

Does Fixing Break Security? An Empirical Study of Security Degradation in Iterative LLM-Driven Infrastructure-As-Code Repair

Benjamin Agyekum ✉ 

Department of Electrical Engineering, Colorado State University, Fort Collins, CO, USA

Fabio Santos ✉ 

Department of Computer Science, Colorado State University, Fort Collins, CO, USA

Abstract

Background. Iterative feedback loops have become the dominant paradigm for improving LLM-generated Infrastructure-as-Code (IaC): validators such as Checkov and `terraform validate` feed error signals back to the model for successive repair attempts. Prior work reports *cumulative-best* metrics, which are monotonically non-decreasing by construction, so the raw per-iteration security trajectory has never been examined in the IaC domain.

Aims. We study *security regression* (a previously-passing CIS Benchmark check that fails after a repair iteration) to determine whether, and how often, iterative LLM repair degrades security while fixing other issues.

Method. We analyze 5,968 scenario timelines from the IaC-Eval benchmark, each one scenario run through one configuration for up to 5 repair iterations. The 15 configurations comprise six model-specific RAG and nine model-aggregated non-RAG configurations, three temperatures each, and together they yield 4,440 iteration transitions with Checkov data on both sides. We track 30 individual CIS check IDs and classify regression root causes from code diffs, under two detection modes: *standard* (inclusive) and *strict* (exclusive check failures only).

Results. Under standard (inclusive) detection, 13.8% of scenarios (24.8% of transitions) exhibit at least one regression. Under strict detection, which counts only unambiguous, exclusive check failures, the rate falls to 3.3% of scenarios (5.2% of transitions). This gap indicates that most apparent regressions are multi-resource measurement artifacts rather than genuine exclusive failures. Resource restructuring (79.0%) is the dominant root cause. Regression transitions show $2.6\times$ more code churn (Cohen's $d = 0.90$) and $4.9\times$ higher strict-mode check volatility ($d = 1.49$). Of standard-mode regressions, 36.6% self-correct within an average of 1.2 iterations, and iteration 3 is the optimal stopping point.

Conclusions. Iterative IaC repair does introduce security regressions, but most *apparent* regressions are multi-resource measurement artifacts. The conservative, defensible rate is approximately 3.3% of scenarios. Our findings motivate security-aware feedback-loop design and provide actionable iteration-budget guidance.

2012 ACM Subject Classification Software and its engineering → Automatic programming; Security and privacy → Software security engineering

Keywords and phrases Infrastructure as Code, Security Regression, LLM Code Repair, Terraform, CIS Compliance, Iterative Feedback

Digital Object Identifier 10.4230/LIPICs.ESEM.2026.50

Category Technical Track Paper

Related Version *Full Version:* <https://arxiv.org/abs/2608.13404>

Supplementary Material

Dataset (Replication Package): <https://doi.org/10.5281/zenodo.20265184> [1]



© Benjamin Agyekum and Fabio Santos;
licensed under Creative Commons License CC-BY 4.0

20th International Symposium on Empirical Software Engineering and Measurement (ESEM 2026).

Editors: Robert Feldt, Maria Paasivaara, Daniel Mendez, Stefan Wagner, and Marvin Muñoz Barón;

Article No. 50; pp. 50:1–50:20



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Infrastructure-as-Code (IaC) has become the standard for managing cloud infrastructure, with Terraform serving as the leading declarative provisioning tool [14]. Ensuring security compliance of generated code remains a critical challenge, even after the rise of large language models (LLMs) that are increasingly adopted for automated IaC generation [20, 28]. Studies report that, depending on prompting strategy and validation method, 12–65% of LLM-generated code violates secure coding standards [2]. IaC-specific evaluations find only 7% of generated scripts are secure without explicit security guidance [12].

An increasingly explored approach to improve generated IaC is *iterative feedback*: running static analysis tools (e.g., Checkov [5], terraform validate) on the generated code and feeding error messages back to the LLM for successive repair attempts [20, 28]. Prior work reports promising results from such feedback loops [20, 28]. However, these studies uniformly report *cumulative-best* metrics: the highest compliance achieved across all iterations. These are monotonically non-decreasing by construction and may mask important dynamics in the repair trajectory.

Recent studies on general-purpose code repair have raised alarms about *security degradation*: the phenomenon where iterative LLM repair introduces new vulnerabilities while fixing existing issues. Shukla et al. [25] found a 37.6% vulnerability increase after 5 iterations in C and Java, and Chen et al. [7] demonstrated that 43.7% of GPT-4o iteration chains contain more vulnerabilities than the baseline. Crucially, Chen et al. showed that Static Application Security Testing (SAST)-based gating can *worsen* latent degradation, raising questions about whether Checkov-based feedback in IaC suffers similar effects.

Security regression is a known problem in traditional software evolution. Braz et al. [4] found that regression vulnerabilities in Mozilla are introduced by bug fixes themselves, and security is rarely discussed while the fix is made. Felderer and Fournet [11] provide a taxonomy of security regression testing approaches. However, no study has investigated security regression specifically in the IaC domain with CIS compliance tracking. IaC differs fundamentally from general code: it is declarative rather than imperative, security checks map to specific configuration properties rather than data/control flow, and validators like Checkov operate at the resource level with enumerable check IDs. This resource-level structure is not a cosmetic difference: because one check can apply to several resources within a single configuration, IaC admits a class of *multi-resource ambiguity* that general-code, CWE-based degradation studies cannot exhibit. As we show, this ambiguity accounts for the majority of apparent regressions. Studying regression in IaC therefore requires a detection methodology without direct analogue in prior general-code work: the *standard/strict* distinction we formalize in Section 4.

This paper presents the first large-scale empirical study of security regression in iterative LLM-driven IaC repair. We analyze 5,968 scenario timelines from the IaC-Eval benchmark [16] across 15 configurations (two RAG models and three non-RAG prompting strategies, each run at three temperatures), tracking 30 individual Checkov check IDs mapped to CIS AWS Foundations Benchmark controls [9] across up to 5 repair iterations. Our study addresses four research questions:

- **RQ1:** Does iterative LLM repair introduce security regressions in IaC?
- **RQ2:** What types of security controls are most vulnerable to regression?
- **RQ3:** How do prompting strategy, model, and temperature affect regression?
- **RQ4:** Is there a security-correctness trade-off in iterative IaC repair?

We make four contributions. This is the first large-scale empirical study of security regression in IaC repair, covering 5,968 scenario timelines, 4,440 transitions, and 30 check IDs across 15 configurations, based on single runs per configuration. From it we derive a taxonomy of regression root causes: in standard mode, resource restructuring accounts for 79.0%, configuration drift 15.5%, argument removal 3.6%, and 1.9% remain unclassified. We quantify the security-correctness trade-off, showing that regression transitions carry $2.6\times$ more code churn and $4.9\times$ higher strict-mode check volatility. Finally, we show that the standard and strict detection modes support qualitatively different conclusions, since the Mistral gap vanishes and RAG reverses between them, and we draw practical guidance from this: stop at iteration 3, expect 36.6% of regressions to correct themselves, and prefer RAG where exclusive regressions matter most.

2 Background and Motivation

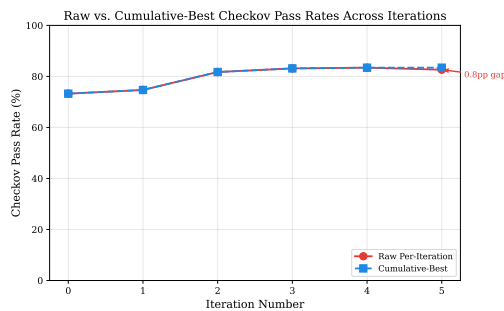
CIS Benchmarks and Checkov. The Center for Internet Security (CIS) AWS Foundations Benchmark v1.5.0 [9] defines security controls for AWS infrastructure. Checkov [5] is a widely-adopted static analysis tool that maps these controls to concrete checks on Terraform configurations. Each Checkov check has a unique ID (e.g., `CKV_AWS_145` for S3 encryption) and can either PASS or FAIL for each resource in a Terraform plan. We map Checkov checks to six security categories: *encryption*, *access control*, *logging*, *networking*, *data protection*, and *other*.

Iterative LLM-Based IaC Repair. The standard IaC feedback loop generates Terraform code from a natural-language prompt, validates it (`terraform validate` for syntax, Checkov for security), summarizes any errors back to the LLM for regeneration, and repeats for up to K iterations. Each pair of consecutive iterations (e.g., $0 \rightarrow 1$) constitutes an *iteration transition*, the unit at which we measure regression. This paradigm is used by TerraFormer [15] and Palavalli et al. [20].

The Hidden Problem of Cumulative-Best Metrics. Prior work on iterative IaC repair universally reports *cumulative-best* metrics: the highest compliance rate achieved up to iteration k . These are monotonically non-decreasing by construction. The *raw per-iteration trajectory*, however, may contain regressions. Figure 1 illustrates this gap. Across 5,968 scenario timelines the cumulative-best CIS pass rate rises from 73% to 83%, yet the raw trajectory *dips* at iteration 5 (82.6% vs. the 83.4% peak at iteration 4) as previously-passing checks fail.

3 Related Work

Security of AI-Generated Code. AI-generated code consistently shows security weaknesses: 29.5% of Python Copilot snippets are vulnerable [13], zero-shot security accuracy is only 37.44% [17], and 12–65% of LLM code violates secure-coding standards [2]. Sajadi et al. [23] found LLMs introduce distinctive vulnerability patterns in automated patches, and Yan et al. [29] found that although over 98% of GPT-4o vulnerability explanations are accurate, guided repair does not always eliminate the underlying issue. We instead track security *changes across iterations* rather than static snapshots, in the IaC domain.



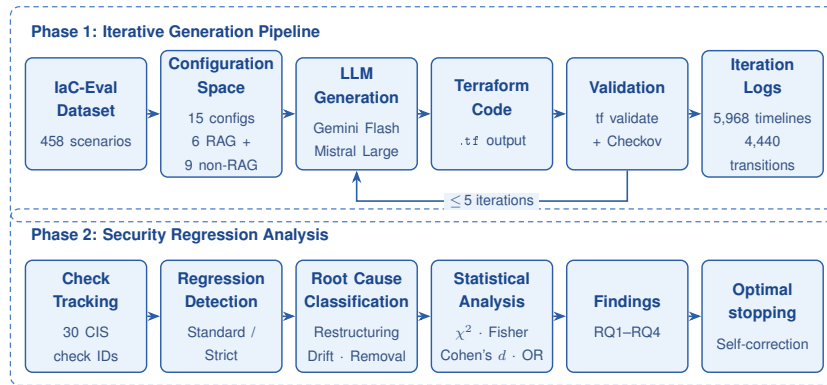
■ **Figure 1** Raw per-iteration vs. cumulative-best Checkov pass rate (5,968 scenario timelines). The raw trajectory *dips* at iteration 5 (82.6% vs. 83.4%), a regression cost masked by cumulative-best reporting.

RAG for Secure Code Generation. RAG improves LLM code security by grounding generation in verified references. Sriram et al. [26] combined RAG with multi-tool feedback (compiler, CodeQL, KLEE) and found retrieval context significantly reduces vulnerabilities, and RESCUE [24] retrieves secure coding patterns at generation time for substantial CWE reduction. Both show RAG’s value for *initial* security but not its stability across iterative repair. Our RAG pipeline grounds generation in CIS documentation and Terraform schema chunks. We find this yields more multi-resource outputs (raising standard-mode counts) but fewer exclusive failures (lowering strict-mode regressions).

IaC-Specific Security and Compliance. Recent IaC work spans generation, validation, and security. TerraFormer [15] combines fine-tuning with policy-guided verifier feedback (+15.94% on IaC-Eval), and Zhang et al. [30] use DevOps simulation as feedback to improve deployment success. TerraFault [28] uses LLM agents to find bugs in IaC updates via plan-level diffs, complementary to our check-level analysis. Firouzi et al. [12] found only 7% of LLM-generated IaC secure without explicit guidance. GenSIaC [18] and Diaz et al. [10] pursue security-aware generation and self-healing, while Palavalli et al. [20] showed feedback effectiveness decays exponentially to a plateau. To the best of our knowledge, we are the first to analyze the per-iteration security trajectory of IaC repair, tracking individual check regressions rather than aggregate compliance.

Iterative Code Repair with LLMs. Iterative LLM feedback is widely applied to code repair. Self-Refine [19] showed ~20% gains from LLM self-critique. RepairAgent [3] reaches state-of-the-art bug repair on Defects4J via multi-tool orchestration, and LLMLOOP [21] improves functional correctness through generation–test feedback. Tang et al. [27] model repair as an exploration–exploitation trade-off in which LLMs under-explore, and Cheng [8] proposes a Detect-Repair-Verify pipeline that evaluates repair side effects, finding repair can introduce new vulnerabilities. These works target functional correctness. None analyze whether iterative repair preserves security properties in IaC, which is our focus.

Security Degradation in Iterative LLM Code Generation. Security regression is well documented in traditional software evolution. Braz et al. [4] studied 78 regression vulnerabilities at Mozilla and found they are introduced by bug fixes themselves. Felderer and Fourneret [11] give a taxonomy of security regression testing. For LLMs, Shukla et al. [25] found a 37.6% vulnerability increase after five iterations of AI code generation, and Chen et al. [7] showed that 43.7% of GPT-4o iteration chains exceed baseline vulnerability counts and that SAST-based gating can *worsen* latent degradation (12.5%→20.8%).



■ **Figure 2** Study overview. Phase 1 generates Terraform through an iterative feedback loop (up to five iterations) across 15 configurations, yielding 5,968 timelines with 4,440 transitions. Phase 2 tracks 30 CIS check IDs to detect, classify, and quantify regressions under standard and strict modes.

Our study is best read as a conceptual replication and extension of Shukla et al. [25] in the IaC domain. We preserve the core design: iterative LLM repair with security measured after every iteration, making degradation visible inside the repair trajectory. We adapt the measurement apparatus to IaC, from imperative C and Java to declarative Terraform, and from multi-tool static analysis with manual review to Checkov’s enumerable check IDs mapped to CIS controls. We extend the design with components absent from the original study: transition-level detection that judges each iteration against its immediate predecessor rather than cumulative vulnerability growth, dual standard/strict detection, a diff-based root-cause taxonomy, self-correction and oscillation analysis, retrieval-augmented configurations, and optimal-stopping analysis.

4 Study Design

Figure 2 presents an overview of our two-phase methodology.

4.1 Dataset and Configurations

We used the IaC-Eval benchmark [16] containing 458 Terraform generation scenarios spanning various AWS services. Each scenario specifies a natural-language infrastructure requirement to be translated into Terraform code.

We evaluated four prompting strategies at three temperatures, which, with RAG analyzed separately per model, give 15 analysis configurations.

- **Prompting strategies** (4): Zero-Shot, Few-Shot, Chain-of-Thought (CoT), and RAG (Retrieval-Augmented Generation with CIS benchmark context). RAG is analyzed separately for each model, giving two groups, while each of the other three forms a single group, for five groups in total.
- **Temperatures** (3): 0.1, 0.4, and 0.7

These five groups at three temperatures give 15 configurations: six RAG (2 models × 3 temperatures), analyzed per model, and nine non-RAG (3 strategies × 3 temperatures), which are not. Both models were run with all four prompting strategies and neither was discarded. The difference is in what the experiment recorded. Every RAG run is stored together with the model that generated it, whereas the non-RAG runs are not, so an individual non-RAG

run cannot be traced back to Gemini or Mistral. Where a scenario was executed more than once under the same non-RAG strategy and temperature, we keep the most recent execution, so each non-RAG timeline is one run from one model rather than an average of the two. For the 36% of non-RAG timelines whose model can still be determined, the split is 66% Gemini to 34% Mistral (893 and 461 of 1,354), so these groups lean toward Gemini rather than being balanced. We analyze each as a single group whose model composition is estimated rather than known.

Running one scenario under one configuration produces a *timeline*: the ordered sequence of code versions from the initial generation (iteration 0) through up to five repair iterations. A timeline ends early once the code passes all checks or the iteration budget is exhausted, so timelines range from a single generation to six versions.

Configuration rationale. The four strategies span the guidance spectrum of prior IaC-generation work: Zero-Shot provides only the task and a security-oriented system prompt. Few-Shot adds three input–output examples. Chain-of-Thought adds an explicit reasoning trace to essentially the same examples. RAG moves to per-query retrieved context from CIS documentation and Terraform schemas. The first three steps each add a single ingredient, permitting descriptive comparison. RAG changes several things at once (its own system-prompt template, no static examples, retrieved context), so we compare it as a paradigm rather than one more increment.

Gemini 2.0 Flash and Mistral Large Latest are production-oriented models from independent vendors, both publicly available at data-collection start (December 2024). Temperatures 0.1, 0.4, and 0.7 cover the lower and middle part of the 0–1 range explored in code-generation evaluations, where low temperatures suit the single-sample setting we use, and higher ones mainly help when many samples are drawn per task [6]. The negligible effect we observe (Cramér’s $V = 0.049$) is consistent with reports that sampling temperature has little influence on task accuracy [22].

We excluded agent-based and self-refinement approaches because they replace the validator-in-the-loop paradigm whose per-iteration behavior we measure (Section 2).

We ran each configuration up to five feedback iterations per scenario, validating with Terraform v1.6.0 (`terraform validate`) and Checkov v3.2.392. Versions were locked at data collection start, December 2024, for reproducibility. Every iteration was recorded with its full Terraform code and individual check results. We limited the feedback loop to five iterations, following prior IaC feedback work that found repair effectiveness plateaus after roughly five iterations [20]. Our regression analysis reconstructs timelines from these per-iteration records, which is where the model attribution described above is lost. We therefore analyzed the six RAG configurations at the model level, Gemini and Mistral, and, for each non-RAG strategy and temperature, analyzed one timeline per scenario. This precludes a model-controlled comparison outside the RAG setting (see Section 7).

The three non-RAG strategies shared a single system prompt that instructed the model to generate CIS-compliant Terraform, combining general directives (least-privilege IAM, encryption at rest, logging, restricted network access) with resource-specific guidance (e.g., configuring S3 public-access controls through a separate `aws_s3_bucket_public_access_block` resource). Few-Shot appended three input–output pairs that map a natural-language request to complete HCL (e.g., “Create an AWS RDS instance ... with randomly generated id and password”). Chain-of-Thought appended essentially the same three example scenarios, each preceded by a reasoning trace that enumerated the required resources, filled in their attributes, and then wired the resources together. RAG used a separately maintained system-prompt

template covering the same core security directives, into which the retrieved CIS-control and resource-schema chunks are injected at query time. The complete templates are included in the replication package [1]. Both models were run with all four strategies. As noted above, only the RAG logs record which model produced each iteration.

4.2 Regression Detection

We define a *security regression* as follows:

► **Definition 1** (Security Regression). *Given consecutive iterations i and $i+1$ of a scenario where both have Checkov results, a regression occurs for check c if $c \in PASSED(i)$ and $c \in FAILED(i+1)$.*

Transitions are formed only between consecutive iterations with *distinct* indices. Where a scenario has two records sharing an iteration index, that pair is excluded: it reflects a repeated attempt at the same step rather than a repair step. We employed two detection modes. **Standard**: any check ID that is in the passed set at iteration i and in the failed set at iteration $i+1$ counts as a regression, regardless of whether it also appears in the passed set (multi-resource scenarios). **Strict**: A check must be *exclusively* passed (not in failed) in iteration i and *exclusively* failed (not in passed) in iteration $i+1$. This excludes ambiguous multi-resource cases. Strict detection trades sensitivity for unambiguity: it can miss genuine degradations that affect only a subset of a configuration’s resources. For example, suppose two S3 buckets pass CKV_AWS_145 (S3 encryption) at iteration i ; at $i+1$ the LLM restructures the code, and one bucket now fails the check. *Standard* mode counts this as a regression (the check is in both the passed and failed sets). *Strict* mode does not (CKV_AWS_145 is not exclusively failed: it still passes for one bucket). We return to this trade-off in Section 7.

4.3 Metrics

Five metrics quantify regression prevalence, the size of the code changes that accompany it, and how often the loop recovers. The *scenario regression rate* is the fraction of scenarios with ≥ 1 regression event, and the *transition regression rate* the fraction of transitions with ≥ 1 regression. *Code churn* covers both lines changed and the text similarity ratio between consecutive iterations, where 0 is completely different and 1 identical. *Check volatility* is $V = |C_{\text{new}}| + |C_{\text{removed}}| + |C_{\text{flipped}}|$, where C_{new} are checks appearing for the first time at iteration $i+1$, C_{removed} are checks present at iteration i but absent at $i+1$, and C_{flipped} are checks changing status, either pass→fail or fail→pass. The *self-correction rate* is the fraction of regressed checks that return to passing later in the same scenario timeline, reported in Section 6.

For RQ4 we additionally examined two derived transition classes. A *post-syntax-fix transition* is one whose prior iteration was syntactically invalid and whose current iteration is valid for the first time, isolating the effect of syntax repair on security. A *syntax-struggled timeline* is a scenario whose iteration sequence contained at least one invalid-syntax iteration, as opposed to a *clean timeline* that was syntactically valid throughout. We compared regression rates between these two groups.

The self-correction rate, root-cause taxonomy, and temporal patterns are reported as deeper analyses in Section 6.

■ **Table 1** Dataset Summary.

Metric	Value
IaC-Eval scenarios	458
Configurations analyzed	15
Max iterations per scenario	5
Total iteration log files parsed	36,736
Unique scenario timelines	5,968
Transitions with Checkov data	4,440
Individual Checkov check IDs tracked	30
CIS security categories	6

4.4 Root Cause Classification

For each regression transition, we classify the root cause from the code diff using four mutually exclusive categories, in order of priority. A transition is *resource restructuring* if the number of Terraform resource blocks changes by ≥ 2 , or if resource types appear or disappear between iterations. The threshold of 2 separates structural rewrites from single-resource edits, where a ± 1 block change is common and would over-trigger the category. Failing that, it is *argument removal* if security-relevant arguments were deleted: the classifier scans removed diff lines for patterns in five groups, covering encryption (`encrypt*`, `kms_key`, `storage_encrypted`, `ssl_policy`), access control (`policy`, `iam_role`, `acl`, `block_public`), logging (`logging`, `flow_log`, `cloudwatch`, `trail`), networking (`security_group`, `cidr_block`, `ingress`, `waf`), and versioning or backup (`versioning`, `backup`, `deletion_protection`). The full regular-expression set is in the replication package [1]. Failing both, it is *configuration drift* if text similarity between iterations exceeds 0.85, meaning small changes flipped a check’s status. The 0.85 cutoff separates near-identical edits, such as a single property change, from substantial rewrites, and we treat it as a heuristic (Section 7). Anything left, such as a diff that failed to parse, is *unclassified*.

4.5 Statistical Methods

We used chi-squared tests for categorical associations, Mann-Whitney U tests for continuous variables (code churn, volatility), Fisher’s exact test for 2×2 tables, and Wilson score intervals for confidence intervals. Effect sizes are reported as Cramér’s V (categorical), Cohen’s d (continuous, used for interpretability despite non-parametric testing), and odds ratios (binary). All tests used $\alpha = 0.05$. We apply Bonferroni family-wise error rate control within each research question. RQ2 corrects $k = 6$ post-hoc category comparisons ($\alpha_{\text{adj}} = 0.0083$). RQ3 corrects $k = 4$ omnibus configuration-factor tests (model, strategy, RAG vs. non-RAG, temperature) and RQ4 $k = 4$ independent tests, both at $\alpha_{\text{adj}} = 0.0125$. All reported effects remain significant after correction. Detailed test specifications are in the replication package [1].

4.6 Scale

Table 1 summarizes our dataset. Of a theoretical maximum of $458 \times 15 = 6,870$ scenario-configuration pairs, 902 (13%) are absent due to API rate-limit errors, provider timeouts, or scenarios that produced no parseable output. Applying the deduplication and same-iteration rules (Sections 4.1 and 4.2) leaves 5,968 scenario timelines and 4,440 iteration transitions with Checkov data on both sides.

■ **Table 2** Regression events by security category (Standard / Strict).

Category	Standard		Strict	
	Count	%	Count	%
Access control	1,017	38.5	26	9.2
Data protection	684	25.9	7	2.5
Encryption	357	13.5	61	21.6
Networking	375	14.2	120	42.6
Logging	152	5.8	38	13.5
Other	54	2.0	30	10.6
Total	2,639		282	

5 Results

The four research questions build on one another: RQ1 measures how often regressions occur, RQ2 where they concentrate, RQ3 which configurations amplify them, and RQ4 what code-level behavior accompanies them. Section 6 then examines causes, timing, and stopping policy.

5.1 RQ1: Does Iterative LLM Repair Introduce Security Regressions?

In the standard analysis, of the 5,968 scenario timelines, 823 (13.8%, 95% CI: [12.9%, 14.7%]) exhibit at least one security regression event. Across the 4,440 transitions with Checkov data on both sides, 1,103 (24.8%, CI: [23.6%, 26.1%]) contain at least one regression, producing 2,639 total regression events.

Under strict detection (excluding ambiguous multi-resource cases), 194 scenarios (3.3%, CI: [2.8%, 3.7%]) exhibit regression, with 282 total events across 231 transitions (5.2% of transitions, CI: [4.6%, 5.9%]). Scenarios fall 4.2 $\times$, from 823 to 194. Transitions fall comparably, 1,103 to 231. Approximately 76% of standard-mode-flagged scenarios are not flagged in strict mode, consistent with multi-resource ambiguity rather than exclusive check failures. We use the scenario-level 76% as the headline artifact fraction.

We measure the fraction of scenarios with ≥ 1 regression event (13.8% standard, 3.3% strict), while Shukla et al. [25] measure cumulative vulnerability growth (37.6% after 5 iterations in C and Java). Metric differences limit direct comparison. The lower IaC rates may reflect its declarative nature and structured validation, but a definitive comparison requires metric alignment across domains.

Finding: 13.8% of IaC scenario timelines experience security regression in iterative repair, lower than general code (37.6%) but still substantial. Strict analysis shows 3.3%, indicating most regressions involve multi-resource ambiguity.

5.2 RQ2: What Types of Security Controls Are Most Vulnerable?

RQ1 established how often regressions occur. RQ2 asks where they land. Table 2 shows the regression distribution across security categories. Regression events are highly concentrated rather than spread evenly across the six categories: a chi-squared goodness-of-fit test against a uniform reference rejects an even split ($\chi^2 = 1445$, $p < 0.001$, Cramér’s $V = 0.33$, medium-to-large effect). We report this test only to quantify the skew (uniformity is not a theoretically motivated null) and use the per-category breakdown below to identify its drivers.

In standard mode, *access control* dominates (38.5%), driven by three IAM-related checks: CKV_AWS_356 (298 events), CKV_AWS_111 (297), and CKV_AWS_109 (291). These checks evaluate IAM policy conditions, wildcard permissions, and role trust boundaries, properties frequently modified during code restructuring.

Strikingly, the category ranking *reverses* in strict mode: *networking* (42.6%) and *encryption* (21.6%) dominate, while access control drops to 9.2%. This reversal indicates that access control regressions are predominantly multi-resource ambiguity, while networking and encryption represent genuine exclusive check failures. The mechanism is structural: IAM scenarios typically generate multiple resources (e.g., `aws_iam_role` and `aws_iam_policy`) where a check like CKV_AWS_356 may pass for one resource but fail for another, creating standard-mode regressions without exclusive failure. In contrast, networking checks (e.g., VPC endpoints) and encryption checks (e.g., KMS key rotation) typically map one-to-one to resources, making multi-resource ambiguity rare.

Operational impact. Not all regressing checks carry equal risk. The three most frequently regressing checks all govern IAM privilege boundaries, among the highest-impact misconfigurations operationally: CKV_AWS_356, CKV_AWS_111, and CKV_AWS_109 (886 events, 33.6% of all standard-mode regressions). Aggregating by category, the high-impact categories (access control, encryption, networking) account for 1,749 events (66.3%), while advisory logging checks account for only 5.8%. This coarse lens indicates that regressions concentrate in the controls practitioners care about most, not in low-stakes advisory checks. A full severity- or exploitability-weighted analysis is future work.

Finding: Access control dominates standard regressions (38.5%) but drops to 9.2% in strict mode. Networking and encryption represent the most genuine regression risks, and the most frequent regressing checks (886 events, 33.6%) all govern IAM privilege.

5.3 RQ3: How Do Configuration Factors Affect Regression?

Knowing how often (RQ1) and where (RQ2) regressions strike, we next ask which configurations amplify them, examining four factors: model, prompting strategy, retrieval augmentation, and temperature. One caveat applies throughout. The logs record the generation model only for RAG runs, so non-RAG results aggregate over an unidentified model and may partly reflect model behavior rather than the named factor. We flag this in each affected result and return to it in Section 7.

In standard mode, the model effect is dramatic: Gemini’s scenario regression rates range from 1.9% to 4.7% across temperatures, while Mistral ranges from 32.6% to 41.2%. Mistral scenarios are over 17 times more likely to regress ($OR = 17.29$, $p < 0.001$). However, this gap **vanishes entirely** in strict mode: neither RAG+Gemini nor RAG+Mistral produces a single strict-mode scenario regression (0 of 1,297 and 0 of 929 respectively), indicating that Mistral’s elevated standard-mode rate stems from multi-resource ambiguity rather than exclusive check failures. RAG+Mistral transitions also exhibit substantially larger code changes than RAG+Gemini (median 32 vs. 8 lines changed), consistent with Mistral regenerating whole resource blocks and thus producing more multi-resource configurations.

Regression rates differ significantly across the five strategy/model groups. A 5×2 test of the five strategy/model groups in Table 3 against regressed vs. non-regressed scenario counts gives $\chi^2 = 596$, $dof = 4$, $p < 0.001$, Cramér’s $V = 0.32$ (medium effect). Among non-RAG strategies, Chain-of-Thought achieves the lowest regression rates (8.2–9.6%), followed by Few-Shot (11.6–12.9%) and Zero-Shot (11.6–15.5%), which would suggest that more structured

■ **Table 3** Scenario regression rates by configuration and temperature. N = scenarios analyzed. Reg = scenarios with ≥ 1 regression. RAG rows are reported per model. Non-RAG rows are not split by model (Section 4).

Configuration	$t = 0.1$			$t = 0.4$			$t = 0.7$		
	N	Reg	Rate	N	Reg	Rate	N	Reg	Rate
RAG + Gemini	413	8	1.9%	447	21	4.7%	437	15	3.4%
RAG + Mistral	138	45	32.6%	439	161	36.7%	352	145	41.2%
Zero-Shot	354	55	15.5%	458	53	11.6%	458	62	13.5%
Few-Shot	419	54	12.9%	336	39	11.6%	458	53	11.6%
Chain-of-Thought	453	37	8.2%	369	33	8.9%	437	42	9.6%

prompting reduces regression risk. That reading does not survive the model composition of these groups. Among the non-RAG timelines whose model can be determined (Section 4), Chain-of-Thought is 79% Gemini, Few-Shot 73%, and Zero-Shot only 46%, the same order as the regression ranking. Because Gemini regresses far less often than Mistral under RAG (1.9–4.7% vs. 32.6–41.2%), a ranking of this shape is what differing model mixes alone would produce. We therefore report the ordering as descriptive of the non-RAG configurations as deployed and draw no causal conclusion about prompting strategy.

The RAG effect *reverses* between modes. In standard mode, RAG has a *higher* regression rate: a 2×2 test of RAG vs. non-RAG scenarios against regressed-vs-non-regressed counts gives $OR = 1.67$, $p < 0.001$. In strict mode, however, RAG records *zero* scenario regressions, versus 4–6% for the non-RAG strategies. This means RAG produces *fewer exclusive regressions* but generates more multi-resource code structures that inflate the standard count (see Section 7 for the retrieval mechanism explanation). Note that this comparison is not controlled for model: the non-RAG timelines are not attributable to a specific model in our logs, whereas RAG includes both Gemini and Mistral.

Temperature has a statistically significant but practically negligible effect (a 3×2 contingency test over the three temperatures: $\chi^2 = 14.3$, $p < 0.001$, Cramér’s $V = 0.049$). Higher temperatures show slightly elevated regression rates, but the effect size is minimal.

Finding: The model gap (Mistral $17\times$ worse than Gemini in standard mode) vanishes entirely in strict mode (neither RAG model produces a single strict regression), revealing it as a multi-resource artifact. RAG reverses from worse to better between standard and strict analysis.

5.4 RQ4: The Security-Correctness Trade-off

The remaining question is mechanism: RQ4 tests whether fixing correctness comes at a security cost, from four angles: the code churn and check volatility of regressing transitions, the regression rate of transitions that immediately follow a syntax fix, and the regression rate of timelines that struggled with syntax.

We compare code churn between all Checkov-bearing transitions that contain a regression and those that do not, pooled across every strategy, model, and temperature. Transitions with regressions show significantly more code modification than those without: the average number of lines changed between consecutive iterations is 140.1 vs. 53.8, a $2.6\times$ difference. The text similarity ratio between iterations also drops from 0.847 to 0.791, indicating that regression transitions involve more extensive rewrites. Mann-Whitney U tests confirm significance ($p < 0.001$, Cohen’s $d = 0.90$ for lines changed, large effect).

Check volatility measures the total number of check-level changes between consecutive iterations (Section 4). It is strongly elevated in regression transitions. In standard mode, average volatility is 5.17 with regression vs. 2.09 without, a $2.5\times$ difference. In strict mode

■ **Table 4** Root cause classification of regression transitions.

Root Cause	Standard		Strict	
	Count	%	Count	%
Resource restructuring	871	79.0	158	68.4
Configuration drift	171	15.5	46	19.9
Argument removal	40	3.6	19	8.2
Unclassified	21	1.9	8	3.5

it is even stronger: 11.55 vs. 2.38 ($4.9\times$), with Cohen’s $d = 1.49$, a very large effect and the single strongest signal in our study, making check volatility the strongest indicator of security regression.

In transitions immediately following a syntax fix, the prior iteration had invalid syntax, and the current iteration first achieves validity. These show a 21% higher regression rate than transitions between two already-valid iterations (28.1% vs. 23.3%, $OR = 1.29$, $p < 0.001$). The two groups partition the 4,440-transition universe: 1,449 and 2,991 transitions. Their regression counts, 407 and 696, sum to 1,103, the count underlying the 24.8% aggregate rate. This effect is not significant in strict mode after Bonferroni correction ($p = 0.045$, above the adjusted $\alpha = 0.0125$), suggesting that post-fix regressions are predominantly multi-resource ambiguity introduced when the LLM restructures code to achieve validity.

Finally, we analyze the syntax struggle paradox. Counter-intuitively, scenarios that *always* had valid syntax (“clean” timelines) show a *higher* regression rate than those that struggled with syntax errors: clean timelines regress in 20.8% of cases versus 12.1% for syntax-struggled timelines. A 2×2 test (clean vs. syntax-struggled timelines $\times$ scenarios with ≥ 1 regression vs. none) gives $OR = 1.90$ ($p < 0.001$). We discuss two candidate mechanisms for this counter-intuitive result in Section 7.

Finding: Larger code changes strongly predict regression ($2.6\times$ more churn, $d = 0.90$). Check volatility is the strongest predictor ($d = 1.49$ strict). Syntax-fixing transitions carry 21% higher regression risk.

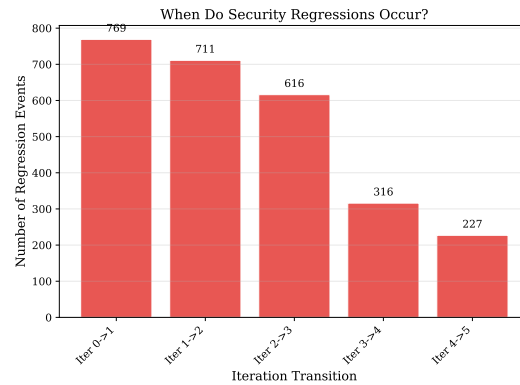
6 Deep Dive Analysis

This section reports analyses that extend the four research questions rather than introducing new ones: the root-cause taxonomy and temporal patterns deepen RQ1 and RQ2 by characterizing *how* and *when* regressions arise, while the optimal-stopping analysis operationalizes the RQ4 security-correctness trade-off into a concrete iteration budget.

6.1 Root Cause Taxonomy

We classified the root cause of every regression transition (1,103 standard and 231 strict) from the code diff between consecutive iterations, using resource-block changes, text similarity, and deleted security arguments (Section 4). Table 4 shows the distribution.

Resource restructuring. (79.0%) is the dominant cause. The LLM adds, removes, or renames resource blocks while fixing validation errors, and the restructured code loses security configurations from the prior iteration. This is especially common in access control, reflecting how tightly IAM policies are coupled to resource structure.



■ **Figure 3** Regression events by iteration transition: counts peak in early-to-middle transitions, then decline as code stabilizes.

Configuration drift. (15.5%) involves small modifications that flip a check’s status, typically changing a single property value or adding/removing an attribute. It most often affects networking and encryption checks, where small changes to security-group rules or KMS key references can fail specific checks.

Argument removal. (3.6% standard, 8.2% strict) represents explicit deletion of security-relevant Terraform arguments. Its proportion more than doubles in strict mode, indicating these are genuine security regressions rather than ambiguous multi-resource artifacts.

Temporal Patterns. Beyond *which* checks regress, we examine *when* regressions occur across the iteration sequence, whether they are subsequently undone, and how often checks flip repeatedly.

When Do Regressions Occur? Figure 3 shows the distribution of regression events across iteration transitions. The highest counts occur early (0→1: 769 events, 1→2: 711, 2→3: 616), reflecting the larger number of scenarios still iterating. Normalized by per-slot transition counts, however, the per-transition *rate* peaks at 2→3 (29.5%) and 1→2 (26.6%), while 0→1 is lower (22.8%), consistent with the 24.8% aggregate rate. One plausible explanation, unconfirmed without per-transition modification analysis, is that early iterations make conservative fixes while middle iterations restructure more aggressively. Late iterations (4→5: 20.8%) show lower rates, possibly reflecting code stabilization.

Self-Correction. Of 2,639 standard-mode regressions, 967 (36.6%) self-correct in a subsequent iteration, meaning the regressed check returns to passing status later in the same scenario’s repair timeline. On average, self-correction takes 1.2 iterations after the regression occurs (median: exactly 1), with 80.4% correcting within a single step (i.e., at the next repair iteration). This indicates that the feedback loop often catches and reverses its own mistakes. Self-correction is at least as common in strict mode (44.0%, mean 1.2 iterations). We report the standard-mode 36.6% as the headline figure because its far larger event pool (2,639 vs. 282 events) gives a more reliable estimate.

Check Oscillation. 28.5% of all scenarios (1,698 of 5,968) exhibit *oscillating* checks, i.e., checks that pass, then fail, then pass again (or vice versa) across iterations. The rate is identical under strict detection, as oscillation is a timeline property independent of the

■ **Table 5** Compliance vs. regression trade-off by iteration (standard mode). N varies as scenarios terminate upon full compliance or budget exhaustion. Average cumulative regressions may decrease when high-regression scenarios terminate earlier, leaving a lower-regression survivor pool.

Iter	N	Avg Pass	Avg Cum. Reg	% w/ Reg
0	2,649	73.2%	0.00	0.0%
1	3,151	74.7%	0.28	13.0%
2	2,726	81.7%	0.49	17.8%
3	1,878	83.1%	0.68	20.8%
4	1,393	83.4%	0.64	18.8%
5	1,029	82.6%	0.94	23.3%

multi-resource threshold. The three most oscillation-prone checks are all IAM-related: CKV_AWS_356 (925 timelines), CKV_AWS_111 (922), and CKV_AWS_109 (832), suggesting the LLM lacks stable strategies for complex IAM configurations and cycles between implementations satisfying different subsets of checks.

Optimal Stopping Point. Table 5 combines average pass rate and cumulative regression data to identify the optimal stopping point. The “N” column counts timelines with Checkov data at each iteration: it rises from iteration 0 to 1 (2,649 $\rightarrow$ 3,151) as initially invalid scenarios become checkable, then falls as scenarios reach full compliance or exhaust their budget.

Iteration 3 offers the best trade-off. We anchor this recommendation on the pass-rate trajectory, which is the more reliable of the two signals: the pass rate reaches 83.1% at iteration 3, within 0.3pp of the 83.4% maximum at iteration 4, and then *decreases* to 82.6% at iteration 5. If we define marginal gains as <0.5pp (with ~ 30 tracked checks, 0.5pp $\approx$ 0.15 checks), the iteration 3 $\rightarrow$ 4 gain falls below this threshold, so iteration 3 captures essentially all attainable improvement.

The average cumulative-regression column tells a consistent but weaker story, and we treat it as secondary because it is *not monotone*: it rises 0.00 $\rightarrow$ 0.68 through iteration 3, dips to 0.64 at iteration 4, then rises to 0.94 at iteration 5. The dip is a survivor-bias artifact, since high-regression scenarios disproportionately terminate by iteration 3 and leave a lower-regression survivor pool, so the column alone cannot serve as a stopping criterion. It does show that regressions keep accumulating past iteration 3. This matches Palavalli et al.’s [20] finding of exponential feedback decay and suggests that adaptive stopping policies (e.g., halt when $\Delta\text{pass_rate} < 0.5\text{pp}$ for two consecutive iterations) could optimize the compliance-regression trade-off per scenario.

7 Discussion

Our results tell a single story. Iterative repair does break previously-passing checks (RQ1), concentrated in IAM, networking, and encryption controls (RQ2). The damage tracks how much code is rewritten far more than any configuration knob (RQ3–RQ4). Much of what standard detection flags is multi-resource bookkeeping rather than exclusive failure. The implications below follow from this reading. The following design hypotheses for IaC tooling are motivated by our observations. None has been implemented or evaluated in this study. We offer them as directions for future tooling rather than validated solutions.

Security anchoring. The dominant root cause, resource restructuring (79.0%), suggests that LLMs should be constrained to make *minimal modifications* during repair rather than regenerating entire resource blocks. A “security anchor” mechanism could lock passing checks and only modify failing ones.

Check-aware feedback. Current feedback loops report all errors without distinguishing new failures from pre-existing ones. Explicitly flagging regressions (“Warning: CKV_AWS_145 was passing but now fails”) could prevent the LLM from unknowingly degrading security.

Iteration budgets. Our finding that iteration 3 is the optimal stopping point supports a conservative iteration budget. Tools should not blindly iterate to convergence. Three iterations capture most of the compliance improvement while keeping regression risk manageable.

Choosing a configuration. Our results support cost- and risk-differentiated guidance rather than a single recommendation. Where retrieval infrastructure is unavailable or too costly, Chain-of-Thought showed the lowest non-RAG regression rates (8.2–9.6%) at no infrastructure cost beyond a longer prompt. That ranking tracks the model mix of the non-RAG groups rather than the strategies themselves (Section 5), so it is descriptive rather than causal. Where stability of individual security properties is paramount, RAG is preferable, since it produced zero strict-mode regressions. Its price is maintaining a retrieval index and triaging the standard-mode alerts that its multi-resource output generates. In either case the iteration budget above applies. Check volatility is cheap to compute online and is the strongest regression signal in our data ($d = 1.49$, strict mode). Halting or flagging a repair loop when volatility spikes offers regression protection at negligible cost. A single regression event, by contrast, is a premature halting signal: 36.6% of regressions self-correct within the loop.

Standard vs. Strict: Two Valid Perspectives. The dramatic differences between standard and strict analysis (13.8% vs. 3.3% scenario rate, Mistral $17\times$ worse in standard mode vs. zero RAG regressions in strict) reflect a fundamental tension in how regression is defined for multi-resource IaC. Standard mode captures the practitioner view: any degradation of a check for *any* resource counts, while strict mode captures the analyst view: only unambiguous, exclusive regressions count. Strict mode’s exclusivity requirement carries a security cost: a check that fails for several resources while still passing for one is invisible to strict detection, so a transition that degrades most, but not all, instances of a control is not counted. Strict mode is thus a measurement-validity device, not an operational security criterion: it isolates regressions that cannot be explained by multi-resource bookkeeping, at the deliberate price of under-counting genuine partial degradations. We therefore recommend reporting both modes in distinct roles: the strict figures anchor the conservative point estimate we defend as our headline result, while the standard figures give the inclusive outer bound an operator should act on: production monitoring should triage standard-mode alerts precisely because strict detection can miss partial degradations, whereas research comparison and cross-study benchmarking should use strict mode.

How Real Is the Phenomenon? A skeptical reading must confront the gap between our standard and strict figures. The standard-mode rates (13.8% of scenarios, 24.8% of transitions) are an *inclusive upper bound*: roughly three quarters (76% of standard-mode regression scenarios, a comparable 79.1% reduction at the transition level) are multi-resource bookkeeping artifacts rather than exclusive check failures (Section 5). The conservative, defensible claim is the strict rate: approximately 3.3% of scenarios (5.2% of transitions) exhibit an unambiguous, exclusive regression.

50:16 Does Fixing Break Security?

Three caveats compound around even the strict rate. We ran each configuration once rather than repeating the experiment, so run-to-run variance is unquantified. The root-cause taxonomy is automated diff-classifier output, never validated against human judgment. Some timelines may mix iterations from separate executions. Together these mean the strict 3.3% should be read as an order-of-magnitude estimate rather than a precise point value. What the data *does* support robustly is the qualitative claim: iterative IaC repair sometimes degrades a previously-passing security check, the effect is not negligible, and cumulative-best reporting hides it entirely.

Self-Correction and Oscillation. The 36.6% self-correction rate implies that single-iteration regression is an imperfect halting signal: roughly one-third of regressions are transient, resolved by the same feedback mechanism that caused them. The 28.5% oscillation rate, however, shows many corrections are themselves unstable. A check may regress, recover, and regress again, so not all self-corrections are stable recoveries. The three most oscillation-prone checks (CKV_AWS_356, CKV_AWS_111, CKV_AWS_109) are all IAM policy checks, suggesting the LLM cycles between implementations satisfying different subsets of IAM constraints. Stabilizing them needs more structured retrieval context or constrained generation.

RAG Reversal and Temperature Effects. The RAG reversal effect (RAG appears *worse* in standard mode but *better* in strict mode) reveals that RAG generates more multi-resource configurations while producing fewer exclusive regressions. We attribute this to retrieval: a query mentioning “VPC with logging” retrieves schema chunks for `aws_vpc`, `aws_flow_log`, and related log-storage resources. That context encourages comprehensive multi-resource solutions, increasing the opportunity for check ambiguity in standard mode. However, the same contextual grounding provides stability for individual security properties, explaining why RAG achieves fewer exclusive regressions in strict mode. The negligible temperature effect (Cramér’s $V = 0.049$) likewise suggests regression is driven by structural generation patterns rather than sampling randomness.

Why Clean Timelines Regress More. A counter-intuitive RQ4 result is that scenarios with consistently valid syntax (“clean” timelines) regress *more* often than those that struggled with syntax errors (OR = 1.90). We identify two non-mutually-exclusive explanations. First, an *opportunity effect*: clean timelines have more consecutive valid-to-valid transitions, providing more opportunities for Checkov-level regressions to occur, since only syntactically valid code undergoes security evaluation. Second, a *feedback purity effect*: when the LLM receives only security feedback (no syntax errors), it aggressively restructures already-valid code in pursuit of compliance, inadvertently breaking passing checks. Disentangling these mechanisms would require an ablation study varying feedback types, which we leave to future work.

What Generalizes Beyond Terraform and Checkov? Which findings are specific to Terraform and Checkov, and which plausibly characterize iterative LLM repair at large? Multi-resource ambiguity, and with it the standard/strict gap and the mode-dependent reversals of RQ3, arises because one check can apply to several resources in a configuration. We expect it to transfer to other IaC ecosystems (CloudFormation, Ansible, tfsec, Terrascan), which likewise evaluate checks per resource, but not to general-purpose code, where findings are not multiplied across resource instances. By contrast, three observations are candidates for domain-general behavior: large rewrites are strongly associated with regressions, restructuring

rather than minimal editing dominates, and repair gains diminish across iterations. The first two are consistent with the degradation Shukla et al. [25] and Chen et al. [7] observe in general-purpose code, and the third is replicated within IaC by Palavalli et al.’s feedback plateau [20]. Whether self-correction and oscillation recur outside IaC is unexamined. These are hypotheses: confirming them requires replications that vary domain and validator while holding the repair loop fixed.

From a human-factors perspective, the contrast is instructive. Braz et al. [4] found that developers focus on the complexity of the bug at hand and work under community pressure to deliver, while security goes undiscussed during the fix. The mechanism we observe differs: LLM regressions arise within a single repair step, when the model rewrites structure wholesale rather than editing minimally. The blind spot, however, is shared. Just as security went unexamined during Mozilla’s bug fixes, current feedback loops never tell the model which checks its previous iteration was already passing.

Threats to Validity

Internal. Our root-cause classification relies on automated diff analysis: it was not validated against human labels, may misclassify boundary cases (e.g., resource renaming), and its thresholds (>0.85 similarity for drift, ≥ 2 resource-count change for restructuring) were chosen conservatively rather than empirically tuned. Alternatives could shift the distributions. We mitigate this with the dual standard/strict framework, but manual validation of a stratified sample remains future work. The per-iteration logs record the generation model for RAG runs but not non-RAG runs, precluding a model-controlled RAG-vs-non-RAG comparison and the isolation of retrieval-augmentation effects from model behavior. We partially mitigate this by reporting RAG+Gemini and RAG+Mistral separately (Table 3). Where a scenario ran more than once we keep the later execution and drop same-iteration transitions, yet consecutive iterations in such a timeline may still come from different attempts. The 902 absent scenario-configuration pairs (13%) stem from API errors, timeouts, and unparseable outputs, and are unlikely to bias our estimates. We use single runs per configuration for computational reasons, since repeated trials (e.g., $n = 3$) would require roughly 1.3M additional API calls, and although the negligible temperature effect ($V = 0.049$) suggests stability, point estimates may vary ± 2 –5pp. Finally, we treat transitions as independent for Mann-Whitney tests despite possible within-timeline autocorrelation, our iteration-3 stopping point rests on a subjective <0.5 pp threshold, and the post-hoc root-cause taxonomy and exploratory subgroup analyses were not pre-registered.

External. Our results are specific to Terraform and Checkov. Other IaC tools (Ansible, CloudFormation) and validators (tfsec, Terrascan) may exhibit different regression patterns. Our dataset (IaC-Eval v1.0, 458 scenarios, accessed December 2024) covers AWS only. We ran 24 experiments in total, all four strategies on both models at three temperatures. Only the six RAG experiments record which model produced each run, so the eighteen non-RAG experiments aggregate into nine model-aggregated configurations, giving the 15 we analyze (Section 4). Attributing all 24 would enable cleaner causal inference. Our findings may also partly reflect the repair capability of the two studied models. More recent or reasoning-oriented models may restructure less aggressively or preserve passing checks more reliably. Replication with newer models is needed to separate paradigm-level from model-level effects.

Construct. Checkov provides a proxy for security. Real security assessment requires deployment-time evaluation (`terraform plan`, sandbox testing). We use static validation rather than deployment for scalability (5,968 deployments would incur substantial cost),

reproducibility (outcomes vary by account state), and safety (some scenarios create production resources). Some Checkov checks may be overly strict or context-dependent. Our study also compares RAG holistically without isolating individual components (CIS docs vs. Terraform schema vs. retrieval depth). Future ablation studies should test these variants separately.

8 Conclusion

We presented the first large-scale empirical study of security regression in iterative LLM-driven IaC repair, analyzing 5,968 scenario timelines across 15 configurations with up to 5 repair iterations each. Six findings emerge. (1) **Regression is real:** 13.8% of scenarios regress in standard mode, 3.3% in strict mode. (2) **Detection mode changes conclusions:** access control dominates standard mode while networking and encryption are the genuine strict-mode risks, the 17× Mistral-vs-Gemini gap vanishes, and RAG reverses from worst to best. Multi-resource ambiguity drives standard-mode differences. (3) **Resource restructuring** is the dominant root cause (79.0%), ahead of configuration drift (15.5%). (4) **Code churn and check volatility are indicators of regressions**, with strict-mode volatility the strongest signal ($d = 1.49$). (5) **Self-correction is common but unstable:** 36.6% of standard-mode regressions self-correct, yet 28.5% of scenarios oscillate. (6) **Iteration 3 is the optimal stopping point**, balancing an 83.1% pass rate against manageable regression risk. Together, these findings provide actionable guidance for designing security-aware IaC feedback loops.

Data Availability Statement

Our replication package is openly available [1]. It contains the analysis scripts, the derived results that reproduce every table, figure, and statistic reported here, the complete prompt templates, a README documenting the pipeline, and a slimmed subset of the raw per-iteration logs so the pipeline can be exercised end to end.

References

- 1 Benjamin Agyekum and Fabio Santos. Replication package: Does fixing break security? an empirical study of security degradation in iterative llm-driven infrastructure-as-code repair. Zenodo, 2026. doi:10.5281/zenodo.20265184.
- 2 Enna Basic and Alberto Giarretta. From vulnerabilities to remediation: A systematic literature review of LLMs in code security. *arXiv preprint arXiv:2412.15004*, 2024. doi:10.48550/arXiv.2412.15004.
- 3 Islem Bouzenia, Premkumar Devanbu, and Michael Pradel. RepairAgent: An autonomous, LLM-based agent for program repair. In *Proceedings of the 47th IEEE/ACM International Conference on Software Engineering (ICSE)*, pages 2188–2200. IEEE/ACM, 2025. doi:10.1109/ICSE55347.2025.00157.
- 4 Larissa Braz, Enrico Fregnan, Vivek Arora, and Alberto Bacchelli. An exploratory study on regression vulnerabilities. In *Proceedings of the 16th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 12–22, 2022. doi:10.1145/3544902.3546250.
- 5 Checkov. Checkov: Infrastructure as code static analysis. Available: <https://www.checkov.io>. Accessed: Mar. 25, 2025.
- 6 Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. doi:10.48550/arXiv.2107.03374.

- 7 Yi Chen, Yun Bian, Haiquan Wang, Shihao Li, and Zhe Cui. SCAFFOLD-CEGIS: Preventing latent security degradation in LLM-driven iterative code refinement. *arXiv preprint arXiv:2603.08520*, 2026. doi:10.48550/arXiv.2603.08520.
- 8 Cheng Cheng. Detect repair verify for securing LLM generated code: A multi-language empirical study. *arXiv preprint arXiv:2603.00897*, 2026. doi:10.48550/arXiv.2603.00897.
- 9 Cis benchmark. Available: <https://www.cisecurity.org/cis-benchmarks>. Accessed: Mar. 25, 2025.
- 10 J. Diaz-de Arcaya, J. López-de Armentia, G. Zárate, and A. I. Torre-Bastida. Towards the self-healing of infrastructure as code projects using constrained llm technologies. In *Proceedings of the 5th ACM/IEEE International Workshop on Automated Program Repair*, pages 1–4, 2024. doi:10.1145/3643788.3648014.
- 11 Michael Felderer and Elizabeta Fournieret. A systematic classification of security regression testing approaches. *International Journal on Software Tools for Technology Transfer*, 17(3):305–319, 2015. doi:10.1007/s10009-015-0365-2.
- 12 Ehsan Firouzi, Shardul Bhatt, and Mohammad Ghafari. Can developers rely on LLMs for secure IaC development? *arXiv preprint arXiv:2602.03648*, 2026. doi:10.48550/arXiv.2602.03648.
- 13 Yujia Fu, Peng Liang, Amjed Tahir, Zengyang Li, Mojtaba Shahin, Jiaxin Yu, and Jinfu Chen. Security weaknesses of Copilot-generated code in GitHub projects: An empirical study. *ACM Transactions on Software Engineering and Methodology*, 34(8):218:1–218:34, 2025. doi:10.1145/3716848.
- 14 HashiCorp. Terraform by hashicorp. Available: <https://www.terraform.io/>, 2024.
- 15 Prithwish Jana, Sam Davidson, Bhavana Bhasker, Andrey Kan, Anoop Deoras, and Laurent Callot. TerraFormer: Automated infrastructure-as-code with LLMs fine-tuned via policy-guided verifier feedback. *arXiv preprint arXiv:2601.08734*, 2026. doi:10.48550/arXiv.2601.08734.
- 16 Patrick Tser Jern Kon, Jiachen Liu, Yiming Qiu, Weijun Fan, Ting He, Lei Lin, Haoran Zhang, Owen M. Park, George S. Elengikal, Yuxin Kang, Ang Chen, Mosharaf Chowdhury, Myungjin Lee, and Xinyu Wang. IaC-Eval: A code generation benchmark for cloud infrastructure-as-code programs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi:10.52202/079017-4273.
- 17 Xinghang Li, Jingzhe Ding, Chao Peng, Bing Zhao, Xiang Gao, Hongwan Gao, and Xinchun Gu. SafeGenBench: A benchmark framework for security vulnerability detection in LLM-generated code. *arXiv preprint arXiv:2506.05692*, 2025. doi:10.48550/arXiv.2506.05692.
- 18 Yikun Li, Matteo Grella, Daniel Nahmias, Gal Engelberg, Dan Klein, Giancarlo Guizzardi, Thijs van Ede, and Andrea Continella. GenSIaC: Toward security-aware infrastructure-as-code generation with large language models. *arXiv preprint arXiv:2511.12385*, 2025. doi:10.48550/arXiv.2511.12385.
- 19 Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. doi:10.52202/075280-2019.
- 20 Mayur Amarnath Palavalli and Mark Santolucito. Using a feedback loop for LLM-based infrastructure as code generation. *CoRR*, abs/2411.19043, 2024. doi:10.48550/arXiv.2411.19043.
- 21 Ravin Ravi, Dylan Bradshaw, Stefano Ruberto, Gunel Jahangirova, and Valerio Terragni. LLMLOOP: Improving LLM-generated code and tests through automated iterative feedback loops. In *Proceedings of the 41st IEEE International Conference on Software Maintenance and Evolution (ICSME), Tool Demonstration Track*. IEEE, 2025. arXiv preprint arXiv:2603.23613. doi:10.1109/ICSME64153.2025.00109.
- 22 Matthew Renze. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356. Association for Computational Linguistics, 2024. doi:10.18653/v1/2024.findings-emnlp.432.

- 23 Amirali Sajadi, Kostadin Damevski, and Preetha Chatterjee. How safe are AI-generated patches? a large-scale study on security risks in LLM and agentic automated program repair on SWE-bench. *arXiv preprint arXiv:2507.02976*, 2025. doi:10.48550/arXiv.2507.02976.
- 24 Jiahao Shi and Tianyi Zhang. RESCUE: Retrieval augmented secure code generation. *arXiv preprint arXiv:2510.18204*, 2025. doi:10.48550/arXiv.2510.18204.
- 25 Shivani Shukla, Himanshu Joshi, and Romilla Syed. Security degradation in iterative AI code generation: A systematic analysis of the paradox. In *Proceedings of the IEEE International Symposium on Technology and Society (IEEE-ISTAS)*. IEEE, 2025. arXiv preprint arXiv:2506.11022. doi:10.1109/ISTAS65609.2025.11269659.
- 26 Vidyut Sriram, Sawan Pandita, Achintya Lakshmanan, Aneesh Shamraj, and Suman Saha. Improving LLM-assisted secure code generation through Retrieval-Augmented-Generation and multi-tool feedback. *arXiv preprint arXiv:2601.00509*, 2026. doi:10.48550/arXiv.2601.00509.
- 27 Hao Tang, Keya Hu, Jin Peng Zhou, Sicheng Zhong, Wei-Long Zheng, Xujie Si, and Kevin Ellis. Code repair with LLMs gives an exploration-exploitation tradeoff. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. arXiv preprint arXiv:2405.17503. doi:10.52202/079017-3746.
- 28 Yiming Xiang, Zhenning Yang, Jingjia Peng, Hermann Bauer, Patrick Tser Jern Kon, Yiming Qiu, and Ang Chen. Automated bug discovery in cloud infrastructure-as-code updates with llm agents. In *2025 IEEE/ACM International Workshop on Cloud Intelligence & AIOps (AIOps)*, pages 20–25. IEEE, 2025. doi:10.1109/AIOps66738.2025.00011.
- 29 Hao Yan, Swapneel Suhas Vaidya, Xiaokuan Zhang, and Ziyu Yao. Guiding AI to fix its own flaws: An empirical study on LLM-driven secure code generation. *arXiv preprint arXiv:2506.23034*, 2025. doi:10.48550/arXiv.2506.23034.
- 30 Tianyi Zhang, Shidong Pan, Zejun Zhang, Zhenchang Xing, and Xiaoyu Sun. Deployability-centric infrastructure-as-code generation: Fail, learn, refine, and succeed through llm-empowered devops simulation. *arXiv preprint arXiv:2506.05623*, 2025. doi:10.48550/arXiv.2506.05623.