigation from consumers and competitors.

Product Safety and Liability

Product-safety classifications determine whether items comply with:

- CPSC (U.S.) child-product safety rules,
- EU General Product Safety Regulation (GPSR),
- chemical bans under EU REACH and California Proposition 65,
- flammability, choking hazard, and toxicity standards.

Incorrect LLM classification of safety-critical attributes can lead to forced recalls or liability for harm.

Restricted Goods Regulation

LLM agents must accurately detect:

- age-restricted goods (alcohol, tobacco alternatives, knives),
- hazardous materials (lithium batteries, corrosives),
- unregistered pesticides,
- controlled chemical products,
- ingestible goods requiring region-specific approvals.

Jurisdiction misalignment (e.g., allowing EU customers to access U.S.-legal but EU-banned pesticides) is a high-severity violation.

Related Work

LLM Safety Research

Existing AI-safety research primarily focuses on general-purpose concerns such as:

- toxic content generation,
- jailbreak resistance,
- hallucination mitigation,
- content filtering.

These approaches do not address retail-specific constraints.

Legal AI and Compliance Systems

Legal-AI systems explore automated contract analysis, case-law reasoning, and high-level regulatory modeling. However, they rarely provide:

- real-time compliance enforcement,
- production-level constraints,
- region-specific model overrides,
- verifiable audit traces for regulatory defense.

SAFE-AGENT-L differs by embedding legal rules directly into the LLM action pipeline.

Retail AI and Marketplace Systems

Retail AI research historically addresses:

- ranking and personalization,
- recommendation systems,
- content retrieval,
- logistics optimization.

No known system operationalizes real-time legal constraints over generative agents.

SAFE-AGENT-L Architecture

SAFE-AGENT-L is built on three interlocking layers that ensure autonomous LLM agents comply with legal and operational constraints.

Layer 1: Grounded Legal Alignment

At the first layer, SAFE-AGENT-L integrates legal rules into the generation process using:

- jurisdiction-tagged JSON schemas,
- explicit prohibited-claims lists,
- approved vocabulary dictionaries,
- cross-product category compliance rules,
- evidence-linked attribute generation,
- region-aware claim suppression.

This grounding ensures that the model cannot generate content outside the legally permissible space.

Layer 2: Risk-Aware Action Governance

Each candidate model output is scored using:

- Uncertainty Analysis $U(a)$: entropy, softmax variance, or LLM self-evaluation.
- Violation Predictor $V(a, s)$: classifiers trained on historical compliance violations.
- Constraint-sensitive detectors: identify high-risk attributes (medical claims, pricing claims).

The risk model:

$$R(a, s) = \alpha U(a) + \beta V(a, s)$$

prioritizes recall for high-severity legal violations.

Layer 3: Multi-Stage Compliance Guardrails

Outputs undergo deterministic checks:

- hard-block rules for prohibited claims,
- numerical consistency checks for pricing,
- region-specific suppression,
- routing to human review if $R(a, s)$ exceeds threshold τ,
- safe fallbacks (neutral, factual statements).

This layer ensures zero-tolerance for legally sensitive errors.

Embodied Interpretation of Retail Agents

Retail autonomy resembles embodied decision-making because agents operate in structured, constraint-rich environments where:

- actions cause irreversible state transitions (publishing product data),
- inputs represent environment signals (catalog data, price graphs),
- outputs affect downstream processes (inventory, ranking, promotions),
- jurisdiction rules shape feasible actions,
- policy constraints act as "physics-like" boundaries.

Thus, retail autonomy requires the same safety rigor as embodied robotics systems.

Evaluation Framework

To evaluate SAFE-AGENT-L, we propose metrics aligned with regulatory audits:

- Compliance Violation Rate: proportion of outputs violating legal rules.
- False Negative Rate: violations not detected by the system.
- Safe Output Yield: proportion of legally compliant outputs.
- Guardrail Success Rate: percent of prohibited claims intercepted.
- Latency Overhead: additional runtime cost introduced by safety layers.

Metric	Target
Compliance Violation Rate	$< 0.01\%$
Safe Output Yield	$> 92\%$
Guardrail Success Rate	$> 99.9\%$
Latency Overhead	< 2s

Case Studies

This section presents three detailed case studies illustrating how SAFE-AGENT-L governs real-world risk in legally regulated retail environments. These case studies were selected because they represent high-frequency, high-severity failure modes that directly create statutory exposure or reputational damage. They also map cleanly to established enforcement domains such as advertising compliance, consumer protection, marketplace safety, and product liability. In each scenario, we analyze (1) how autonomous LLM agents typically fail, (2) the legal rules implicated, and (3) how SAFE-AGENT-L prevents or mitigates these failures.

Attribute Enrichment

Attribute enrichment is one of the most critical and risk-sensitive LLM tasks in contemporary retail platforms. In this workflow, an agent receives unstructured seller text,

noisy metadata, or multimodal product imagery and generates structured attributes such as material composition, usage instructions, age suitability, health-related properties, allergens, chemical disclosures, or regulatory certifications.

Without a legal-compliance framework, LLMs frequently hallucinate regulated attributes, including:

- Health claims ("clinically proven") governed by FTC and FDA.
- Safety claims ("non-toxic," "safe for infants") regulated under CPSA and EU GPSR.
- Organic / eco certifications ("USDA Organic," "ISO-certified") that require third-party verifiable proof.
- Dietary restrictions ("gluten-free," "nut-free") without validated testing.
- Ingredient disclosures that misrepresent allergens or banned substances.

Each hallucination is a potential regulatory violation. Under FTC policy, any unsubstantiated claim is considered deceptive advertising. Under the EU's Unfair Commercial Practices Directive, unverifiable claims are automatically illegal. Thus, an autonomous agent enriching attributes at scale could unintentionally produce thousands of illegal product listings in minutes.

SAFE-AGENT-L prevents this through:

- Grounded Legal Alignment: Schema constraints explicitly disallow claims belonging to regulated categories unless documentary evidence is present.
- Risk-Aware Governance: Violation prediction models score proposed attributes against a library of known restricted claims.
- Guardrails: Deterministic filters block high-risk attribute classes entirely (e.g., medical claims, pesticide claims, baby-safety claims).
- Fallbacks: Instead of hallucinating a restricted attribute, the agent outputs "Information Not Provided" or escalates to human review.

In a simulation of 50,000 random retail SKUs, SAFE-AGENT-L reduced illegal attribute hallucinations from 8.1% (baseline LLM) to 0.04%. This performance meets the internal compliance target commonly used for regulated marketplace surfaces.

Restricted Goods Classification

Many product categories are regulated across global regions, including:

- Age-restricted goods (alcohol, knives, tobacco alternatives, OTC medications)
- Hazardous materials (lithium batteries, corrosive chemicals, flammable liquids)
- Controlled supplements and compounds
- Pesticides, insecticides, and agricultural products
- Child safety–sensitive goods (sleepwear, cribs, high chairs)

LLMs often misclassify these categories due to:

- misleading seller descriptions,
- ambiguous product titles,
- hallucinations from incomplete metadata.

A misclassified restricted good can instantly violate:

- U.S. EPA FIFRA (pesticide registration),
- EU REACH chemical restrictions,
- US CPSC child-product safety standards,
- UK Trading Standards age-restriction laws.

SAFE-AGENT-L addresses restricted goods risks by:

- Constraint-first prompting: The agent receives explicit jurisdiction-specific rules prior to processing.
- Restricted Goods Classifier: A lightweight model predicts whether an item falls within a legally sensitive class.
- Risk Scoring: Items with ambiguous or incomplete metadata receive a high $R(a, s)$ score and are automatically blocked or escalated.
- Region-Aware Overrides: For example, certain pesticides banned in the EU but legal in the U.S. are automatically suppressed in EU regions.

This approach reduced uncontrolled restricted-goods visibility by 97% in test environments—a result far exceeding traditional keyword-based filters.

Pricing Representations

Pricing is one of the most heavily litigated and regulated domains in retail. Autonomous LLM agents that generate promotional text, callouts, comparison pricing, or discount explanations frequently introduce legal exposure.

Common LLM pricing errors include:

- Incorrect "Was" pricing (e.g., inflated reference prices)
- Misleading discount labels ("50% off" when actual discount is less)
- Impermissible "lowest price ever" claims
- Regional price misalignment (currency mixing, VAT errors)
- Legally restricted phrases ("guaranteed savings")

These errors violate:

- FTC Pricing Truth-in-Advertising Guidelines,
- California's Comparison Pricing Law,
- UK CMA Pricing Practices Guide,
- EU Omnibus Directive on discount transparency.

SAFE-AGENT-L mitigates these risks by:

- performing numeric consistency checks between generated text and real-time price inputs,
- using hard constraints and schemas that forbid unverifiable superlatives,
- mapping output phrases to a legally allowed vocabulary,
- enforcing regional rule overrides for discount representations.

During internal evaluations, SAFE-AGENT-L reduced pricing misrepresentation rates from 3.5% to effectively 0%.

Ablation Experiments

We performed ablation analyses to quantify the contribution of each SAFE-AGENT-L layer. Experiments were conducted using 20,000 product samples across apparel, OTC health products, household chemicals, and electronics.

Without Grounded Legal Alignment

Removing legal grounding increased:

- medical claim hallucinations by 312%,
- restricted goods misclassification by 187%,
- pricing claim violations by 91%.

Without Risk-Aware Governance

Removing the risk model increased overall compliance violations from 0.04% to 2.7%. High-uncertainty cases accounted for most errors.

Without Multi-Stage Guardrails

When deterministic guardrails were disabled:

- false negatives increased sharply,
- prohibited product claims bypassed validation,
- restricted goods surfaced in multiple unsafe regions.

Societal Impact

Deployment of autonomous LLM agents in retail has broad implications:

- Consumer Risk Reduction: SAFE-AGENT-L prevents false medical claims, misleading pricing, and dangerous product mislabeling.
- Marketplace Integrity: Ensures sellers cannot exploit AI-generated content to introduce deceptive claims.
- Fair Competition: Compliant agents reduce anticompetitive practices related to deceptive promotions.
- Regulatory Accountability: Provides auditability aligned with FTC, CMA, EU DSA, and state consumer-protection enforcement.

Without such systems, LLM-driven retailers risk systemic legal violations, large-scale consumer harm, and repeated penalty actions.

Limitations

SAFE-AGENT-L has several limitations:

- Ambiguity Handling: Some legal categories remain inherently ambiguous (e.g., "wellness claims").
- Jurisdiction Explosion: Full global compliance imposes a combinatorial regulatory burden.
- Dependency on Upstream Data: Poor metadata quality reduces LLM accuracy even with safety layers.
- Human Escalation Load: Excessively cautious settings may increase manual review volume.

Conclusion

SAFE-AGENT-L is the first end-to-end compliance-assured governance framework for autonomous LLM agents operating in legally regulated retail environments. By combining grounded legal alignment, risk-aware action governance, and deterministic guardrails, it enables verifiable and trustworthy AI-driven retail automation. As retail platforms transition toward autonomous decision engines, SAFE-AGENT-L provides a legally robust blueprint for ensuring that innovation aligns with regulatory obligations, marketplace fairness, and consumer protection.

References

Federal Trade Commission. Federal Trade Commission Act, Section 5.

U.S. Food and Drug Administration. Federal Food, Drug, and Cosmetic Act.

U.S. Consumer Product Safety Commission. Consumer Product Safety Act.

European Union. Unfair Commercial Practices Directive (2005/29/EC).

European Union. Omnibus Directive (EU) 2019/2161.

UK Competition and Markets Authority. Pricing Practices Guide.

European Union. General Product Safety Regulation (EU) 2023/988.

European Chemicals Agency. REACH Regulation (EC) No 1907/2006.