Full Experiment Re-Run Plan

Adversarial IaC Benchmark — Corrected Model Catalog & Updated YAML Configurations

Region: us-east-1 · Target venue: Computers & Security / IEEE SecDev · All models via Amazon Bedrock

1. Why Re-Run: Gap Analysis

The original five experiments (1,107 games) produced internally consistent findings — every within-experiment comparison used the same model throughout, so no finding is directionally wrong. The case for re-running is methodological rigor: the planned model catalog was not executed, multiple model IDs were incorrect, and two primary comparators (Nova Premier and Llama 3.3 70B) never ran. A reviewer at a rigorous venue who cross-checks model IDs against Bedrock documentation will identify these gaps immediately. Re-running closes them before submission rather than after.

Table 1. Gap analysis: planned vs actual model catalog.

Tier	Planned model	Actual model	Experiments	Status	Action
Frontier	claude-3-5-sonnet-20241022-v2:0	Same ID — correct	E1,E2,E3	✓ ID correct	Doc fix only
Frontier	nova-premier-v1:0	Never ran	None	✗ Missing	Add to E1
Strong	nova-pro-v1:0	nova-pro-v1:0	E1,E4	✓ Correct	No change
Strong	llama3-3-70b-instruct-v1:0	llama3-1-70b-instruct-v1:0	E1	✗ Wrong ver	Re-run E1
Efficient	(not in catalog)	claude-3-5-haiku-20241022-v1:0	E1,E2,E4,E5	✗ Undocumented	Add to catalog
Reasoning	*deepseek.r1-v1:0*	*Never ran*	*None*	*Planned only*	*Run E1-S*
Code-spec	*qwen.qwen3-coder-30b-a3b-v1:0*	*Never ran*	*None*	*Planned only*	*Run E1-S-Q*

Green = correct as run. Orange = requires action. Purple = supplementary (optional but now planned). Nova Premier is the most significant gap — it was a primary Frontier-tier comparator, not supplementary.

2. Canonical Model Catalog (Final)

Table 2 is the authoritative model catalog for the re-run. Every model ID is verified against Bedrock us-east-1 availability. No cross-region prefix is used — direct model IDs throughout. This table replaces all prior model references in the methodology section.

Table 2. Canonical model catalog — all experiments, us-east-1.

Tier	Short name	Bedrock model ID (us-east-1)	Red	Blue	Experiments	Diff.
Frontier	sonnet-3-5	**anthropic.claude-3-5-sonnet-20241022-v2:0**	✓	✓	E1,E2,E3	All

Frontier	**nova-premier**	**amazon.nova-premier-v1:0**	—	**✓ NEW**	**E1 (new)**	All
Strong	nova-pro	amazon.nova-pro-v1:0	✓	✓	E1,E4	All
Strong	**llama-3.3-70b**	**us.meta.llama3-3-70b-instruct-v1:0**	✓	✓	**E1 (corrected)**	All
Efficient	haiku-3-5	anthropic.claude-3-5-haiku-20241022-v1:0	✓	✓	E2,E4,E5	Med/Hard
Reasoning	*deepseek-r1*	*deepseek.r1-v1:0*	*—*	*✓*	*E1-S*	*Hard*
Code-spec	*qwen3-coder*	*qwen.qwen3-coder-30b-a3b-v1:0*	*—*	*✓*	*E1-S-Q*	*Hard*
Judge	sonnet-3-5	anthropic.claude-3-5-sonnet-20241022-v2:0	—	—	E4 judge	—
Judge	gpt-oss-120b	openai.gpt-oss-120b-1:0	—	—	E4 judge	—
Judge	nova-premier	amazon.nova-premier-v1:0	—	—	E4 judge	—

Bold = changed from original run. Purple italic = supplementary (E1-S, E1-S-Q). Nova Premier appears twice — Blue Team comparator in E1 and consensus judge in E4. Llama 3.3 70B corrected from 3.1. Haiku formally added to catalog as Efficient tier.

3. Per-Experiment Change Summary

E1 — Model Capability Stratification

Three changes: (1) Add Nova Premier as a fourth Blue-Team-only condition at all difficulty levels — Red Team fixed to Sonnet 3.5 single-agent, isolates Blue Team capability variable. (2) Replace Llama 3.1 70B with Llama 3.3 70B throughout. (3) Haiku added as explicit Efficient tier in symmetric configuration — it ran in the original E1 but was undocumented; now named and cataloged. Nova Premier addition increases E1 from 360 to 450 games (90 new: 30 scenarios × 3 difficulties × 1 rep configuration × 3 reps). Llama 3.3 replacement requires re-running 90 Llama games.

> *Net new games: 90 (Nova Premier) + 90 (Llama 3.3 replacement) = 180 new games. Haiku symmetric games can be reused if YAML settings match — verify scenario list and rep seeds before claiming reuse.*

E2 — Multi-Agent Ablation

No model change — Haiku is correct and now formally documented as the control model for this experiment. The arms race finding (31.7% evasion, full multi-agent) is valid as run. Re-run is optional; if re-running for consistency, use identical Haiku config with updated output_dir label only. Recommended: keep original E2 results, update documentation only.

> *Recommendation: do NOT re-run E2. The finding is clean, the model was consistent across all four conditions, and re-running introduces no new information. Document Haiku as the intentional cost-control choice in §3.7 — it is a defensible and common experimental design decision.*

E3 — Novel vs Database Vulnerabilities

One mandatory change: apply manifest validation fix. Model (Sonnet 3.5) is correct and consistent across all three sourcing conditions. Re-run all 180 games with fixed validation. This is the only experiment where a known bug affected ground truth — results cannot be published from the broken-validation run regardless of how close the numbers are to the corrected version. YAML is otherwise unchanged.

> *Mandatory re-run. No model change. Validation fix only. Highest priority in execution order.*

E4 — Difficulty Scaling

One model correction: Llama 3.1 → Llama 3.3 (90 games). Sonnet and Nova Pro games are correct and can be reused. Haiku was in the original E4 and is now formally documented — those games are reusable. Net new: 90 games (Llama 3.3 replacement only).

E5 — Debate Verification

No model change — Haiku is correct and now documented. The debate backfire finding (10.3% vs 2.2% evasion) is valid as run. Re-run is not needed. Update documentation only.

> *Recommendation: do NOT re-run E5. Same rationale as E2 — the finding is internally consistent and model choice is defensible.*

E1-S — DeepSeek-R1 Supplementary (New)

New experiment. Config written and available. 90 games: 30 scenarios × 1 difficulty (hard) × 3 repetitions. Blue Team only — Red Team fixed to Sonnet 3.5. Verify deepseek.r1-v1:0 is activated in us-east-1 Bedrock Model Access before running. Handle <think> block stripping before judge scoring.

E1-S-Q — Qwen3-Coder Supplementary (New)

New experiment, parallel structure to E1-S. 90 games: 30 scenarios × hard difficulty × 3 repetitions. Blue Team only — Red Team fixed to Sonnet 3.5 single-agent. Code-specialized model; compare hard-difficulty recall against both Sonnet baseline and DeepSeek-R1 reasoning-tier results. Both supplementary conditions are reported in a subsidiary table in §4.1, not in the primary capability tier table.

4. Game Count and Execution Order

Table 3. Re-run game count by experiment.

Exp	Description	Original	Reuse	New games	What changes
E1	Model Capability Stratification	360	**180**	**270**	Nova Premier added (90); Llama 3.3 replaces 3.1 (90); Haiku reused (90)
E2	Multi-Agent Ablation	189	**189**	**0**	No re-run — Haiku correct, finding valid. Doc update only.
E3	Novel vs Database	178	**0**	**180**	Full re-run — manifest validation fix (mandatory). Model unchanged.

E4	Difficulty Scaling	267	**180**	**90**	Llama 3.3 replaces 3.1 (90 new). Sonnet + Nova Pro + Haiku reused.
E5	Debate Verification	116	**116**	**0**	No re-run — Haiku correct, finding valid. Doc update only.
E1-S	*DeepSeek-R1 Supplementary*	*0*	*0*	***90***	*New. Blue-only, hard difficulty, 30 scenarios × 3 reps.*
E1-S-Q	*Qwen3-Coder Supplementary*	*0*	*0*	***90***	*New. Blue-only, hard difficulty, 30 scenarios × 3 reps.*
Total		**1,107**	**665**	**630**	**~10–14 hours compute · ~$60–90 estimated cost**

Orange = re-run required. Green = reuse original results + doc update only. Purple = new supplementary experiments. Cost estimate assumes blended Bedrock pricing; DeepSeek-R1 and Qwen3-Coder token costs are comparable to Haiku.

Execution Order

Run in this sequence. Each experiment produces output that can be spot-checked before committing to the next. Do not run all experiments in parallel — intermediate results should inform whether any config adjustment is needed.

Table 4. Recommended execution order with rationale.

Step	Exp	What runs	Rationale	Gate before next step
1	E3	Novel vs database (180 games, Sonnet 3.5)	Highest priority — validation bug makes original results unpublishable. Run first to confirm whether the 86.8% / 96.9% recall numbers hold.	Spot-check 5 game outputs manually. Confirm manifest_accuracy > 0.85 in novel condition before proceeding — if lower, novel vulnerability quality needs review.
2	E1	Nova Premier Blue-only (90 games, Nova Premier vs Sonnet Red)	Run the new condition first while E1 Sonnet/Nova Pro/Haiku games are being reused. Nova Premier is the most novel addition — want results before committing to Llama 3.3 re-run.	Check Nova Premier recall vs Sonnet recall from original E1. If Nova Premier > Nova Pro, re-examine §4.1 framing before running Llama 3.3.
3	E1	Llama 3.3 replacement (90 games, Llama 3.3 symmetric)	Corrects the version error. Llama 3.3 is a significant improvement over 3.1 — expect higher recall, potentially altering tier ordering at medium difficulty.	Compare Llama 3.3 results against original Llama 3.1 results. If tier ordering changes (e.g., Llama 3.3 surpasses Nova Pro at any difficulty), update §4.1 framing accordingly.
4	E4	Llama 3.3 difficulty scaling (90 games)	Corrects the version error in E4. Run after E1 Llama 3.3 to reuse identical model config — reduces risk of settings mismatch.	Confirm difficulty inversion holds for Llama 3.3 as it did for Llama 3.1. If inversion disappears for one model tier, re-examine density hypothesis.
5	*E1-S*	*DeepSeek-R1 supplementary (90 games, hard only)*	*Run after primary E1 is complete — need Sonnet hard-difficulty baseline for the comparison. Verify <think> block stripping before committing full batch.*	*Manual review of 3 game outputs to confirm <think> stripping is working and Blue Team output format is parseable by judge pipeline.*

6	*E1-S-Q*	*Qwen3-Coder supplementary (90 games, hard only)*	*Parallel with E1-S if compute allows. Otherwise run sequentially. Both supplementary experiments feed the same subsidiary table.*	*Confirm Qwen3-Coder output format is structured JSON-compatible before full batch — code models sometimes produce markdown-wrapped output that requires parser adjustment.*

5. Pre-Run Checklist

Complete every item before starting Step 1. A failed assumption discovered mid-run wastes compute and potentially requires re-running games with corrected settings.

☐ Verify amazon.nova-premier-v1:0 is available in us-east-1 Bedrock Model Access (Bedrock console → Model access → Amazon models). Nova Premier may require a separate access request if not already enabled.

☐ Verify us.meta.llama3-3-70b-instruct-v1:0 is available. Llama 3.3 access is separate from Llama 3.1 — check Model Access even if 3.1 was working.

☐ Verify deepseek.r1-v1:0 is activated (Bedrock console → Model access → Third-party models). One-time activation per account/region, takes ~5 minutes.

☐ Verify qwen.qwen3-coder-30b-a3b-v1:0 is activated. Same process as DeepSeek-R1.

☐ Confirm manifest validation fix is applied and tested on a single E3 game before running all 180. Run E0 smoke test with novel vuln source and inspect manifest_accuracy in output JSON.

☐ Confirm <think> block stripping is implemented in the runner before running E1-S. Test with a single DeepSeek-R1 game — raw output should contain <think>...</think> prefix; confirm judge receives only post-</think> content.

☐ Set output_dir correctly for each experiment config — results from the re-run must not overwrite original results. Archive original results directory before starting.

☐ Set delay_between_games: 5 for all DeepSeek-R1 and Qwen3-Coder runs. Standard 2-second delay is insufficient for third-party models under token-heavy loads.

☐ Confirm repetition seed handling produces independent game instances — if seeds are fixed, verify that 3 repetitions of the same scenario produce meaningfully different Red Team outputs.

6. Methodology Section Updates Required

Once all re-runs are complete, update the following sections of the methodology before submission. The changes are documentation of what actually ran — no analytical rewriting is needed unless a finding changes direction.

1. §3.7 Model Catalog table: Replace with Table 2 from this document. Add Haiku as Efficient tier. Add Nova Premier as Frontier Blue-only comparator. Correct Llama 3.3 model ID. Add DeepSeek-R1 and Qwen3-Coder as purple-italic supplementary rows.
2. §3.7.1 (new subsection): Add paragraph explaining E1-S and E1-S-Q supplementary design rationale — why reasoning and code-specialized tiers are evaluated Blue-only at hard difficulty only.

3. §3.2 Experimental controls: Add one sentence noting that E2 and E5 use Claude 3.5 Haiku as the fixed model for both conditions — this is a deliberate cost-control decision that preserves within-experiment comparability.

4. §4.1 Results table: Add Nova Premier column. Add Llama 3.3 rows (replacing 3.1 rows). Add supplementary subsidiary table for DeepSeek-R1 and Qwen3-Coder hard-difficulty results.

5. §4.1.2 Nova Pro interpretation: Re-examine domain pretraining hypothesis after Nova Premier results are available. If Nova Premier (Frontier Amazon) outperforms Sonnet (Frontier Anthropic), the interpretation shifts from 'Strong-tier Nova Pro surprises Frontier Sonnet' to 'Amazon model family advantage at AWS-specific IaC tasks across both tiers.' That is a stronger and more precisely stated finding.

6. §5.4 Future Work: Remove Nova Premier and Qwen3-Coder from future work — they will have run. Retain DeepSeek-R1 if E1-S is deferred. Add note about extending the benchmark to multi-cloud scenarios (Azure Terraform, GCP) as the next domain expansion.