Guardrails in Large Language Models for Domain-Specific AI Assistants: A Systematic Literature Review

Daniel S. A. Araújo[1], **Lhayana V. Macedo**[1]

[1]Digital Metropolis Institute (IMD) – Federal University of Rio Grande do Norte (UFRN)
Natal/RN – Brazil – https://www.metropoledigital.ufrn.br/portal/

daniel@imd.ufrn.br, lhayana.vieira.073@ufrn.edu.br

Abstract. *Guardrails are runtime input/output filters that enforce domain restrictions and safety policies at inference time. Since 2022, they have become critical infrastructure for Large Language Model (LLM) deployments, while recent surveys map LLM safeguards broadly [Dong et al. 2025], a synthesis focused on domain-restricted assistants and on the operational security–performance trade-off (latency, cost, over-defense) is still missing. Following Kitchenham's protocol across five databases, this SLR examines guardrails along five dimensions—off-topic detection, harmful-content filtering, PII protection, adversarial mitigation, and security-performance trade-offs—retaining 22 of 756 studies (2022–2026) at a quality threshold $\geq$4.5/5.0. Findings show that off-topic detection has matured from keyword filters to Representation Offset Analysis (ROA) and Neural Barrier Functions (NBF); adversarial mitigation is most effective when defensive suffixes, lightweight classifiers, and semantic deviation analysis are combined; and lightweight classifiers add only tens of milliseconds of latency. Adaptive routing can, in principle, keep most interactions on compact models, since its theoretical upper bound delegates only $\approx 5\%$ of inputs to a larger safety model, but trained routers only approximate this ideal. A critical emerging concern is that guardrails themselves are vulnerable to backdoor poisoning and denial-of-service exploitation. Cross-modal evaluation and multi-agent security frameworks represent the field's most urgent open problems.*

1. Introduction

Since 2022, the release of ChatGPT has been a turning point in the accessibility of instructional Large Language Models (LLMs), bringing conversational AI to the general public at an unprecedented scale [Nazir and Wang 2023]. Built on architectures such as GPT-3.5 and refined through Reinforcement Learning from Human Feedback (RLHF), these models have quickly become available across practical applications ranging from corporate assistants to customer service systems.

However, the same natural interface that broadened access also opened a profound new attack surface. The literature soon catalogued failures ranging from hallucinations to critical vulnerabilities such as prompt injection and jailbreaking [Borji 2023], which exploit the model's instruction-following to bypass its own safety guidelines. In practice these translate into multidimensional risks: training-data bias, topic drift, and PII leakage. Whereas pre-2022 work focused on bias reduction in research models

[Weidinger et al. 2021], attention has since shifted to production guardrails that filter inputs and outputs in real time [Dong et al. 2025].

To face these threats, guardrails (runtime moderation filters that block domain violations or harmful content) have emerged as a key measure. While broad surveys of LLM safeguards already exist [Dong et al. 2025], they synthesize the field at a general level; what remains missing is a focused characterization of guardrails in domain-restricted assistants and an actionable account of the security–performance trade-off (latency, computational cost, and over-defense), which this review provides.

To address this gap, this Systematic Literature Review (SLR) was designed to achieve a fivefold objective: (i) identify the most effective techniques for off-topic detection; (ii) map the predominant architectural approaches for harmful content filtering; (iii) analyze PII protection strategies; (iv) evaluate mechanisms for adversarial attack mitigation; and (v) characterize the trade-off between security and operational performance in guardrail-equipped systems.

To fulfill the stated objectives, this paper performs a SLR following the methodological guidelines outlined in [Kitchenham and Charters 2007], with searches across five scientific databases (ACM Digital Library, IEEE Xplore, Scopus, ScienceDirect, and SpringerLink) and a structured three-phase protocol for study selection and quality assessment.

The remainder of this paper is organized as follows: Section 2 describes the methodology; Section 3 presents the study selection results; Section 4 discusses the findings in response to the research questions; and Section 5 presents conclusions and directions for future work.

2. Methodology

The review followed four phases: defining research questions, formulating the search strategy, selecting and quality-assessing studies, and extracting and synthesizing data.

2.1. Research Objective and Research Questions

This review investigates the implementation of guardrails in LLM-based AI assistants—the constraints imposed on functional domain and the measures adopted to prevent harmful output and protect sensitive data—motivating the research questions (RQs) below.

- **RQ1:** Which techniques (such as intent classifiers and prompt filters) does the literature identify as most effective for detecting and blocking off-topic deviations in domain-specific chatbots?
- **RQ2:** What are the predominant architectural approaches for filtering and preventing the generation of hate speech, toxicity, and inappropriate content by AI systems?
- **RQ3:** How are guardrails being used to detect, mask, or block Personally Identifiable Information (PII) and ensure the protection of user data?
- **RQ4:** Which guardrail techniques and architectures are cited in the literature as the most effective in detecting and mitigating intentional attacks, such as prompt injection, jailbreaking, and adversarial attacks?

- **RQ5:** What is the impact of applying guardrails on the trade-off between security and performance, specifically in terms of latency, computational cost, and degradation in response quality?

RQ1–RQ3 address content governance: mitigating accidental violations, model biases, and domain drift through semantic filtering and moderation. RQ4 targets adversarial security, where malicious actors deliberately subvert alignment; because such attacks bypass surface-level filters, they warrant a dedicated question. RQ5 evaluates operational viability, since runtime defenses introduce latency, cost, and over-defense that must remain compatible with real-world deployment.

2.2. Search String and Temporal Delimitation

Searches ran across the five databases, combining LLM and conversational-agent terms with security and domain-restriction terms, per the main string below:

```
 ("Large Language Model*" OR "LLM*" OR "Conversational AI"
OR "Chatbot*" OR "AI Assistant*" OR "Generative AI") AND
("Guardrail*" OR "Safety Filter*" OR "Input Moderation" OR
"Output Moderation" OR "Content Filter*") AND ("Harmful
Content" OR "Toxicity" OR "Hate Speech" OR "Privacy" OR "Data
Protection" OR "PII" OR "Data Leak*" OR "Domain Specific"
OR "Topic Restriction" OR "Off-topic" OR "Topic Drift" OR
"Prompt Injection" OR "Jailbreak*" OR "Adversarial Attack*" OR
"Trade-off" OR "Latency" OR "Computational Cost")
```

ScienceDirect, which limits query length, used a reduced string:

```
("LLM" OR "Large Language Model" OR "Generative AI") AND ("Guardrail"
OR "Safety Filter" OR "Prompt Injection" OR "Jailbreak")
```

The search covers publications from 2022 onward, when the democratization of instructional LLMs shifted research focus from bias reduction to production guardrails with real-time filtering [Dong et al. 2025].

2.3. Inclusion and Exclusion Criteria

2.3.1. Inclusion Criteria

Studies were included that met all of the following criteria:

- Written entirely in English;
- Explicitly focused on the proposal, evaluation, or analysis of guardrail mechanisms (input/output filters) applied to LLMs or AI assistants;
- Address the application of guardrails for domain restriction (e.g., preventing off-topic responses) or for safety (e.g., toxicity, data privacy);
- Published in peer-reviewed scientific journals or peer-reviewed conference proceedings.

2.3.2. Exclusion Criteria

Studies that met any of the following criteria were excluded:

- Technical reports, non-scientific articles, short papers, and posters;
- Duplicate results across the databases consulted;
- Works without the full text available for download, even via institutional access;
- Works that do not address any of the defined research questions;
- Publications prior to 2022;
- Secondary studies (systematic reviews, surveys, and meta-analyses);
- Studies focused on multimodal inputs or outputs;
- Purely adversarial research or attack frameworks;
- Exclusively theoretical or conceptual studies, without the presentation of quantitative performance metrics.

This criterion targets works whose primary contribution is a novel attack; it does not exclude studies that evaluate the robustness of existing guardrails (e.g., [Tyshchenko and Drushchak 2025, Zhang et al. 2025a, Tong et al. 2025]), which we retain because they characterize defensive failure modes rather than proposing attacks as an end in themselves.

2.4. Quality Assessment

The preselected studies underwent a quality assessment based on five criteria:

- **QC1 (Quantitative metrics):** Does the study report quantitative safety performance metrics?
- **QC2 (Defensive Architecture):** Are the structural components and operational flows of the defensive mechanism explicitly defined?
- **QC3 (Benchmarks or Public Datasets):** Do the experiments use standardized benchmarks or publicly available datasets?
- **QC4 (Discussion of limitations):** Do the authors explicitly discuss the limitations of the proposed approach?
- **QC5 (Operational efficiency):** Does the study address operational efficiency, such as latency, throughput, or computational cost?

"Safety performance" (QC1) encompasses the trade-off between security and utility: a defense that indiscriminately blocks interactions suffers severe over-defense, impractical for deployment. The corpus itself illustrates this—[Zhang et al. 2025a] weaponizes a guardrail's false positives for denial-of-service, and PIGuard [Li et al. 2025] is explicitly motivated by undue blocking of benign prompts. QC1 therefore requires both threat-mitigation effectiveness (e.g., Attack Success Rate) and precision on normal interactions (e.g., False Positive Rate), with hardware and compute cost isolated in QC5. positives to trigger denial-of-service, while PIGuard [Li et al. 2025] is explicitly motivated by the undue blocking of benign prompts (over-defense). Therefore, QC1 ensures that the evaluated studies report both their effectiveness in mitigating threats (e.g., Attack Success Rate) and their precision in preserving normal interactions (e.g., False Positive Rate). By isolating hardware and computational costs in QC5, QC1 strictly evaluates the filtering quality and the balance between protection and usability.

Each criterion was scored 0, 0.5, or 1 (maximum 5.0). Only studies scoring $\geq$4.5 were analyzed in full, ensuring each met all five criteria at least partially.

Applying the 4.5 threshold retained 22 of the 52 eligible studies; the 30 excluded scored between 1.5 and 4.0. The dominant reason for exclusion was QC5 (operational efficiency): 24 of the 30 excluded studies (80%) failed it fully or partially, and QC5 alone accounts for 42% of the quality points lost across the excluded set — more than any other criterion. In other words, most excluded works did propose a defensive architecture (mean QC2 of 0.82) but did not report latency, throughput, or computational cost. As a sensitivity check, relaxing the threshold to 4.0 would admit 13 additional studies (35 in total); however, 9 of these 13 also fail QC5, so they would broaden architectural coverage without strengthening the operational characterization on which RQ5 depends. The answers to RQ1–RQ5 would thus remain substantively unchanged, and the systematic under-reporting of QC5 itself reinforces the motivation for this review.

2.5. Data Extraction and Synthesis

Studies passing the quality assessment underwent a standardized extraction protocol (technique, architecture, safety metrics, operational impact), synthesized qualitatively to identify recurring patterns and gaps.

2.6. Execution

The review followed three sequential phases. In Phase 1, the 756 studies retrieved from the databases were screened by title and abstract, resulting in 69 candidates.[1] In Phase 2, these studies were reviewed for their introduction and conclusion to verify eligibility, reducing the set to 52 studies. In Phase 3, the eligible studies were evaluated according to quality criteria QC1–QC5 (Section 2.4), comprising the final corpus of 22 articles. Table 1 summarizes the inclusion and exclusion numbers for each phase.

Table 1. Works included and excluded at each phase

	Input	Removed	Remaining
Stage 1: Initial Screening	756	687	69
Stage 2: Eligibility	69	17	52
Stage 3: Quality Assessment	52	30	22

3. Results

This section presents the final corpus, its temporal distribution, and the answers to the research questions derived from the 22 retained studies.

3.1. Overview of the papers

Table 2 lists the 22 articles with their publication venue[2], sorted by year. All fall between 2024 and 2026, reflecting the review window and the novelty of guardrails for instructional LLMs.

The distribution concentrates in 2025 (15 of 22, 68%), with 5 works from 2026 already published at search time and 2 from 2024—the most established in the corpus.

[1]The 687 records excluded in Stage 1 include both off-topic results and cross-database duplicates.

[2]Tiers follow standard bibliometric rankings, with Q1–Q4 denoting Scimago/JCR journal quartiles and A*/A/B the CORE conference ranking; workshops and preprints are labeled as such.

Table 2. Overview of selected papers sorted by year

Id	Citation	Year	Venue (type, tier)	Title
S1	[Rabieinejad et al. 2026]	2026	IPM (Journal, Q1)	Beyond the prompt: Log-based threat detection and attribution for multi-Agent LLMs
S2	[Liu et al. 2026]	2026	IPM (Journal, Q1)	Defending LLMs against jailbreak attacks through representation offset detection
S3	[Cevallos-Moreno et al. 2026]	2026	EAAI (Journal, Q1)	Ethical treatment of language models against harmful inference-time interventions
S4	[Hu et al. 2026]	2026	TMLR 2026 (Journal)	Steering Dialogue Dynamics for Robustness against Multi-turn Jailbreaking Attacks
S5	[Yan and Wan 2026]	2026	AIFM '25 (Conf., ACM)	LMDMI: A Lightweight Multilevel Defense against Malicious Inputs for Generative Language Models
S6	[Yi et al. 2025]	2025	Neurocomputing (Journal, Q1)	Fine-grained detoxification framework via instance-level prefixes for large language models
S7	[Zhang et al. 2025b]	2025	TOSEM (Journal, A*)	JailGuard: A Universal Detection Framework for Prompt-based Attacks on LLM Systems
S8	[Lee et al. 2025a]	2025	PVLDB (Journal, A*)	Semantic Integrity Constraints: Declarative Guardrails for AI-Augmented Data Processing Systems
S9	[Yang et al. 2025]	2025	ASE 2025 (Conf., A*)	AdaptiveGuard: Towards Adaptive Runtime Safety for LLM-Powered Software
S10	[Lee et al. 2025b]	2025	ACL 2025 Findings (Conf.)	SafeRoute: Adaptive Model Selection for Efficient and Accurate Safety Guardrails in Large Language Models
S11	[Xiong et al. 2025]	2025	ACL 2025 Findings (Conf.)	Defensive Prompt Patch: A Robust and Generalizable Defense of Large Language Models against Jailbreak Attacks
S12	[Tyshchenko and Drushchak 2025]	2025	FLLM 2025 (Conf., IEEE)	Towards Safe and Secure LLM Systems: Defense Strategies and Assessment Against Inference-Time Security Risks
S13	[Krishnan and Siddiquie 2025]	2025	IEEE ICA 2025 (Conf.)	Policy Over Tokens: Enforcing Declarative Governance Constraints in Cost-Aware LLM Deployments
S14	[Zheng et al. 2025]	2025	COLING 2025 (Conf., A)	Lightweight Safety Guardrails Using Fine-tuned BERT Embeddings
S15	[Li et al. 2025]	2025	ACL 2025 Main (Conf., A*)	PIGuard: Prompt Injection Guardrail via Mitigating Overdefense for Free
S16	[Zhang et al. 2025a]	2025	LAMPS '25 (Workshop, ACM)	LLM Safeguard is a Double-Edged Sword: Exploiting False Positives for Denial-of-Service Attacks
S17	[Sreedhar et al. 2025]	2025	EMNLP 2025 Findings (Conf.)	Safety Through Reasoning: An Empirical Study of Reasoning Guardrail Models
S18	[Tong et al. 2025]	2025	ICLR 2025 (Conf., A*)	BadJudge: Backdoor Vulnerabilities of LLM-As-A-Judge
S19	[Chen et al. 2025]	2025	ICML 2025 (Conf., A*)	ShieldAgent: Shielding Agents via Verifiable Safety Policy Reasoning
S20	[Piet et al. 2025]	2025	AISec '25 (Workshop)	JailbreaksOverTime: Detecting Jailbreak Attacks Under Distribution Shift
S21	[Hu et al. 2024]	2024	NeurIPS 2024 (Conf., A*)	Gradient Cuff: Detecting Jailbreak Attacks on Large Language Models by Exploring Refusal Loss Landscapes
S22	[Yuan et al. 2024]	2024	ICML 2024 (Conf., A*)	RigorLLM: Resilient Guardrails for Large Language Models against Undesired Content

4. Discussion

This section answers the following research questions:

RQ1 — Off-topic detection.

Off-topic detection has evolved from fragile keyword filters to approaches com-

bining semantic understanding with architectural control [Krishnan and Siddiquie 2025, Yan and Wan 2026]. Reasoning-based guardrails generate an intermediate chain of reasoning to judge topical alignment before responding [Sreedhar et al. 2025], while Representation Offset Analysis (ROA) measures the geometric distance between input and intended intent in embedding space to block camouflaged deviations [Liu et al. 2026]. For extended dialogues, Semantic Integrity Constraints combined with Neural Barrier Function steering constrain the dialogue trajectory to the approved topic space, catching the gradual multi-turn drift that single-turn classifiers miss [Lee et al. 2025a, Hu et al. 2026].

RQ2 — Harmful-content filtering.

The literature groups solutions into three families: native alignment, external filters, and decoding interventions. Alignment via RLHF and SFT integrates ethical guidelines into model weights [Xiong et al. 2025, Zhang et al. 2025a]. External runtime mechanisms include input defenses such as the Defensive Prompt Patch (DPP) and lightweight TinyBERT/DeBERTa classifiers for toxic intent [Xiong et al. 2025, Li et al. 2025, Yan and Wan 2026], plus representation-analysis frameworks—JailGuard uses KL divergence to detect instabilities, while Representation Offset Defense (ROD) monitors latent-space distance [Zhang et al. 2025b, Liu et al. 2026]. At generation time, FGDILP subtracts subtoxicity vectors during decoding [Yi et al. 2025], Gradient Cuff inspects the refusal-loss landscape, and AdaptiveGuard and JailbreaksOverTime adapt continually to new threats [Hu et al. 2024, Yang et al. 2025, Piet et al. 2025].

RQ3 — PII protection.

PII protection combines semantic classifiers, active masking, and behavioral auditing. Preventive blocking uses Sentence-BERT with SVM or neural classifiers to intercept sensitive data before the model [Zheng et al. 2025]. Policy Over Tokens anonymizes via declarative pre-tokenization and post-generation layers [Krishnan and Siddiquie 2025], while Semantic Integrity Constraints enforce PII-exclusion rules through abstraction to preserve utility [Lee et al. 2025a]. For agents, SHIELDAGENT compiles regulations such as GDPR into runtime monitoring circuits [Chen et al. 2025], and commercial APIs (e.g., Amazon Comprehend) mitigate exfiltration [Tyshchenko and Drushchak 2025]. ALTEDA flags anomalies from telemetry without reading message content [Rabieinejad et al. 2026]. Protection is further reinforced indirectly by robustness against jailbreaks and prompt injection [Liu et al. 2026, Zhang et al. 2025b, Li et al. 2025, Cevallos-Moreno et al. 2026].

RQ4 — Adversarial-attack mitigation.

The literature indicates that the mitigation of malicious attacks on Large Language Models has evolved toward approaches that combine defensive suffixes, lightweight classifiers, and semantic deviation analysis. Strategies such as Defensive Prompt Patch (DPP) and RigorLLM use genetic algorithms and minimax optimization to generate mathematical suffixes that, when appended to user prompts, force the rejection of malicious intentions without compromising the system's utility [Xiong et al. 2025, Yuan et al. 2024]. In parallel, lightweight classifier frameworks such as PIGuard and LMDMI, based on models like DeBERTaV3 and TinyBERT, have proven effective at filtering prompt injection and jailbreaking across multiple detection levels with low latency [Li et al. 2025, Yan and Wan 2026]. JailGuard exploits attack instability detection through input muta-

tions and divergence analysis, while Gradient Cuff and Representation Offset Defense (ROD) focus on measuring rejection loss and vector space mapping to identify hidden intentions [Zhang et al. 2025b, Hu et al. 2024, Liu et al. 2026].

For dynamic threats and prolonged dialogues, AdaptiveGuard uses anomaly detection for continual learning, while Neural Barrier Function steering applies control theory to keep the dialogue out of vulnerable zones [Yang et al. 2025, Hu et al. 2026], and reasoning models neutralize persuasive tactics via Chain-of-Thought [Sreedhar et al. 2025]. The guardrail models themselves (LLM-as-a-Judge) are also vulnerable to backdoor poisoning, which Model Merging mitigates by restoring evaluator reliability at low cost [Tong et al. 2025].

Finally, for multi-agent systems, SHIELDAGENT enforces verifiable safety-policy circuits and ALTEDA detects agent hijacking from behavioral logs without reading message content [Chen et al. 2025, Rabieinejad et al. 2026].

RQ5 — Security–performance trade-off.

Guardrails inevitably trade safety against performance. Over-defensive designs raise false-positive rates that frustrate users [Li et al. 2025, Zhang et al. 2025a], while heavy architectures add latency [Zheng et al. 2025, Sreedhar et al. 2025], degrade native reasoning [Cevallos-Moreno et al. 2026], and inflate token and memory cost [Zhang et al. 2025b]. Recent work, however, rebalances this equation [Sreedhar et al. 2025, Lee et al. 2025b].

To mitigate latency, the literature suggests the use of lightweight semantic classifiers: a distilBERT-based classifier ($\approx$67M params) adds $\approx$50 ms [Zheng et al. 2025], the TinyBERT-based LMDMI ($\approx$11M params) runs at $\approx$10 ms even on CPU [Yan and Wan 2026], and PIGuard (DeBERTaV3) at $\approx$15 ms [Li et al. 2025]; further gains come from cascaded hierarchical filters and the enforcement of "Reasoning Budgets", intercepting dangerous commands in mere milliseconds [Sreedhar et al. 2025]. The high computational cost has been addressed through intelligent routing. Its theoretical upper bound (the SafeRoute *oracle*, which assumes access to ground-truth labels) delegates only $\approx 5\%$ of inputs to the larger safety model while the compact model handles the remaining $\approx 95\%$; the trained router only approximates this selection, reaching an average routing F1 of ≈ 0.50 rather than attaining the bound directly [Lee et al. 2025b]. This cost is further reduced by agile predictive algorithms such as KNN, offline training, and telemetry-based analysis [Yuan et al. 2024, Piet et al. 2025, Rabieinejad et al. 2026]. Finally, external modules such as Mitigating Over-defense for Free (MOF) and Low-Rank Adaptation (LoRA) correct undue blockages without retraining the main model, blocking attacks while preserving reasoning on benign inputs [Li et al. 2025, Yang et al. 2025, Liu et al. 2026, Yuan et al. 2024].

4.1. Comparative Analysis

Table 3 compares the main architectures by how security meets performance. We group works sharing an architectural foundation into ten paradigms, ordered by pipeline stage. At the **input level**: *Defensive Suffixes*, appending optimized refusal tokens [Xiong et al. 2025, Yuan et al. 2024]; *Lightweight Classifiers* filtering malicious intent [Zheng et al. 2025, Li et al. 2025, Yan and Wan 2026]; and *Mutations & Instability* via input perturbation [Zhang et al. 2025b]. At the **representation level**: *Representa-*

tion Offset (single pass) [Liu et al. 2026] and *Refusal-Loss Landscape* (multi-query) [Hu et al. 2024]. During **generation**: *Chain-of-Thought* [Sreedhar et al. 2025], *Decoding Interventions* [Yi et al. 2025, Cevallos-Moreno et al. 2026], and *Declarative Steering* for topic/PII control [Krishnan and Siddiquie 2025, Lee et al. 2025a, Hu et al. 2026]. At the **orchestration level**: *Intelligent Routing* [Lee et al. 2025b], *Continual Learning* [Yang et al. 2025, Piet et al. 2025], and *Multi-Agent Telemetry* [Chen et al. 2025, Rabieinejad et al. 2026].

These ten paradigms cover 19 of the 22 studies. The three remaining works were excluded on distinct grounds: S16 [Zhang et al. 2025a] and S18 [Tong et al. 2025] characterize attack vectors against guardrails rather than proposing filtering architectures, whereas S12 [Tyshchenko and Drushchak 2025] assesses and benchmarks existing mechanisms (and contributes the PII strategy discussed in RQ3) rather than introducing a single architecture with its own latency/cost profile.

Table 3. Performance and efficacy of LLM safety architectures. Values are measured in the primary studies; each paradigm is mapped to its source(s) in Section 4.1 and Table 2.

Architecture / Paradigm	Latency (source)	Computational Cost	Efficacy & impact (ASR↓ / FPR where reported)
Mutations & Instability	High (requires N sequential queries)	High (multiplies token usage by N variants)	JailGuard: acc. 81.7%/79.5% (text/image); attack recall 78.0%; FPR 13.8% on benign.
Chain-of-Thought	High (generates reasoning tokens first)	Moderate/High (requires reasoning models)	F1 0.841 (safety), 0.878 on custom policies (vs. 0.845 without CoT); saturates at ≈5k samples.
Continual Learning	Minimal (zero inference overhead)	Moderate (LoRA adaptation)	AdaptiveGuard: OOD F1 96.1%, adapts in ≈2 steps, 85% F1 retained; JailbreaksOverTime: self-retrain cuts FNR 4.0→0.3%.
Defensive Suffixes	Minimal (adds few input tokens)	Low runtime; High training	DPP: ASR 2.0–3.8% (13.0% adaptive); RigorLLM: ToxicChat F1 0.749, 100% HDR under attack.
Representation Offset (single pass)	Low (1.10×; ≈10% wall-clock, 200-tok prompts)	Low (no base-model retraining)	ASR 2.1–3.9%; FPR 2% (↑16% multiling.); 8 jailbreaks, Vicuna/Llama2.
Refusal-Loss Landscape (multi-query)	High (2.3× early-exit; up to 11.2× full benign eval)	Low (110 queries, $N{=}P{=}10$; no retraining)	ASR 11.7%/25.7% (Llama2/Vicuna); FPR 2.2–3.4%; 6 jailbreaks.
Declarative Steering	Low (+18–22 ms NBF filter)	Low (+4.2% tokens vs. baseline)	Topic/PII scope control; SIC not empirically benchmarked.
Decoding Interventions	Minimal (attention-layer intervention)	Low (modifies activations)	E.T.: post-attack toxicity 0.64 (vs. 0.89); retains 83% utility, GSM8K to 37%.
Intelligent Routing	Low (bypasses heavy model for simple inputs)	Low (oracle routes ≈5% to large model)	Router F1 ≈0.50 (oracle 0.81), WildGuardMix.
Multi-Agent Telemetry	Asynchronous (runs in logs)	Low (no full-text LLM eval)	SHIELDAGENT: ACC 90.4–91.7%, rule recall 90.1% @4.8% FPR; ALTEDA: log-based F1 0.893, FP cut >76%.
Lightweight Classifiers	Very low: ≈50 ms (distilBERT, 67M), ≈10 ms (TinyBERT, 11M, CPU), ≈15 ms (DeBERTaV3)	Minimal (CPU-capable; 11M–86M params)	distilBERT: F1 0.89 (AEGIS); LMDMI: 98.7% acc.; PIGuard: over-defense 64→87% (MOF).

4.2. Threats to Validity

As with any systematic review, the design of this study involves methodological choices whose implications are worth making explicit. Screening and extraction were carried out by a single researcher, so we relied on the standardized protocol of Sections 2.3–2.5—explicit inclusion criteria and a fixed extraction form—to keep judgments consistent

and auditable, rather than on inter-rater agreement. Our scope was deliberately bounded to peer-reviewed work in five established databases (ACM DL, IEEE Xplore, Scopus, ScienceDirect, SpringerLink), a choice that favors the reliability of the evidence base over exhaustive coverage; excluding preprints trades some recency for the assurance that every included study has passed peer review. The restriction to English-language venues is a real limitation in this particular field, since a substantial and fast-growing share of LLM research originates in China and often appears first in Chinese-language venues before any English version circulates. Finally, the focus on single-modality text guardrails is what enabled the operational comparison of Section 4; the cross-modal evaluation flagged in Section 5 therefore lies just outside this scope by design, not as a gap in the field.

5. Conclusion and Future Work

This systematic review investigated how guardrails are implemented in AI assistants based on LLMs, focusing on domain restriction and protection against harmful content and data leaks. By analyzing 22 studies published between 2024 and 2026, we were able to identify established patterns and significant gaps in the literature.

Across the research questions: (RQ1) off-topic detection has moved from keyword filters to vector-space analysis and reasoning guardrails, notably ROA and NBF; (RQ2) harmful-content filtering spans three layers—RLHF/SFT alignment, external guardrails, and decoding-time interventions such as FGDILP; (RQ3) PII protection combines semantic classifiers, declarative masking, and telemetry auditing (SHIELDAGENT, Policy Over Tokens); (RQ4) adversarial mitigation is strongest when defensive suffixes, lightweight classifiers, and semantic-deviation analysis are combined (DPP, RigorLLM, JailGuard); and (RQ5) the security–performance balance is sustained by compact classifiers ($\approx$10–50 ms), intelligent routing (oracle bound $\approx$5% escalation, only approximated by trained routers), and MOF/LoRA modules that preserve quality.

Future work should extend guardrails beyond single-model text pipelines. Defensive mechanisms must be evaluated on multimodal LMMs, where visual and audio jailbreaks remain far less studied than text moderation [Zhang et al. 2025b].

Multi-agent ecosystems also need standard guardrails that audit continuously operating agents for hijacking or protocol violations without prohibitive overhead [Rabieinejad et al. 2026, Chen et al. 2025].

Finally, the defenses themselves are becoming attack surfaces: evaluator models and guardrails are vulnerable to backdoor poisoning [Tong et al. 2025] and DoS via induced false positives [Zhang et al. 2025a]. Continual-learning and OOD anomaly detection are promising directions for guardrails that adapt to concept drift without full retraining [Piet et al. 2025, Yang et al. 2025].

References

Borji, A. (2023). A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494*.

Cevallos-Moreno, J. F., Rizzardi, A., Sicari, S., and Coen-Porisini, A. (2026). Ethical treatment of language models against harmful inference-time interventions. *Engineering Applications of Artificial Intelligence*, 172:114317.

Chen, Z., Kang, M., and Li, B. (2025). Shieldagent: Shielding agents via verifiable safety policy reasoning. In *Forty-second International Conference on Machine Learning*.

Dong, Y., Mu, R., Zhang, Y., Sun, S., Zhang, T., Wu, C., Jin, G., Qi, Y., Hu, J., Meng, J., et al. (2025). Safeguarding large language models: A survey. *Artificial intelligence review*, 58(12):382.

Hu, H., Robey, A., and Liu, C. (2026). Steering dialogue dynamics for robustness against multi-turn jailbreaking attacks. *Transactions on Machine Learning Research*.

Hu, X., Chen, P.-Y., and Ho, T.-Y. (2024). Gradient cuff: Detecting jailbreak attacks on large language models by exploring refusal loss landscapes. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37.

Kitchenham, B. A. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01, Keele University and Durham University Joint Report.

Krishnan, S. S. and Siddiquie, D. (2025). Policy over tokens: Enforcing declarative governance constraints in cost-aware llm deployments. In *2025 IEEE International Conference on Agentic AI (ICA)*, pages 202–209. IEEE.

Lee, A. W., Chan, J., Fu, M., Kim, N., Mehta, A., Raghavan, D., and Çetintemel, U. (2025a). Semantic integrity constraints: Declarative guardrails for AI-augmented data processing systems. *Proc. VLDB Endow.*, 18(11):4073–4080.

Lee, S., Lee, D. B., Wagner, D., Kang, M., Seong, H., Bocklet, T., Lee, J., and Hwang, S. J. (2025b). SafeRoute: Adaptive model selection for efficient and accurate safety guardrails in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2053–2069, Vienna, Austria. Association for Computational Linguistics.

Li, H., Liu, X., Zhang, N., and Xiao, C. (2025). Piguard: Prompt injection guardrail via mitigating overdefense for free. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30420–30437, Vienna, Austria. Association for Computational Linguistics.

Liu, S., Cheng, X., Zheng, Z., and Su, S. (2026). Defending llms against jailbreak attacks through representation offset detection. *Information Processing & Management*, 63(5):104662.

Nazir, A. and Wang, Z. (2023). A comprehensive survey of chatgpt: advancements, applications, prospects, and challenges. *Meta-radiology*, 1(2):100022.

Piet, J., Huang, X., Jacob, D., Chow, A., Alrashed, M., Zhao, G., Hu, Z., Sitawarin, C., Alomair, B., and Wagner, D. (2025). Jailbreaksovertime: Detecting jailbreak attacks under distribution shift. In *Proceedings of the 2025 Workshop on Artificial Intelligence and Security (AISec '25)*. ACM.

Rabieinejad, E., Zarrinkalam, F., and Dehghantanha, A. (2026). Beyond the prompt: Log-based threat detection and attribution for multi-agent llms. *Information Processing & Management*, 63:104768.

Sreedhar, M. N., Rebedea, T., and Parisien, C. (2025). Safety through reasoning: An empirical study of reasoning guardrail models. In *Findings of the Association for Com-*

putational Linguistics: EMNLP 2025, Suzhou, China. Association for Computational Linguistics.

Tong, T., Wang, F., Zhao, Z., and Chen, M. (2025). Badjudge: Backdoor vulnerabilities of LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*.

Tyshchenko, V. and Drushchak, N. (2025). Towards safe and secure llm systems: Defense strategies and assessment against inference-time security risks. In *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*, pages 172–179. IEEE.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Xiong, C., Qi, X., Chen, P.-Y., and Ho, T.-Y. (2025). Defensive prompt patch: A robust and generalizable defense of large language models against jailbreak attacks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 409–437.

Yan, S. and Wan, C. (2026). Lmdmi: A lightweight multilevel defense against malicious inputs for generative language models. In *Proceedings of the 2025 International Conference on Artificial Intelligence and Foundation Model (AIFM '25)*, pages 66–71, New York, NY, USA. Association for Computing Machinery.

Yang, R., Fu, M., Tantithamthavorn, C., Arora, C., Gulmammadova, G., and Chua, J. (2025). Adaptiveguard: Towards adaptive runtime safety for llm-powered software. In *2025 40th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 3381–3391. IEEE Press.

Yi, X., Wang, L., Wang, X., and He, L. (2025). Fine-grained detoxification framework via instance-level prefixes for large language models. *Neurocomputing*, 611:128684.

Yuan, Z., Xiong, Z., Zeng, Y., Yu, N., Jia, R., Song, D., and Li, B. (2024). Rigorllm: Resilient guardrails for large language models against undesired content. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235. PMLR.

Zhang, Q., Xiong, Z., and Mao, M. (2025a). Llm safeguard is a double-edged sword: Exploiting false positives for denial-of-service attacks. In *Proceedings of the 2025 Workshop on Large AI Systems and Models with Privacy and Security Analysis (LAMPS '25)*. ACM.

Zhang, X., Zhang, C., Li, T., Huang, Y., Jia, X., Hu, M., Zhang, J., Liu, Y., Ma, S., and Shen, C. (2025b). Jailguard: A universal detection framework for prompt-based attacks on llm systems. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 35(1):Article 8.

Zheng, A., Rana, M., and Stolcke, A. (2025). Lightweight safety guardrails using fine-tuned bert embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING): Industry Track*, pages 689–696.