Novelty Assessment Report

Paper: Building a Foundational Guardrail for General Agentic Systems via Synthetic Data
PDF URL: https://openreview.net/pdf?id=M47SWYubR5
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

While LLM agents can plan multi-step tasks, intervening at the planning stage—before any action is executed—is often the safest way to prevent harm, since certain risks can lead to severe consequences once carried out. However, existing guardrails mostly operate post-execution, which is difficult to scale and leaves little room for controllable supervision at the plan level. To address this challenge, we highlight three critical gaps in current research: data gap, model gap, and evaluation gap. To close the data gap, we introduce AuraGen, a controllable engine that (i) synthesizes benign trajectories, (ii) injects category-labeled risks with calibrated difficulty, and (iii) filters outputs via an automated reward model, producing large and reliable corpora for pre-execution safety. To close the guardian model gap, we propose a foundational guardrail Safiron, combining a cross-planner adapter with a compact guardian model. The adapter unifies different input formats, while Safiron flags risky cases, assigns risk types, and generates rationales; trained in two stages with a broadly explored data recipe, Safiron achieves robust transfer across settings. To close the evaluation gap, we release \texttt{Pre-Exec Bench}, a realistic benchmark covering diverse tools and branching trajectories, which measures detection, fine-grained categorization, explanation, and cross-planner generalization in human-verified scenarios. Extensive experiments demonstrate consistent gains over strong baselines on Pre-Exec Bench, and ablations further distill actionable practices, providing a practical template for safer agentic systems.

Core Task Landscape

This paper addresses: **Pre-Execution Safety Guardrails for LLM-Based Agentic Systems**

A total of **50 papers** were analyzed and organized into a taxonomy with **28 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Guardrail Architectures and Enforcement Mechanisms**
- **Safety Evaluation and Benchmarking**
- **Threat Models and Attack Vectors**
- **Domain-Specific Safety Applications**
- **Human-Agent Interaction and Oversight**
- **Adaptive and Learning-Based Safety Mechanisms**
- **Formal Verification and Logic-Based Safety**
- **Multi-Agent Safety and Moderation**
- **Foundational Concepts and Surveys**
- **Auxiliary and Cross-Cutting Topics**

Complete Taxonomy Tree

- Pre-Execution Safety Guardrails for LLM-Based Agentic Systems Survey Taxonomy
- Guardrail Architectures and Enforcement Mechanisms
  - Multi-Stage Guardrail Frameworks ★ (4 papers)
  - [0] Building a Foundational Guardrail for General Agentic Systems via Synthetic Data (Anon et al., 2026) View paper
  - [1] Llamafirewall: An open source guardrail system for building secure ai agents (Sa-hana Chennabasappa, 2025) View paper
  - [10] Trustagent: Towards safe and trustworthy llm-based agents through agent constitution (Hua, 2024) View paper
  - [16] Trustagent: Towards safe and trustworthy llm-based agents (Hua Wenyue, 2024) View paper
  - Specification-Based Runtime Enforcement (2 papers)
  - [12] AgentSpec: Customizable Runtime Enforcement for Safe and Reliable LLM Agents (Wang Haoyu, 2025) View paper
  - [19] Secure and Efficient Access Control for Computer-Use Agents via Context Space (Gong Haochen, 2025) View paper
  - Constitution-Based Agent Frameworks (2 papers)
  - [43] Personalized Constitutionally-Aligned Agentic Superego: Secure AI Behavior Aligned to Diverse Human Values (Eleanor 'Nell' Watson, 2025) View paper
  - [45] PSG-Agent: Personality-Aware Safety Guardrail for LLM-based Agents (Wu, 2025) View paper
  - Proactive and Predictive Enforcement (2 papers)
  - [39] Pro2Guard: Proactive Runtime Enforcement of LLM Agent Safety via Probabilistic Model Checking (Wang Haoyu, 2025) View paper
  - [49] Ensuring safety in llm-driven robotics: A cross-layer sequence supervision mechanism (Ziming Wang, 2024) View paper
- Safety Evaluation and Benchmarking
  - Comprehensive Multi-Domain Safety Benchmarks (3 papers)
  - [5] Openagentsafety: A comprehensive framework for evaluating real-world ai agent safety (Sanidhya Vijayvargiya, 2025) View paper

    - [9] Aegis2.0: A Diverse AI Safety Dataset and Risks Taxonomy for Alignment of LLM Guardrails (Ghosh, 2025) View paper
    - [21] Agent-safetybench: Evaluating the safety of llm agents (Zhang, 2024) View paper
    - Embodied and Task-Planning Safety Benchmarks (3 papers)
    - [14] Safeagentbench: A benchmark for safe task planning of embodied llm agents (Yin Sheng, 2024) View paper
    - [15] A Framework for Benchmarking and Aligning Task-Planning Safety in LLM-Based Embodied Agents (Huang Yu-ting, 2025) View paper
    - [46] AGENTSAFE: Benchmarking the Safety of Embodied Agents on Hazardous Instructions (Wang Le, 2025) View paper
    - GUI and Mobile Agent Safety Evaluation (3 papers)
    - [18] Mla-trust: Benchmarking trustworthiness of multimodal llm agents in gui environments (Yang Xiao, 2025) View paper
    - [26] Webguard: Building a generalizable guardrail for web agents (Zheng Boyuan, 2025) View paper
    - [33] Mobilesafetybench: Evaluating safety of autonomous agents in mobile device control (Lee, 2024) View paper
    - Human-Level and Adversarial Safety Evaluation (2 papers)
    - [30] Agentauditor: Human-level safety and security evaluation for llm agents (Luo Hanjun, 2025) View paper
    - [35] AgentGuard: Repurposing Agentic Orchestrator for Safety Evaluation of Tool Orchestration (Chen Ji-Zhou, 2025) View paper
- Threat Models and Attack Vectors
    - Jailbreaking and Prompt Injection Attacks (2 papers)
    - [8] Guardians of the agentic system: Preventing many shots jailbreak with agentic system (Rahman, 2025) View paper
    - [50] Adversarial Reinforcement Learning for Large Language Model Agent Safety (Wang, 2025) View paper
    - Control-Flow Hijacking and Orchestration Attacks (1 papers)
    - [29] Breaking and Fixing Defenses Against Control-Flow Hijacking in Multi-Agent Systems (Jha, 2025) View paper
    - Comprehensive Risk Taxonomies and Threat Landscapes (2 papers)
    - [3] TRiSM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems (Raza, 2025) View paper
    - [23] A Survey on Autonomy-Induced Security Risks in Large Model-Based Agents (Su Hang, 2025) View paper
- Domain-Specific Safety Applications
    - Robotics and Physical Embodiment Safety (2 papers)
    - [7] Safety Guardrails for LLM-Enabled Robots (Ravichandran, 2025) View paper
    - [13] Safety aware task planning via large language models in robotics (Khan, 2025) View paper
    - Scientific Discovery and High-Stakes Research (2 papers)
    - [6] SafeScientist: Toward Risk-Aware Scientific Discoveries by LLM Agents (Zhu Kunlun, 2025) View paper
    - [22] Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path? (Bengio, 2025) View paper
    - Critical Infrastructure and Industrial Control (3 papers)
    - [4] Osprey: Production-Ready Agentic AI for Safety-Critical Control Systems (Thorsten Hellert, 2025) View paper
    - [27] InfraMind: A Novel Exploration-based GUI Agentic Framework for Mission-critical Industrial Management (Zhu Zhao-Meng, 2025) View paper
    - [47] Grid-Agent: An LLM-Powered Multi-Agent System for Power Grid Control (Zhang Yan, 2025) View paper
    - Financial and Multi-Agent Team Systems (1 papers)
    - [24] From Tasks to Teams: A Risk-First Evaluation Framework for Multi-Agent LLM Systems in Finance (Z Chen, 2025) View paper
- Human-Agent Interaction and Oversight
    - Human-in-the-Loop Agentic Systems (1 papers)
    - [2] Magentic-UI: Towards Human-in-the-loop Agentic Systems (Mozannar, 2025) View paper
    - Proactive Inquiry and Assistance Mechanisms (1 papers)
    - [38] Inquiremobile: Teaching vlm-based mobile agent to request human assistance via reinforcement fine-tuning (Qihang Ai, 2025) View paper
- Adaptive and Learning-Based Safety Mechanisms
    - Reinforcement Learning for Safety Alignment (1 papers)
    - [17] Agent Safety Alignment via Reinforcement Learning (Sha, 2025) View paper
    - Adversarial Learning and Risk Pattern Discovery (1 papers)
    - [44] ALRPHFS: Adversarially Learned Risk Patterns with Hierarchical Fast\& Slow Reasoning for Robust Agent Defense (Xiang Shi-yu, 2025) View paper
    - Dynamic Thought Correction and Trajectory Alignment (1 papers)
    - [40] Think Twice Before You Act: Enhancing Agent Behavioral Safety with Thought Correction (Jiang Changyue, 2025) View paper
- Formal Verification and Logic-Based Safety
    - Logic-Based Action Verification (1 papers)
    - [34] VeriSafe Agent: Safeguarding Mobile GUI Agent via Logic-based Action Verification (Lee Jung-Jae, 2025) View paper
    - Knowledge-Enabled Reasoning Guardrails (2 papers)
    - [32] Guardagent: Safeguard llm agents by a guard agent via knowledge-enabled reasoning (Xiang Zhen, 2024) View paper
    - [48] Guardagent: Safeguard llm agents via knowledge-enabled reasoning (Z Xiang, 2025) View paper
- Multi-Agent Safety and Moderation (1 papers)
    - [28] Agentic Moderation: Multi-Agent Design for Safer Vision-Language Models (Ren Juan, 2025) View paper
- Foundational Concepts and Surveys
    - Agentic AI Surveys and Taxonomies (3 papers)
    - [11] A Survey on (M)LLM-Based GUI Agents (Tang Fei, 2025) View paper
    - [36] Vibe Coding vs. Agentic Coding: Fundamentals and Practical Implications of Agentic AI (Sapkota, 2025) View paper
    - [41] Agentic large-language-model systems in medicine: A systematic review and taxonomy (Abdul Mohaimen Al Radi, 2025) View paper
    - Guardrail Techniques and Fine-Tuning Reviews (1 papers)
    - [37] Advancing the Safety, Performance, and Adaptability of Large Language Models: Review of Fine-Tuning and Guardrails (Joshi, 2025) View paper
    - Real-Time Failure Detection and Monitoring (1 papers)
    - [42] Prioritizing real-time failure detection in AI agents (M Srikumar, 2025) View paper
- Auxiliary and Cross-Cutting Topics
    - Self-Improvement and Meta-Optimization (2 papers)

- [20] StorageXTuner: An LLM Agent-Driven Automatic Tuning Framework for Heterogeneous Storage Systems (Lin Qi, 2025) View paper
- [25] GÃ¶del agent: A self-referential agent framework for recursively self-improvement (Xunjian Yin, 2025) View paper
- Consistency and Logical Coherence Checking (1 papers)
- [31] Consistency Checks for Language Model Forecasters (Paleka, 2024) View paper

Narrative

Core task: pre-execution safety guardrails for LLM-based agentic systems. The field has organized itself around ten major branches that collectively address how to prevent harmful agent actions before they occur. Guardrail Architectures and Enforcement Mechanisms explores the structural designs and multi-stage frameworks that intercept risky behaviors, while Safety Evaluation and Benchmarking develops datasets and metrics to measure guardrail effectiveness across diverse scenarios. Threat Models and Attack Vectors catalogs the adversarial landscape—from prompt injection to control-flow hijacking—that guardrails must defend against. Domain-Specific Safety Applications tailors protections to high-stakes environments such as robotics, healthcare, and web automation, whereas Human-Agent Interaction and Oversight examines how human feedback and constitutional principles can guide safe operation. Adaptive and Learning-Based Safety Mechanisms investigates reinforcement learning and fine-tuning approaches that allow guardrails to evolve, and Formal Verification and Logic-Based Safety applies rigorous mathematical methods to certify agent behavior. Multi-Agent Safety and Moderation addresses coordination and content filtering in systems with multiple interacting agents, while Foundational Concepts and Surveys provide overarching taxonomies and theoretical grounding. Finally, Auxiliary and Cross-Cutting Topics captures emerging themes like personalized safety profiles and real-time failure detection.

Several active lines of work reveal key trade-offs between expressiveness and verifiability: adaptive mechanisms such as Safety Alignment RL[17] promise context-sensitive protection but complicate formal guarantees, whereas logic-based approaches offer provable correctness at the cost of reduced flexibility. Multi-stage frameworks have become particularly prominent, with works like Llamafirewall[1], TrustAgent Constitution[10], and TrustAgent[16] layering constitutional checks, plan verification, and execution monitoring to catch unsafe actions at multiple decision points. Foundational Guardrail[0] sits squarely within this multi-stage cluster, emphasizing a structured pipeline that integrates pre-execution filters with policy-driven oversight. Compared to TrustAgent Constitution[10], which foregrounds human-readable constitutional rules, Foundational Guardrail[0] appears to place greater weight on automated enforcement layers, while sharing TrustAgent[16]'s commitment to transparent, modular guardrail design. This positioning reflects a broader tension in the field: balancing the need for interpretable, human-aligned safety constraints against the demand for scalable, real-time protection in increasingly autonomous systems.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Llamafirewall: An open source guardrail system for building secure ai agents

Authors: Sa-hana Chennabasappa, Song Daniel, Cyrus Nikolaidis, MolnÃ¡r DÃ¡vid, Daniel Song, et al. (24 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Large language models (LLMs) have evolved from simple chatbots into autonomous agents capable of performing complex tasks such as editing production code, orchestrating workflows, and taking higher-stakes actions based on untrusted inputs like webpages and emails. These capabilities introduce new security risks that existing security measures, such as model fine-tuning or chatbot-focused guardrails, do not fully address. Given the higher stakes and the absence of deterministic solutions to mitig...

Relationship Analysis

Both papers belong to the Multi-Stage Guardrail Frameworks category, employing multiple validation stages for pre-execution safety in LLM-based agentic systems. They overlap in addressing pre-execution risk detection through layered architectures: the original paper proposes AuraGen for synthetic data generation and Safiron as a guardian model with cross-planner adaptation, while the candidate paper presents LlamaFirewall with PromptGuard 2 for jailbreak detection, AlignmentCheck for goal hijacking, and CodeShield for insecure code. The key difference is that the original paper focuses on foundational guardrail training via synthetic data and unified input normalization across planners, whereas the candidate paper emphasizes production-ready, modular scanners for specific threat types (prompt injection, alignment drift, code vulnerabilities) with real-time deployment at Meta.

---

2. Trustagent: Towards safe and trustworthy llm-based agents through agent constitution

Authors: Hua, Wenyue, Yang Xian-jun, Wenyue Hua, Jin Ming-yu, et al. (13 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

The rise of LLM-based agents shows great potential to revolutionize task planning, capturing significant attention. Given that these agents will be integrated into high-stake domains, ensuring their reliability and safety is crucial. This paper presents an Agent-Constitution-based agent framework, TrustAgent, with a particular focus on improving the LLM-based agent safety. The proposed framework ensures strict adherence to the Agent Constitution through three strategic components: pre-planning s...

Relationship Analysis

Both papers belong to the Multi-Stage Guardrail Frameworks category, employing multiple validation stages for pre-execution safety in LLM-based agentic systems. They overlap in addressing planning-stage safety through multi-stage architectures: the original paper uses a two-stage training pipeline (SFT + RL) with a cross-planner adapter and guardian model (Safiron) for pre-execution detection, while the candidate paper (TrustAgent) implements a three-stage operational framework (pre-planning, in-planning, post-planning) with constitution-based safety strategies. The key difference is that the original paper focuses on synthetic data generation (AuraGen) and a compact guardian model trained for detection/categorization/explanation, whereas TrustAgent emphasizes Agent Constitution enforcement through prompting, fine-tuning, and post-planning inspection without specialized synthetic data generation.

---

3. Trustagent: Towards safe and trustworthy llm-based agents

Authors: Hua Wenyue, Xianjun Yang, Wenyue Hua, Mingyu Jin, Zelong Li, et al. (8 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

The rise of LLM-based agents shows great potential to revolutionize task planning, capturing significant attention. Given that these agents will be integrated into high-stake domains, ensuring their reliability and safety is crucial. This paper presents an Agent-Constitution-based agent framework, TrustAgent, with a particular focus on improving the LLM-based agent safety. The proposed framework ensures strict adherence to the Agent Constitution through three strategic components: pre-planning s...

Relationship Analysis

Both papers belong to the Multi-Stage Guardrail Frameworks category, employing sequential validation stages for LLM-based agent safety. They overlap in their focus on pre-execution intervention through multi-stage architectures: the original paper (Safiron) uses a three-stage synthetic data pipeline (benign synthesis, risk injection, quality assurance) plus a two-stage training approach (SFT + RL),

while TrustAgent implements three temporal safety strategies (pre-planning knowledge injection, in-planning prompting with retrieval, and post-planning inspection with feedback). The key difference is that Safiron emphasizes foundational guardrail training via synthetic data generation and a unified adapter for cross-planner generalization, whereas TrustAgent focuses on Agent Constitution-based safety through dynamic regulation retrieval and iterative plan inspection without requiring model retraining for deployment.

Contributions Analysis

Overall novelty summary. The paper introduces three interconnected contributions addressing pre-execution safety for LLM agents: AuraGen (a synthetic data engine), Safiron (a foundational guardrail with cross-planner adapter), and Pre-Exec Bench (an evaluation benchmark). It resides in the Multi-Stage Guardrail Frameworks leaf, which contains four papers including Llamafirewall, TrustAgent Constitution, and TrustAgent. This leaf represents a moderately active research direction within the broader Guardrail Architectures and Enforcement Mechanisms branch, focusing on layered validation pipelines that intercept unsafe actions at multiple decision points before execution.

The taxonomy reveals that Multi-Stage Guardrail Frameworks sits alongside three sibling categories: Specification-Based Runtime Enforcement (two papers using formal languages), Constitution-Based Agent Frameworks (two papers embedding explicit safety principles), and Proactive and Predictive Enforcement (two papers employing probabilistic model checking). The paper's cross-planner adapter and multi-stage design connect it to constitution-based approaches, while its emphasis on pre-execution interception distinguishes it from runtime enforcement methods. Neighboring branches address complementary concerns: Safety Evaluation and Benchmarking (thirteen papers across four leaves) and Adaptive and Learning-Based Safety Mechanisms (three papers), suggesting the paper bridges architectural design with evaluation infrastructure.

Among twenty-two candidates examined, none clearly refute the three contributions. AuraGen's synthetic trajectory generation with controllable risk injection examined five candidates with zero refutations, suggesting novelty in combining benign synthesis, category-labeled risk insertion, and automated filtering. Safiron's cross-planner adapter and compact guardian model examined seven candidates with no overlapping prior work, indicating potential originality in unifying heterogeneous planner formats. Pre-Exec Bench examined ten candidates without refutation, though the comprehensive safety benchmark landscape (thirteen papers in the taxonomy) implies this contribution enters a more crowded evaluation space where incremental advances are common.

Based on the limited search scope of twenty-two semantically similar papers, the work appears to offer fresh perspectives on data generation and cross-planner unification, while the benchmark contribution aligns with established evaluation trends. The analysis does not cover exhaustive citation networks or domain-specific literature beyond top-K semantic matches, so definitive novelty claims require broader verification. The taxonomy context suggests the paper occupies a strategic position linking architectural innovation with evaluation infrastructure in a moderately mature research area.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: AuraGen: Synthetic Data Engine for Risky Agent Trajectories

Description: AuraGen is a three-stage synthetic data generation pipeline that addresses data scarcity by producing large-scale, diverse, and controllable corpora of risky agent trajectories. It synthesizes benign trajectories, injects risks through four principled strategies (single-step, multi-step, new branch, and bridged branch), and applies automated quality assurance via a reward model.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Predicting lane-changing risk considering the class imbalance problem: a control method for synthetic samples

URL: View paper

Brief Assessment

Lane Changing Risk[65] addresses class imbalance in traffic safety prediction through controlled synthetic sample generation for lane-changing scenarios. This is fundamentally different from AuraGen's focus on generating risky agent trajectories for LLM-based agentic systems with controllable risk injection strategies.

---

2. A survey on safety-critical driving scenario generationâ□□a methodological perspective

URL: View paper

Brief Assessment

Safety Critical Generation[62] focuses on generating safety-critical driving scenarios for autonomous vehicles, not general agentic systems. The methods target vehicle trajectories in traffic scenarios rather than multi-step agent action sequences with tool invocations.

---

3. TeraSim-World: Worldwide Safety-Critical Data Synthesis for End-to-End Autonomous Driving

URL: View paper

Brief Assessment

TeraSim World[64] focuses on autonomous driving sensor simulation and safety-critical traffic scenarios, not on general agentic system trajectories or LLM agent planning risks. The domains and risk taxonomies are fundamentally different.

---

4. Decoupled diffusion sparks adaptive scene generation

URL: View paper

Brief Assessment

Decoupled Diffusion[63] focuses on scene generation for autonomous driving using diffusion models with decoupled noise states, not on synthetic data generation for risky agent trajectories in general agentic systems. The technical domains and objectives are fundamentally different.

---

5. Automating Safety Enhancement for LLM-based Agents with Synthetic Risk Scenarios

URL: View paper

Brief Assessment

Synthetic Risk Scenarios[66] focuses on generating risk scenarios for LLM-based agents through automated simulation of unsafe user behaviors and self-reflective reasoning, whereas AuraGen specifically addresses synthetic data generation for risky agent trajectories with a three-stage pipeline involving benign trajectory synthesis, principled risk injection via four distinct strategies, and automated quality assurance through a reward model. The candidate's approach differs in its emphasis on user-environment interaction simulation rather than trajectory-level risk injection mechanisms.

Contribution 2: Safiron: Foundational Guardrail with Cross-Planner Adapter

Description: Safiron is a guardian model that combines a unified adapter (normalizing heterogeneous agent outputs) with a compact detection model. It flags risky cases, assigns fine-grained risk types, and generates explanations, trained via a two-stage recipe (supervised fine-tuning followed by GRPO-based reinforcement learning).

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. BayesLoRA: Task-Specific Uncertainty in Low-Rank Adapters

URL: View paper

Brief Assessment

BayesLoRA[53] focuses on uncertainty quantification and rank selection in low-rank adapters for model fine-tuning, not on guardrail systems for detecting risks in agent planning outputs. The technical domains are entirely distinct.

---

2. Adapting to Planning Failures in Lifelong Multi-Agent Path Finding

URL: View paper

Brief Assessment

Planning Failures MAPF[51] addresses planning failures in multi-agent pathfinding for warehouse automation, not guardrail models with adapters for detecting risks in agent planning outputs. The domains are fundamentally different.

---

3. Agentops pattern catalogue: Architectural patterns for safe and observable operations of foundation model-based agents

URL: View paper

Brief Assessment

AgentOps Patterns[55] discusses architectural patterns for safe operations and mentions adapters for heterogeneous stores/APIs and per-step guardrails, but focuses on operational patterns rather than a unified detection model with risk categorization and explanation generation trained via supervised fine-tuning and reinforcement learning.

---

4. Protect: Towards Robust Guardrailing Stack for Trustworthy Enterprise LLM Systems

URL: View paper

Brief Assessment

Protect[52] focuses on multi-modal content moderation (toxicity, sexism, privacy, prompt injection) across text/image/audio inputs, not on pre-execution planning-stage guardrails for agentic systems. The adapter in Protect[52] normalizes different output formats for safety classification, whereas Safiron's cross-planner adapter unifies heterogeneous agent planning outputs before risk detection at the planning stage.

---

5. AGENTSAFE: Benchmarking the Safety of Embodied Agents on Hazardous Instructions

URL: View paper

Brief Assessment

AGENTSAFE[46] focuses on evaluating embodied agents in simulation environments with hazardous instructions, not on building guardrail models with adapters for detecting risks in agent planning outputs.

---

6. Deploying Agentic AI in Enterprise Environments

URL: View paper

Brief Assessment

The candidate paper (Deploying Agentic Enterprise[54]) discusses enterprise deployment of agentic AI with adapters and guardrails, but the provided context fragments are too sparse to establish whether it presents similar technical contributions predating the original paper's unified adapter architecture and two-stage training recipe for Safiron.

---

7. PSG-Agent: Personality-Aware Safety Guardrail for LLM-based Agents

URL: View paper

Brief Assessment

PSG Agent[45] focuses on personality-aware, user-specific safety guardrails for agents, not on cross-planner adapters or unified input normalization for heterogeneous agent outputs.

Contribution 3: Pre-Exec Bench: Benchmark for Pre-Execution Safety Evaluation

Description: Pre-Exec Bench is a benchmark designed specifically for evaluating planning-stage (pre-execution) safety in agentic systems. It is constructed through tool refinement, diverse trajectory generation, and two-phase human verification, providing realistic assessments of detection, categorization, explanation, and generalization capabilities.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Guard: A safe reinforcement learning benchmark

URL: View paper

Brief Assessment

Guard Benchmark[61] focuses on safe reinforcement learning for robot locomotion tasks with physical safety constraints (hazards, collisions), not on pre-execution safety evaluation of LLM-based agentic systems' planning trajectories.

---

2. Why do multiagent systems fail?

URL: View paper

Brief Assessment

Multiagent Failures[57] focuses on failure taxonomy for multi-agent LLM systems across general task completion, not pre-execution safety evaluation in agentic systems with diverse tools.

---

3. Safeagentbench: A benchmark for safe task planning of embodied llm agents

URL: View paper

Brief Assessment

SafeAgentBench[14] focuses on embodied agents executing tasks in interactive simulation environments (AI2-THOR), evaluating safety during task execution in physical environments. Pre-Exec Bench evaluates planning-stage safety before any action is executed in general agentic systems, operating at a different intervention point and scope.

---

4. Evaluation and benchmarking of llm agents: A survey

URL: View paper

Brief Assessment

LLM Agents Evaluation[56] is a survey paper that provides a taxonomy of evaluation objectives and processes for LLM agents. It does not present a specific benchmark for pre-execution safety evaluation with diverse tools and trajectory generation, which is the core contribution of Pre-Exec Bench.

---

5. Safety gymnasium: A unified safe reinforcement learning benchmark

URL: View paper

Brief Assessment

Safety Gymnasium[58] focuses on safe reinforcement learning environments for robot control tasks with safety constraints during execution (e.g., velocity constraints, hazards, pillars). It does not address pre-execution planning-stage safety evaluation in agentic systems with diverse tools, which is the core focus of Pre-Exec Bench.

---

6. Personalized Constitutionally-Aligned Agentic Superego: Secure AI Behavior Aligned to Diverse Human Values

URL: View paper

Brief Assessment

Personalized Superego[43] focuses on constitutional alignment and real-time compliance enforcement for agentic systems, not on creating benchmarks for pre-execution safety evaluation with diverse tools and trajectory generation methodologies.

---

7. Safebench: A benchmarking platform for safety evaluation of autonomous vehicles

URL: View paper

Brief Assessment

SafeBench[59] focuses on autonomous vehicle safety evaluation through driving scenario generation and testing, not on pre-execution safety evaluation of general agentic systems with diverse tools.

---

8. Agentauditor: Human-level safety and security evaluation for llm agents

URL: View paper

Brief Assessment

AgentAuditor[30] focuses on evaluating LLM-based evaluators for agent safety and security through a memory-augmented reasoning framework and introduces AsseBench for this purpose. It does not address pre-execution (planning-stage) safety evaluation of agentic systems with diverse tools, which is the core focus of Pre-Exec Bench.

---

9. Redcode: Risky code execution and generation benchmark for code agents

URL: View paper

Brief Assessment

RedCode[60] focuses on evaluating code agents' safety in executing and generating risky code snippets, not on pre-execution planning-stage safety evaluation of general agentic systems with diverse tools as described in the original paper.

---

10. Openagentsafety: A comprehensive framework for evaluating real-world ai agent safety

URL: View paper

Brief Assessment

OpenAgentSafety[5] focuses on evaluating agents during execution with real tools (file systems, browsers, messaging platforms) across multi-turn interactions, not on pre-execution planning-stage safety evaluation with diverse trajectory generation methods.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Building a Foundational Guardrail for General Agentic Systems via Synthetic Data View paper
- [1] Llamafirewall: An open source guardrail system for building secure ai agents View paper
- [2] Magentic-UI: Towards Human-in-the-loop Agentic Systems View paper
- [3] TRiSM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems View paper
- [4] Osprey: Production-Ready Agentic AI for Safety-Critical Control Systems View paper
- [5] Openagentsafety: A comprehensive framework for evaluating real-world ai agent safety View paper
- [6] SafeScientist: Toward Risk-Aware Scientific Discoveries by LLM Agents View paper
- [7] Safety Guardrails for LLM-Enabled Robots View paper
- [8] Guardians of the agentic system: Preventing many shots jailbreak with agentic system View paper
- [9] Aegis2.0: A Diverse AI Safety Dataset and Risks Taxonomy for Alignment of LLM Guardrails View paper
- [10] Trustagent: Towards safe and trustworthy llm-based agents through agent constitution View paper
- [11] A Survey on (M)LLM-Based GUI Agents View paper
- [12] AgentSpec: Customizable Runtime Enforcement for Safe and Reliable LLM Agents View paper
- [13] Safety aware task planning via large language models in robotics View paper
- [14] Safeagentbench: A benchmark for safe task planning of embodied llm agents View paper
- [15] A Framework for Benchmarking and Aligning Task-Planning Safety in LLM-Based Embodied Agents View paper
- [16] Trustagent: Towards safe and trustworthy llm-based agents View paper
- [17] Agent Safety Alignment via Reinforcement Learning View paper
- [18] Mla-trust: Benchmarking trustworthiness of multimodal llm agents in gui environments View paper
- [19] Secure and Efficient Access Control for Computer-Use Agents via Context Space View paper
- [20] StorageXTuner: An LLM Agent-Driven Automatic Tuning Framework for Heterogeneous Storage Systems View paper

- [21] Agent-safetybench: Evaluating the safety of llm agents View paper
- [22] Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path? View paper
- [23] A Survey on Autonomy-Induced Security Risks in Large Model-Based Agents View paper
- [24] From Tasks to Teams: A Risk-First Evaluation Framework for Multi-Agent LLM Systems in Finance View paper
- [25] GÃ¶del agent: A self-referential agent framework for recursively self-improvement View paper
- [26] Webguard: Building a generalizable guardrail for web agents View paper
- [27] InfraMind: A Novel Exploration-based GUI Agentic Framework for Mission-critical Industrial Management View paper
- [28] Agentic Moderation: Multi-Agent Design for Safer Vision-Language Models View paper
- [29] Breaking and Fixing Defenses Against Control-Flow Hijacking in Multi-Agent Systems View paper
- [30] Agentauditor: Human-level safety and security evaluation for llm agents View paper
- [31] Consistency Checks for Language Model Forecasters View paper
- [32] Guardagent: Safeguard llm agents by a guard agent via knowledge-enabled reasoning View paper
- [33] Mobilesafetybench: Evaluating safety of autonomous agents in mobile device control View paper
- [34] VeriSafe Agent: Safeguarding Mobile GUI Agent via Logic-based Action Verification View paper
- [35] AgentGuard: Repurposing Agentic Orchestrator for Safety Evaluation of Tool Orchestration View paper
- [36] Vibe Coding vs. Agentic Coding: Fundamentals and Practical Implications of Agentic AI View paper
- [37] Advancing the Safety, Performance, and Adaptability of Large Language Models: Review of Fine-Tuning and Guardrails View paper
- [38] Inquiremobile: Teaching vlm-based mobile agent to request human assistance via reinforcement fine-tuning View paper
- [39] Pro2Guard: Proactive Runtime Enforcement of LLM Agent Safety via Probabilistic Model Checking View paper
- [40] Think Twice Before You Act: Enhancing Agent Behavioral Safety with Thought Correction View paper
- [41] Agentic large-language-model systems in medicine: A systematic review and taxonomy View paper
- [42] Prioritizing real-time failure detection in AI agents View paper
- [43] Personalized Constitutionally-Aligned Agentic Superego: Secure AI Behavior Aligned to Diverse Human Values View paper
- [44] ALRPHFS: Adversarially Learned Risk Patterns with Hierarchical Fast\& Slow Reasoning for Robust Agent Defense View paper
- [45] PSG-Agent: Personality-Aware Safety Guardrail for LLM-based Agents View paper
- [46] AGENTSAFE: Benchmarking the Safety of Embodied Agents on Hazardous Instructions View paper
- [47] Grid-Agent: An LLM-Powered Multi-Agent System for Power Grid Control View paper
- [48] Guardagent: Safeguard llm agents via knowledge-enabled reasoning View paper
- [49] Ensuring safety in llm-driven robotics: A cross-layer sequence supervision mechanism View paper
- [50] Adversarial Reinforcement Learning for Large Language Model Agent Safety View paper
- [51] Adapting to Planning Failures in Lifelong Multi-Agent Path Finding View paper
- [52] Protect: Towards Robust Guardrailing Stack for Trustworthy Enterprise LLM Systems View paper
- [53] BayesLoRA: Task-Specific Uncertainty in Low-Rank Adapters View paper
- [54] Deploying Agentic AI in Enterprise Environments View paper
- [55] Agentops pattern catalogue: Architectural patterns for safe and observable operations of foundation model-based agents View paper
- [56] Evaluation and benchmarking of llm agents: A survey View paper
- [57] Why do multiagent systems fail? View paper
- [58] Safety gymnasium: A unified safe reinforcement learning benchmark View paper
- [59] Safebench: A benchmarking platform for safety evaluation of autonomous vehicles View paper
- [60] Redcode: Risky code execution and generation benchmark for code agents View paper
- [61] Guard: A safe reinforcement learning benchmark View paper
- [62] A survey on safety-critical driving scenario generationâ□□a methodological perspective View paper
- [63] Decoupled diffusion sparks adaptive scene generation View paper
- [64] TeraSim-World: Worldwide Safety-Critical Data Synthesis for End-to-End Autonomous Driving View paper
- [65] Predicting lane-changing risk considering the class imbalance problem: a control method for synthetic samples View paper
- [66] Automating Safety Enhancement for LLM-based Agents with Synthetic Risk Scenarios View paper