# 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

## Objectives

SAFE-AGENT-L addresses critical legal compliance challenges in autonomous LLM-driven retail systems:

- Prevent hallucinated medical claims, false pricing, and misclassified restricted goods
- Operationalize statutory constraints from FTC, CPSA, FDCA, EU UCPD
- Enable verifiable, auditable AI autonomy at global scale
- Reduce legal violations from 8.1% to <0.04%

## Introduction

Modern retail platforms delegate high-impact tasks to **autonomous LLM agents**:

- Product description generation
- Attribute enrichment
- Restricted goods classification
- Pricing representation

**The Problem:**

- Hallucinate regulated claims
- Misrepresent pricing
- Misclassify dangerous goods
- Violate advertising standards

**Legal Impact:** Each error is a prosecutable violation triggering multi-million-dollar enforcement actions.

**Scale:** Small error rates produce thousands of violations per hour.

## Legal Environment

**Consumer Protection**

- FTC Act Section 5
- EU UCPD, DSA

**Product Safety**

- CPSC standards
- EU GPSR
- REACH, Prop 65

**Advertising**

- Lanham Act
- EU Omnibus Directive
- UK CMA Guide

**Restricted Goods**

- Age restrictions
- Hazardous materials
- Pesticides

## Related Work

**LLM Safety:** Focuses on toxic content, jailbreaks—not retail constraints

**Legal AI:** Contract analysis, regulatory modeling—lacks real-time enforcement

**Retail AI:** Ranking, recommendations—no legal constraint operationalization

*SAFE-AGENT-L fills this gap*

## Layer 1: Grounded Legal Alignment

Integrates legal rules into generation:

- Jurisdiction-tagged schemas
- Prohibited-claims lists
- Approved dictionaries
- Category compliance rules
- Evidence-linked attributes
- Region-aware suppression

*Model cannot generate content outside legally permissible space*

## Layer 2: Risk-Aware Governance

Composite risk scoring:

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

**Components:**

- $U(a)$: Uncertainty (entropy, variance)
- $V(a,s)$: Violation predictor
- Constraint-sensitive detectors

Optimized for high-severity violations

## Layer 3: Compliance Guardrails

Multi-stage validation:

1. Hard-block prohibited claims
2. Numerical consistency checks
3. Region-specific filters
4. Human escalation if $R(a,s) > \tau$
5. Safe fallback templates

Zero-tolerance for legal errors

## Embodied Retail Autonomy

Retail agents operate like embodied systems:

- Irreversible state transitions
- Environment signals
- Downstream effects
- Jurisdiction-shaped actions
- "Physics-like" boundaries

Requires robotics-level safety rigor

## Key Results

**Production-grade legal compliance achieved:**

- Illegal attribute hallucinations: **8.1% → 0.04%**
- Restricted goods misclassification: **-97%**
- Pricing misrepresentation: **3.5% → 0%**
- Guardrail success rate: **>99.9%**

## Case Study 1: Attribute Enrichment

**Challenge:** Hallucinated regulated attributes

- Health claims ("clinically proven")
- Safety claims ("safe for infants")
- Certifications ("USDA Organic")
- Dietary restrictions

**Legal Risk:** FTC/EU violations

**Solution:**

- Schema constraints
- Violation prediction
- Deterministic filters
- Safe fallbacks

**Result:** 8.1% → 0.04% (50K SKUs)

## Case Study 2: Restricted Goods

**Challenge:** Misclassified regulated products

- Age-restricted goods
- Hazardous materials
- Pesticides

**Legal Risk:** EPA, REACH, CPSC violations

**Solution:**

- Constraint-first prompting
- Restricted goods classifier
- Risk scoring

## Case Study 3: Pricing

**Challenge:** Illegal pricing claims

- Incorrect "Was" pricing
- Misleading discounts
- Impermissible superlatives
- Regional misalignment

**Legal Risk:** FTC, CMA, EU violations

**Solution:**

- Numeric consistency checks
- Hard constraints
- Allowed vocabulary
- Regional overrides

**Result:** 3.5% → 0%

## Evaluation Metrics

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

Aligned with regulatory audits

## Ablation Studies

Testing 20,000 products:

**Without Legal Alignment:**

- Medical claims: +312%
- Restricted goods: +187%
- Pricing: +91%

**Without Risk Governance:**

- Violations: 0.04% → 2.7%

**Without Guardrails:**

- Sharp increase in false negatives
- Prohibited claims bypass validation

⇒ All layers essential

## Societal Impact

**Consumer Protection:**

- Prevents false claims
- Eliminates misleading pricing
- Blocks dangerous mislabeling

**Marketplace Integrity:**

- Prevents seller exploitation
- Ensures fair competition

**Regulatory Accountability:**

- Audit trails for FTC, CMA, EU DSA
- Regulatory defense capability

## Limitations

- Legal ambiguity in some categories
- Combinatorial jurisdiction burden
- Metadata quality dependency
- Manual review escalation load

## Conclusion

SAFE-AGENT-L is the **first end-to-end compliance-assured governance framework** for autonomous LLM agents in retail.

Enables verifiable and trustworthy AI-driven retail automation through grounded legal alignment, risk-aware governance, and deterministic guardrails.

Provides a legally robust blueprint for innovation aligned with regulatory obligations and consumer protection.

## References

- Federal Trade Commission Act, Section 5
- Federal Food, Drug, and Cosmetic Act
- Consumer Product Safety Act
- EU Unfair Commercial Practices Directive (2005/29/EC)
- EU Omnibus Directive (EU) 2019/2161
- UK CMA Pricing Practices Guide
- EU General Product Safety Regulation (EU) 2023/988
- EU REACH Regulation (EC) No 1907/2006

## Contact Information

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