SAFE-AGENT: A Safety-Assured Framework for Embodied LLM Autonomy in Large-Scale Retail & E-Commerce Decision Systems

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

Abstract

Large Language Models (LLMs) are rapidly transitioning from assistants to autonomous decision-making agents deployed across high-stakes operational environments. In global-scale retail and e-commerce systems, LLM agents generate storefronts, curate merchandising layouts, enrich product attributes, and construct navigational experiences for millions of customers. Although these environments are digital, the agents are **embodied** within structured world models that enforce strict constraints, exhibit irreversible state transitions, and demand consistent policy alignment. These characteristics make retail autonomy safety-critical, much like embodied robotic systems.

We introduce **SAFE-AGENT**, the first safety-assured framework for embodied LLM autonomy in retail. The architecture integrates three layers—(1) Grounded Decision Alignment, (2) Risk-Aware Action Governance, and (3) Multi-Stage Guardrail Enforcement—alongside a formal risk-bounded validator that provides deterministic rejection of unsafe actions. This extended abstract articulates a full problem formulation, architectural breakdown, evaluation plan, and case studies demonstrating how SAFE-AGENT enables robotics-grade safety in the next generation of autonomous LLM systems.

1 Introduction

LLMs are increasingly integrated into decision pipelines traditionally reserved for symbolic systems or human editors. In retail, these agents:

- generate category browse pages,
- select and arrange UI modules,
- enrich or correct attributes,
- decide which deals, content, and assets appear,
- construct multi-step customer journeys.

These actions have direct customer impact at planetary scale and are often irreversible once deployed. A hallucinated size chart, fabricated product specification, or non-compliant storefront can lead to:

- regulatory violations,
- contractual breaches with sellers,
- customer harm or confusion,
- millions of dollars in lost revenue.

The safety-criticality of these autonomous agents mirrors that of robotics. Both systems:

- operate under uncertainty,
- must follow strict constraints,
- take actions that modify world state,
- require verifiable control loops,
- risk catastrophic failure if unaligned.

Because of these parallels, we propose applying robotics-grade safety techniques to LLM autonomy in digital retail systems. SAFE-AGENT operationalizes this insight.

2 Motivation: Retail as an Embodied, Safety-Critical Domain

Retail environments are often misunderstood as "low stakes" because they lack physical actuators. However, the consequences of unsafe LLM behavior can be severe.

2.1 Failure Mode Examples

Case 1: Hallucinated Safety-Critical Attribute
An LLM enriches a product with incorrect voltage or material information. Downstream effects include customer harm, product recalls, liability exposure, and regulatory penalties.

Case 2: Non-Compliant Deal Promotion
An LLM selects a deal or category for which promotion is legally restricted in certain regions. Result: fines, contract violations, or brand trust erosion.

Case 3: Broken Customer Journey
An LLM restructures a browse page in a way that hides required legal disclaimers or misroutes customers to unavailable inventory, causing widespread confusion and drop-offs.

Case 4: Seller Policy Violations
Incorrectly linking incompatible ASINs can trigger escalations, legal disputes, or forced takedowns.

These failure modes establish the need for autonomous safety systems akin to those used in robotics.

3 Problem Definition and System Model

We formalize retail autonomy as a constrained embodied decision process.

3.1 State Space

Retail environment state includes:

$$s_t = \{C, P, M, R, U\}$$

where:

- C: catalog metadata,
- P: policy graph,
- M: merchandising context,
- R: regional and legal constraints,
- U: user intent or journey signals.

3.2 Action Space

An LLM agent selects:

$$a_t \in \mathcal{A} = \{\text{layouts}, \text{attributes}, \text{slots}, \text{copy}, \text{recommendations}\}.$$

3.3 Transition Function

The environment transitions as:

$$s_{t+1} = f(s_t, a_t),$$

and transitions are often irreversible due to caching or deployment pipelines.

3.4 Constraints

Constraints K include:

- category schemas,
- brand rules,
- compliance regulations,
- layout grammars,
- safety policies.

An action is valid iff:

$$a_t \models K.$$

4 SAFE-AGENT Architecture

4.1 Layer 1: Grounded Decision Alignment

Injects immutable facts:

- category schemas,
- legal rules,
- policy graphs,
- regional constraints.

Techniques include:

- schema-constrained decoding,
- retrieval-based grounding,
- prompt-level safety encoding.

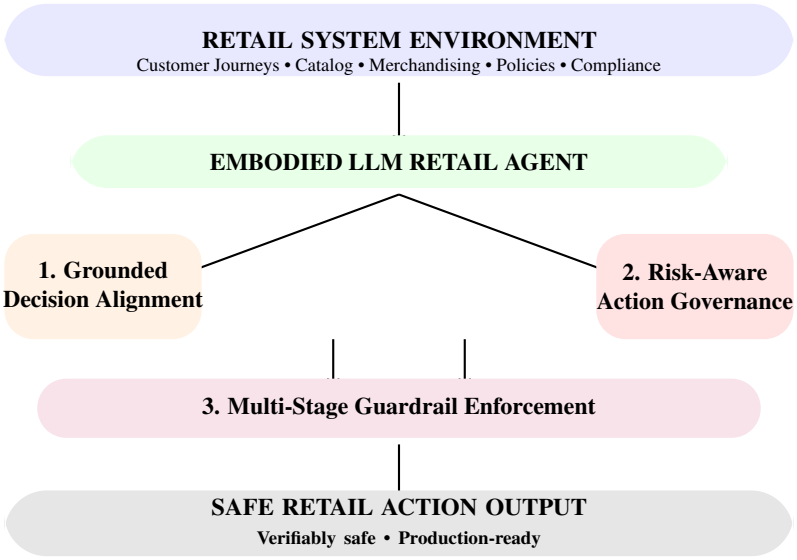

Figure 1: SAFE-AGENT three-layer safety architecture.

4.2 Layer 2: Risk-Aware Action Governance

Predicts violation likelihood:

$$R(a, s) \in [0, 1].$$

Tools:

- uncertainty estimation,
- counterfactual prompting,
- ensemble verification models,
- rule-based detectors.

4.3 Layer 3: Multi-Stage Guardrail Enforcement

Actions are blocked, rewritten, or escalated.

4.4 Formal Validator

$$\pi_{\text{safe}}(a|s) = \begin{cases} 0 & R(a, s) > \tau \\ 1 & R(a, s) \leq \tau. \end{cases}$$

5 Case Studies

5.1 Case Study 1: Hallucinated Specification

An agent invents "100% steel blade" for a toy. SAFE-AGENT blocks the action due to material-policy mismatch.

5.2 Case Study 2: Illegal Regional Deal

A deal advertised in a restricted jurisdiction is blocked by regional constraints in Layer 1 and flagged in Layer 2.

6 Related Work

Embodied AI research focuses on robotics, autonomous navigation, manipulation, and sensorimotor control. Formal verification and shielded control have been proposed in robotics and autonomous driving. LLM safety research explores hallucination mitigation, retrieval grounding, and model-based oversight, but no prior work treats retail as an embodied safety-critical domain. SAFE-AGENT bridges these areas.

7 Evaluation Plan

We simulate a catalog with 20 constraint types and evaluate:

- violation detection,
- safe-output yield,
- latency overhead,
- ablations (no-grounding, no-guardrails).

Metric	Target
Violation Reduction	$> 90\%$
Guardrail Success	$> 99.9\%$
Safe Output Yield	$> 92\%$
Latency Overhead	$< 1.8s$

Table 1: Evaluation targets.

8 Discussion

SAFE-AGENT raises open research needs:

- multi-agent consistency,
- real-time verification,
- long-horizon safety assurance,
- drift monitoring,
- cross-modal (image + text) safety.

9 Ethical & Societal Impact

A safety framework prevents misinformation, consumer harm, unfair seller treatment, and compliance violations. SAFE-AGENT aligns with responsible deployment practices for autonomous systems.

10 Conclusion

SAFE-AGENT introduces a robotics-inspired safety framework for embodied LLM autonomy in retail, providing grounding, risk governance, and guardrails required for dependable large-scale deployment. This work lays the foundation for verifiable, high-confidence LLM autonomy in safety-critical digital ecosystems.