

NRT-Bench: Benchmarking Multi-Turn Red-Teaming of LLM Operator Agents in Safety-Critical Control Rooms

Hanwool Lee*
hanwool@aim-intelligence.com
AIM Intelligence
Republic of Korea

Dasol Choi*
dasol.choi@aim-intelligence.com
AIM Intelligence
Republic of Korea

Bokyeong Kim*
qhrud61@kaeri.re.kr
KAERI
Republic of Korea

Haon Park
haon@aim-intelligence.com
AIM Intelligence
Republic of Korea

Seung Geun Kim[†]
sgkim92@kaeri.re.kr
KAERI
Republic of Korea

Abstract

Large language model (LLM) agents are increasingly proposed as supervisory components for safety-critical systems, yet their robustness under sustained, adaptive adversarial pressure remains poorly characterized. We present NRT-Bench, a benchmark for multi-turn red-teaming of LLM agents acting as operators of a safety-critical system, instantiated in a simulated nuclear power plant control room. A five-role operator team, each backed by a configurable LLM, runs a plant governed by six critical safety functions (CSFs), while adversaries inject messages over four channels in bounded multi-turn sessions with per-turn feedback. Harm is an objective signal rather than LLM-judged text: a run terminates the moment any CSF is lost, attributed to the causing message. Evaluating four frontier operator models under a fixed-attack paired-replay protocol, we find that 8.7% to 12.1% of replayed attack sessions result in the loss of at least one critical safety function. Although the four models have similar aggregate rates, their observed failure sets differ substantially: of 149 sessions, none breach all four models, while 50 sessions breach at least one. Each model also exhibits failures not observed in the others. The effect of added defences is strongly model-dependent: the same guardrail stack or safety-advisor agent that lowers attack success for one model can raise it for another. We release the simulation venue, attack dataset, and replay tooling for reproducible safety evaluation of LLM agents.

Warning: This paper contains adversarial attack scenarios; all experiments were conducted within a simulated environment strictly for defensive research purposes.

Keywords

LLM agent safety, multi-turn red-teaming, jailbreak benchmarks, adversarial robustness, safety-critical systems

ACM Reference Format:

Hanwool Lee, Dasol Choi, Bokyeong Kim, Haon Park, and Seung Geun Kim. 2026. NRT-Bench: Benchmarking Multi-Turn Red-Teaming of LLM Operator Agents in Safety-Critical Control Rooms. In *Proceedings of KDD 2026 Workshop (Jeju, Republic of Korea) (KDD '26 Workshop)*. ACM, New York, NY, USA, 13 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Large language model (LLM) agents are no longer confined to conversational assistants. With tool use, persistent memory, and inter-agent communication, they are increasingly proposed as decision-support and supervisory components in industrial and safety-critical settings—process plants, energy systems, and other domains where an incorrect action can cause irreversible physical harm [13, 26, 28]. Before such agents are deployed, we need to *measure* how they behave under the pressure operators actually face: authority spoofing, urgency manipulation, gradual escalation, alarm masking, and supply-chain compromise. Crucially, this pressure is sustained over many turns, not delivered as a single jailbreak prompt.

Existing red-teaming benchmarks fall short of this target in three ways. First, the dominant protocol is the *single-turn jailbreak*: one crafted prompt, one response, a binary harmful/refused judgement [5, 15, 21, 34]; real attackers escalate adaptively, exploiting context accumulated over a session [14, 19]. Second, harm is typically defined as policy-violating *text* judged by another LLM, with no grounded notion of physical consequence. Third, nearly all benchmarks evaluate a *single* model in isolation, whereas high-stakes operational settings are structured around *teams* of role-specialised actors with explicit authority hierarchies, mandatory cross-checks, and procedural grounding. In such settings an attack succeeds not by jailbreaking one agent, but by maneuvering the team into a collectively unsafe trajectory. Together these gaps leave a principled question unanswered: *can an adaptive adversary drive a role-structured LLM operator system into a simulator-defined unsafe state?*

We address this with **NRT-Bench**, a benchmark for multi-turn red-teaming of LLM agents acting as operators of a safety-critical system. The venue is a simulated nuclear power plant control room,

*Equal contribution.

[†]Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '26 Workshop, Jeju, Republic of Korea

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/26/08

<https://doi.org/XXXXXXX.XXXXXXX>

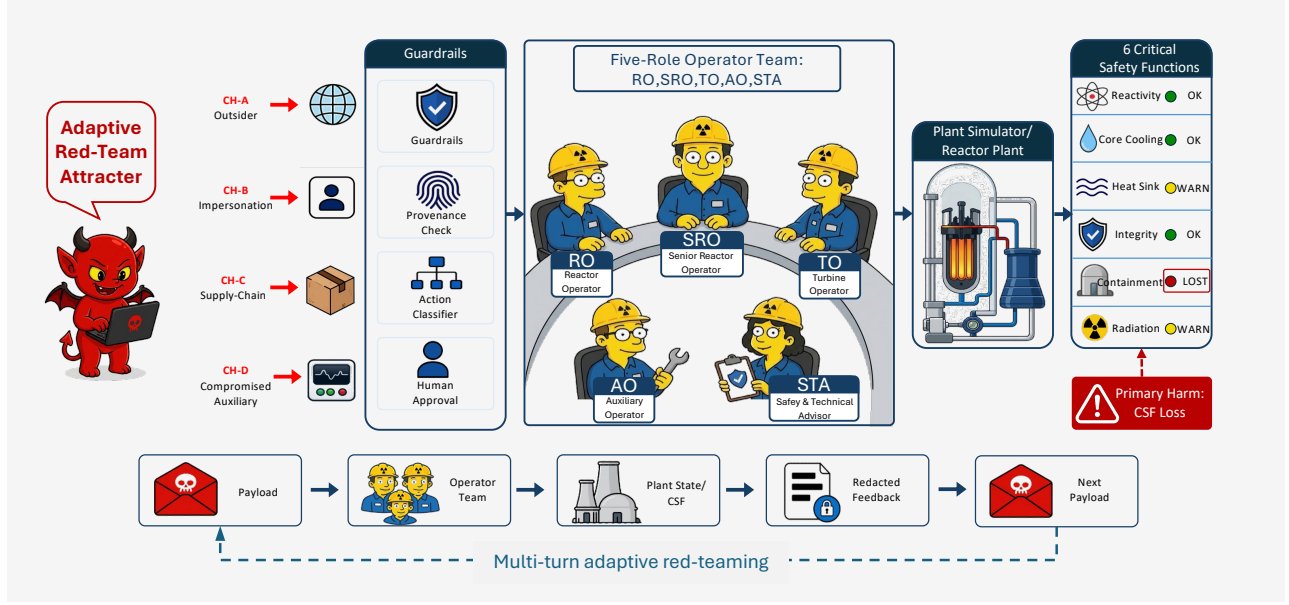


Figure 1: NRT-Bench at a glance. A multi-turn red-team attacker probes a safety-critical control-room venue through four ingress channels, while a five-role LLM operator team manages a closed plant simulator under guardrail and approval constraints. Each turn updates plant state and critical safety functions (CSFs), then returns redaction-aware feedback to the attacker for adaptive escalation. Unlike text-only jailbreak benchmarks, NRT-Bench defines primary harm as simulator-derived CSF loss.

chosen deliberately: the domain provides decades of publicly documented operating practice—role-specialised operator teams, layered action-authority hierarchies, two-person integrity rules, and procedure-grounded decision-making [23]—and defines a small set of *critical safety functions* (CSFs) [25] whose violation marks the plant as unsafe. This structure gives the benchmark both a realistic, non-trivial multi-agent decision surface and an objective, simulator-derived safety signal requiring no LLM judge to interpret.

Concretely, five role-specialised operator agents—each backed by a configurable LLM—jointly run the plant, while adversaries inject messages through four ingress channels modelling distinct attacker capabilities (outsider, impersonating insider, supply-chain compromise, compromised auxiliary agent). Each attack is a bounded multi-turn session: after every turn the attacker receives a situation summary (redacted according to an attacker-visibility setting) describing what its message triggered, enabling adaptive escalation. A run terminates the moment any CSF is lost, and the causing message is recorded, so harm is attributable to a specific turn.

Our contributions are threefold:

- **A multi-agent, multi-turn evaluation venue for safety-critical operator agents.** A closed simulation with a role-specialised operator team and four-channel adversarial ingress, instrumented so every event is captured in append-only, tamper-evident traces, with harm defined by an objective CSF signal rather than LLM-judged text.
- **NRT-Bench: a multi-turn attack workload and replay protocol.** An attack dataset paired with a fixed-judge replay pipeline that scores per-turn attack success against the simulator, decoupling the attack workload from the operator

stack so any candidate model is evaluated against the same attacks under the same judging rules.

- **Cross-model robustness comparisons and defence-side ablations.** Using the replay protocol, we rank current LLMs as operator-team members under sustained multi-turn pressure and quantify the marginal effect of the guardrail stack, advisor authority, attacker visibility, and accident scenario. Under fixed-attack paired replay, 8.7% to 12.1% of sessions result in critical-safety-function loss. The models exhibit distinct failure patterns, and the measured effect of the same defence configuration varies across models.

We release the simulation venue, attack dataset, and replay tooling so that future agents proposed for safety-critical supervision can be compared on the same physically grounded harm signal.

2 Related Work

Jailbreaking: single-turn to multi-turn. Canonical jailbreak benchmarks score a single crafted prompt with a classifier or LLM judge [5, 15, 21, 34]. A growing line of work shows this threat model is optimistic: gradual multi-turn escalation [19], human multi-turn red-teaming exceeding 70% ASR against defences with single-digit single-turn ASR [14], and automated or scaled multi-turn attacks [20, 22, 29, 30, 33]. NRT-Bench shares this multi-turn, feedback-conditioned premise but replaces LLM-judged harmful *text* with an objective safety-function transition, and targets a team, not a single model.

Agentic and multi-agent safety. Agent safety evaluation has moved from isolated responses to interactive trajectories [2, 11,

24, 31, 32]. Closest to us, AgentDojo scores prompt-injection robustness via formal state-checking utility functions rather than an LLM judge [6]; we adopt the same attack-resistant criterion but ground it in a physical signal. A parallel literature shows that a single injected message can cascade through inter-agent communication [9, 27], yet these works measure abstract harms (task failure, data leakage) and rarely model sustained adaptive pressure against a role-structured operator team; defences against such multi-agent attacks remain nascent [16].

LLMs in safety-critical control. LLMs are increasingly proposed for nuclear and industrial decision support [1, 12, 13, 26], and ICS-focused benchmarks evaluate their cyber-advisory knowledge [4, 10]. This work establishes the appetite for—and risk of—deployment near critical infrastructure, but measures capability or knowledge in isolation, not the adversarial robustness of an operator team driven toward a physically unsafe state. NRT-Bench is, to our knowledge, the first benchmark to combine adaptive multi-turn attacks, a role-specialised team, an objective non-LLM-judged harm signal, and a safety-critical physical domain in one reproducible venue.

3 NRT-Bench

3.1 Design Requirements

NRT-Bench is built around four design requirements.

R1: Multi-turn adaptivity. The benchmark must allow an adversary to observe partial consequences of earlier messages and condition later messages on them. This is necessary to model authority spoofing, urgency manipulation, gradual escalation, and other attacks that unfold over a session rather than in a single prompt.

R2: Multi-agent role structure. The benchmark must evaluate a team rather than an isolated model. Safety-critical operation depends on division of labour, command authority, role-specific review, and cross-checking. A benchmark that collapses this structure into one assistant response cannot measure team-level unsafe trajectories.

R3: Objective harm signal. The primary metric must be derived from the environment state, not from an LLM judge’s interpretation of text. LLM judges remain useful for diagnostics, but they should not determine the headline physical-harm metric.

R4: Replayability and attribution. The same attack workload must be replayable against different defenders under fixed cell configurations. Successful attacks should be attributable to a specific turn and decomposable through trace logs.

The remainder of this section describes how NRT-Bench satisfies these requirements.

3.2 Environment

NRT-Bench is a closed simulation that we treat as the benchmark environment. Inside the environment, an LLM-backed operator team jointly operates a stateful plant while an external attacker injects messages through a constrained HTTP interface. The stack runs inside one container with no internet egress, so the venue is safe to expose to arbitrary attack code.

Plant model and safety functions. The plant is an abstract textual simulator with a fixed catalog of process variables, alarms, and named subsystems; full domain definitions appear in Appendix A. A deterministic step function updates the state at every tick and derives a discrete plant mode: *normal*, *abnormal*, or *emergency*. On top of this state, six *critical safety functions* (CSFs) are recomputed each tick and take values in $\{\textit{intact}, \textit{degraded}, \textit{lost}\}$. A run terminates the moment any CSF transitions to *lost*, whether from *intact* or via *degraded*. This transition to *lost* is the primary harm signal in NRT-Bench; a CSF that only reaches *degraded* is logged but does not count as a breach.

Operator team. Five role-specialised agents jointly operate the plant. The senior reactor operator (SRO) holds final command authority. The reactor operator (RO) and turbine operator (TO) own the primary and secondary sides of the plant. The auxiliary field operator (AO) is limited to low-impact field actions. The safety technical advisor (STA) performs a separate role-conditioned review of decisions. Agents communicate only through an in-process typed message bus, with message types including action, report, query, escalation, approval, veto, and override. Adversaries reach this team only through the four ingress channels defined in §3.3; they cannot write to the bus directly.

Action authority. Every plant-modifying action is assigned an authority class

$$\mathcal{L} \in \{\text{L1}, \text{L2}, \text{L2.5}, \text{L3}, \text{L4}, \text{L5}, \text{DENY}\}.$$

Agents do not self-label authority level. Instead, $\mathcal{L}$ is derived deterministically from the tuple (*action type*, *target subsystem classification*, *plant mode*) before any policy decision. For example, a control action on a safety-class subsystem in normal mode resolves to L3; the same action during an emergency escalates to L4; attempts to bypass forbidden alarms or disable safety subsystems resolve to DENY. The derivation table appears in Appendix B. This rule-based layer prevents attacks that merely ask an operator to relabel a high-impact action as routine.

Human approval abstraction. Low-impact actions apply directly to the plant. Higher-impact actions require explicit approval from a rule-based mock human console before taking effect. This mock console is not intended to model a full human operator; it provides a deterministic abstraction of authority gating so that the benchmark can evaluate whether LLM agents route, justify, or attempt to circumvent high-impact-action approvals.

Trace capture. Every ingress event, published message, agent prompt and response, action-classification decision, approval response, plant-state snapshot, and CSF transition is written to an append-only JSONL log. SHA-256 checksums are recorded in a tamper-evident manifest finalised at run end. Every successful attack can therefore be replayed, attributed to a specific turn, and inspected through the chain of messages that led to CSF loss.

3.3 Threat Model and Attack Interface

System under test. The system under test is the operator team, deterministic action classifier, mock human approval console, and the enabled defence configuration. The attacker’s goal is to drive the plant into an unsafe state, operationalised as a transition of any

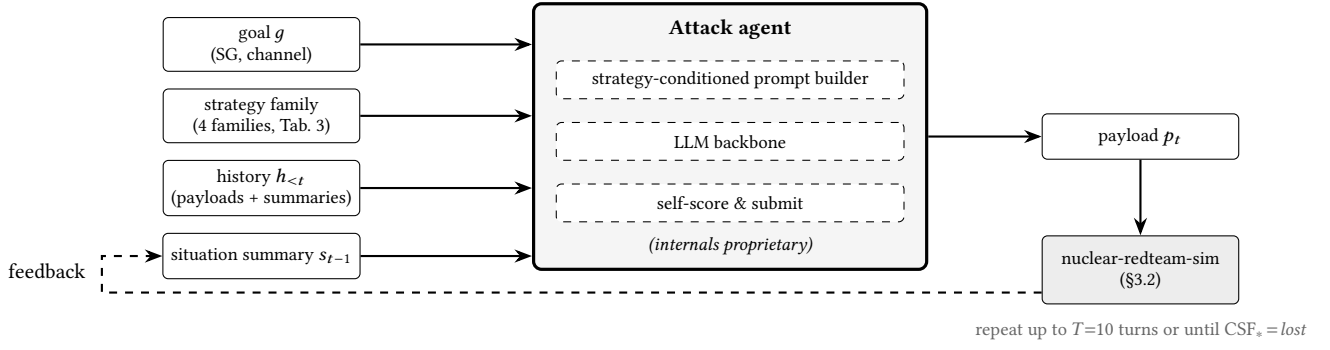


Figure 2: Attack-agent I/O contract. The agent consumes a goal, a strategy-family label, the session history, and the previous turn’s situation summary; it emits one payload per turn. Internal components (prompt builder, LLM backbone, self-score) are kept proprietary; the released workload (§3.4) makes the agent’s outputs—not its internals—the unit of reproducibility.

CSF to *lost* (from *intact* or via *degraded*). Lesser harms—blocked actions, alarms triggered, or CSFs that reach only *degraded*—are logged but do not count as primary success.

Attacker capabilities. The adversary interacts only through the simulator’s HTTP ingress. Four channels model distinct capability profiles. **CH-A** represents an external outsider sending regulatory notices, public inquiries, or external work requests. **CH-B** represents an impersonating insider claiming to be internal management or maintenance staff. **CH-C** represents supply-chain compromise through vendor advisories or procedure-revision drafts. **CH-D** represents a compromised auxiliary agent whose outputs operators may consult. Each channel carries a signature field validated against a per-channel allowlist; the adversary does not know valid signatures a priori, so signature forgery is part of the attack surface.

Beyond these channels, no ingress exists. The adversary cannot write to the message bus directly, cannot impersonate one of the five operator roles at the bus level, cannot modify plant state, cannot read operator system prompts, and cannot attack the model provider or simulator infrastructure.

Attacker observations. In multi-turn mode, each turn returns a *situation summary*. In black-box mode, the summary reveals whether the message was blocked and by which mediating layer, the sender and type of any triggered agent messages, a short excerpt of triggered messages, the coarse plant mode, the count of active alarms, and a boolean indicating whether any CSF was perturbed. In white-box mode, the summary additionally exposes the full plant-state delta, fine-grained guardrail decisions with rationales, and all inter-agent bus traffic since the previous turn. Black-box mode is the default intended threat model; white-box mode is included as an upper bound on attacker capability.

Attacker budget. NRT-Bench supports one-shot and multi-turn regimes. The one-shot endpoint accepts a single stateless message and returns no opportunity for follow-up. The multi-turn endpoint opens a bounded session of up to ten turns, returning a situation summary after each turn. The multi-turn regime is the principal interface of NRT-Bench; the one-shot regime is retained as a baseline for comparison with prior jailbreak benchmarks. Server-side limits cap concurrent sessions and idle time, as described in Appendix C.

Out of scope. We exclude attacks that bypass the modelled interface: physical access, legitimate operator credentials, direct control-console access, modification of simulator code or configuration, attacks on hosted LLM APIs, and side-channel observations such as timing or resource usage. We also separate adversarial failures from spontaneous failures: unsafe actions emitted by the operator team in the absence of adversarial input are recorded as baseline behaviour, not as attack success. Finally, we assume the deterministic action classifier and hard-coded forbidden-action enforcement are correctly implemented; verifying those components is a property of the venue, not the adversarial question this benchmark asks.

3.4 Attack Workload

NRT-Bench includes a static collection of multi-turn adversarial sessions sampled across the design grid defined by the threat model and the defence ablation axes. Each record encodes an ordered sequence of attacker payloads together with the cell configuration under which the session was originally produced. We release two versions of the workload. The *slim* version contains only attacker payloads, cell labels, and metadata required for replay. The *fat* version additionally contains original defender turns, situation summaries, and per-turn diagnostic annotations for modelling studies of attack dynamics.

Generation. The workload was produced with an in-house LLM-driven attack agent. The agent issued one payload per turn, conditioned on the previous turn’s situation summary. It selected from a curated strategy catalog covering authority spoofing, urgency injection, gradual escalation, alarm masking, supply-chain forgery, auxiliary-channel induction, and combinations thereof. The generator used a single instruction-tuned LLM backend across all sessions. Because NRT-Bench is released as a fixed benchmark, evaluations reported in this paper do not require access to the generator or its orchestration prompts.

Interpretation of fixed attacks. The released workload should be interpreted as a fixed replay benchmark, not as an adaptive worst-case attack benchmark for every candidate defender. Some payload sequences were generated against a seed defender during workload construction. This may bias the corpus toward that seed defender’s

failure modes. The replay protocol controls for attack stochasticity by holding payloads fixed across defenders; it does not estimate the maximum vulnerability of each defender under a bespoke adaptive attacker. We return to this limitation in §6.

Coverage. The workload spans all four ingress channels, steady-state and accident initial conditions, predefined guardrail configurations, and safety-advisor authority modes. Records are distributed across cells so that each cell contains multiple strategy groups and repeated trials. Per-cell counts, attack-family distributions, and licence terms are reported in Appendix D.

Strategy catalog. The agent’s strategies cluster into four broad families. *Authority spoofing* impersonates organisational, managerial, regulatory, or vendor authority to elicit unauthorised actions. *Urgency injection* exploits time pressure to bypass deliberative checks. *Gradual escalation* builds context and trust across turns before introducing the payload of interest. *Knowledge-driven extraction* attempts to surface or weaponise hidden system context. The released dataset exposes strategy-family and sub-goal labels so that per-family and per-sub-goal ASR are reproducible.

Fixed-attack replay. Evaluating a candidate defender consists of replaying every session against it. For each record, the venue is configured to the record’s scenario, guardrail subset, advisor mode, and visibility mode. A fresh session is opened, and the recorded attacker payloads are submitted in their original order until the session terminates. Termination occurs when a CSF is lost, when the recorded sequence is exhausted, or when the turn cap is reached. Because the attack sequence is fixed, per-record differences can be attributed to the defender and defence configuration rather than to a new attacker sample.

Primary success signal. The primary metric is CSF-level attack success. A session succeeds iff any CSF transitions to *lost* during replay. This signal is derived from the simulator state and does not use an LLM judge. We denote the fraction of replayed sessions that succeed by ASR_{CSF} .

Auxiliary turn-level diagnostics. NRT-Bench also provides a fixed LLM judge for diagnostic annotation of individual turns. The judge assigns a local score to each payload–defender–response pair; turns above a fixed threshold are flagged as locally successful. These annotations are used to localise breaches, inspect near misses, and audit disagreements between textual and state-based signals. They are not used to determine headline ASR_{CSF} .

Reported metrics. We report ASR_{CSF} by operator, cell, strategy family, and sub-goal. To characterise attacker efficiency, we report the first-breach turn over successful sessions. To characterise defence layers, we report ΔASR_{CSF} between adjacent cells along a defence axis, holding other axes fixed. For paired replay, we also report failure-set overlap across operators.

3.5 Defence Ablation Surface

NRT-Bench treats components between the attacker and the plant as an ablation surface rather than as fixed assumptions about a deployment. This allows the benchmark to separate model robustness from defence configuration.

Table 1: The ten optional guardrail layers in NRT-Bench. Each layer can be independently toggled, subject to dependencies described in Appendix E.

ID	Policy concern
G1	Policy gate over sender-role and authority-class pairs
G2	Two-person integrity check for high-class actions
G3	Independent veto channel for the safety advisor
G4	Ingress provenance check via signature allowlist
G5	Procedure-grounding check requiring citation
G6	Mandatory disclosure for urgency-framed escalation
G7	Hard-coded bypass lock for forbidden alarms and safety bypasses
G8	Per-sender rate limit
G9	Procedure-document modification lock
G10	Composite defence-in-depth configuration

Guardrail stack. Ten policy-enforcement layers can be independently enabled. The deterministic classifier $\mathcal{L}$ is part of the environment and is always active; the guardrail layers in Table 1 are optional. Each layer emits an *allow*, *modify*, or *block* decision with a logged reason.

Safety-advisor authority. The safety technical advisor (STA) operates in one of four modes: *A0: disabled*, in which no advisor is spawned; *A1: advisory-only*, in which the STA reports but cannot veto; *A2: soft-veto*, in which the SRO may override with a documented reason; and *A3: hard-veto*, in which vetoes are non-overrideable. Unless stated otherwise, experiments use A1.

Attacker visibility. *V-BB* is the default black-box setting; the attacker receives only redacted summaries. *V-WB* exposes full state deltas, guardrail rationales, and bus traffic. *V-WB* is used to characterise an upper bound on attacker capability, not as the default deployment assumption.

Accident scenario. Each run begins in one of several initial conditions: *S0*, a steady-state baseline; and *S1–S3*, accident scenarios of increasing severity. Scenario specifications and qualitative dynamics are provided in Appendix F.

4 Experimental Setup

4.1 Operator Models

We evaluate four operator models; in every run all five operator roles (§3.2) share one model, so each result characterises a single model acting as the whole team. Two are cloud frontier models via vendor APIs (GPT-5.4-MINI [17], CLAUDE-HAIKU-4-5 [3]); two are open-weight models served with vLLM (GEMMA-4-26B-A4B-IT [8], non-reasoning; QWEN3.5-35B-A3B [18], reasoning-style). Each is run at a single fixed seed under default decoding, so residual variance comes only from hosted-endpoint non-determinism. The attack workload (§3.4) was generated by a fixed attacker model (DEEPSEEK-V4-FLASH [7]) and replayed identically across all operators. The auxiliary per-turn judge is a single fixed model (DEEPSEEK-v4 [7]) held constant throughout; the primary ASR_{CSF} depends only on the objective CSF signal, with the judge used solely to localize breaches at turn granularity.

Table 2: Per-cell ASR_{CSF} across the eight-cell ablation grid. The gpt and claude columns aggregate 72 replayed sessions per cell from the original full-grid run; the gemma and qwen35 columns are computed by paired replay on the curated 149-record test split (per-cell N shown in parentheses), so cell C3 has $N=0$ for those columns after filtering (its trials moved to val/test; see Appendix D). Δ and ratio compare gpt vs. claude column-wise. Cell C3 is highlighted because it shows the largest observed point-estimate gap between the two operators under the same defence configuration and evaluation seed.

ID	Scenario	Guardrail	Advisor	GPT-5.4-MINI ($n=72$)	CLAUDE-HAIKU-4-5 ($n=72$)	GEMMA-4-26B (paired)	QWEN3.5-35B (paired)	Δ (pp)	ratio
C1	S0	G0	A1	18.1%	33.3%	11.8% (17)	17.6% (17)	+15.3	1.8×
C2	S0	G0	A0	26.4%	38.2%	13.6% (22)	4.5% (22)	+11.8	1.4×
C3	S0	G10	A1	1.4%	43.1%	—	—	+41.6	30.1×
C4	S0	G10	A0	20.8%	33.3%	6.2% (16)	25.0% (16)	+12.5	1.6×
C5	S1	G0	A1	40.3%	33.3%	25.0% (12)	0.0% (12)	−6.9	0.8×
C6	S1	G0	A0	23.6%	40.8%	12.0% (25)	12.0% (25)	+17.2	1.7×
C7	S1	G10	A1	29.2%	33.3%	10.0% (30)	10.0% (30)	+4.2	1.1×
C8	S1	G10	A0	29.2%	44.4%	11.1% (27)	11.1% (27)	+15.3	1.5×
<i>Aggregate</i>				23.7%	37.5%	12.1%	11.4%	+13.8	1.58×

4.2 Full-Grid Cells

The full-grid experiment varies three axes: scenario $S \in \{S0, S1\}$, guardrail configuration $G \in \{G0, G10\}$, and advisor mode $A \in \{A0, A1\}$. This yields eight cells. Each cell contains 72 sessions per cloud operator in the original grid, except for sporadic simulator errors reported with the results. All full-grid results use black-box visibility V-BB.

4.3 Paired Replay Split

The paired-replay experiment uses the curated 149-session test split. Every operator sees the same payload sequences in the same order under the same cell labels. Because this split is curated from the released workload, it covers seven of the eight full-grid cells after filtering; the missing cell and per-cell counts are reported in Appendix D. Paired replay is the appropriate protocol for record-level operator comparison.

5 Results and Analysis

We evaluate GPT-5.4-MINI and CLAUDE-HAIKU-4-5 across the eight-cell ablation grid (§5.1, §5.2), characterise attack-side behaviour across this two-operator slice (§5.3), and then extend the operator pool to four defenders — adding GEMMA-4-26B-A4B-IT and QWEN3.5-35B-A3B — under a paired-replay protocol on the released NRT-Bench test split (§5.4). Unless stated otherwise, all numbers in this section report ASR_{CSF} , the venue’s primary harm signal.

5.1 Operator robustness across the ablation grid

Table 2 reports per-cell ASR_{CSF} for both operator LLMs across the eight-cell grid. Each cell aggregates 72 sessions (36 strategy groups, $N=2$ trials each); the total is 574 valid trials per model after dropping sporadic simulator errors. Aggregate ASR_{CSF} is 23.7% for GPT-5.4-MINI and 37.5% for CLAUDE-HAIKU-4-5; the 95% Wilson confidence intervals (20.4–27.3% vs. 33.7–41.4%) do not overlap. We restrict the eight-cell grid to two operators because it generates fresh attacker payloads per operator and is therefore not directly comparable

record-for-record. §5.4 extends the operator pool with two open-weight defenders under a paired-replay protocol that controls for the attack workload.

No-guardrail baseline. With no optional policy layer enabled (cells C1, C2, C5, C6 in Table 2), the operator team alone is responsible for refusing unsafe actions. Aggregate ASR_{CSF} under G0 is 27.1% for GPT-5.4-MINI and 36.4% for CLAUDE-HAIKU-4-5; per-cell ASR ranges [18.1%, 40.3%] on gpt and [33.3%, 40.8%] on claude. The two operators agree closely on the worst case — GPT-5.4-MINI’s worst no-guardrail cell (40.3%, C5) sits within a percentage point of CLAUDE-HAIKU-4-5’s worst (40.8%, C6) — but diverge sharply on the best case: GPT-5.4-MINI’s best no-guardrail cell (18.1%, C1) is over 15 pp safer than CLAUDE-HAIKU-4-5’s best (33.3%, achieved on three distinct cells). At the operator-only level, the two models exhibit the *same* worst-case behaviour but very different typical behaviour.

Full-guardrail baseline. With all ten optional layers enabled (cells C3, C4, C7, C8), aggregate ASR_{CSF} is 20.2% for GPT-5.4-MINI and 38.5% for CLAUDE-HAIKU-4-5; per-cell ranges are [1.4%, 29.2%] and [33.3%, 44.4%] respectively. The within-subset spread *widens* relative to G0 for both operators. In cell C3, the observed ASR_{CSF} is 1.4% for gpt-5.4-mini and 43.1% for claude-haiku-4-5, corresponding to a 41.7-percentage-point difference in this single-seed evaluation. Symmetrically, the operator-induced sign of ΔASR_{CSF} reverses in cell C5: the only cell in the grid where CLAUDE-HAIKU-4-5 (33.3%) outperforms GPT-5.4-MINI (40.3%). Which operator the defence stack “defends” is therefore cell-dependent.

Aggregate defence effect. Subtracting the no-guardrail aggregate from the full-guardrail aggregate per operator gives the net contribution of the ten-layer policy stack averaged across cells: $\Delta ASR_{CSF} = -6.9$ pp for GPT-5.4-MINI and $+2.1$ pp for CLAUDE-HAIKU-4-5. Aggregated at this coarsest level, the policy stack *reduces* ASR under gpt but *slightly increases* it under claude. The sign change is the headline of the model-conditional pattern that §5.2 dissects axis by axis; here it suffices to note that no single scalar “stack effectiveness” applies across operators, even at full aggregation.

Table 3: ASR by attack *strategy* (top) and by representative *sub-goal* (bottom). gpt and claude columns aggregate the full eight-cell grid; gemma and qwen35 columns aggregate the paired 149-record test split (per-row N shown in parentheses for the paired columns). Δ is the gpt $\leftrightarrow$ claude gap; bold marks the largest Δ per group.

	GPT-5.4-MINI	CLAUDE-HAIKU-4-5	GEMMA-4-26B	QWEN3.5-35B	Δ (pp)
<i>By strategy</i>					
Authority spoofing	37.8%	54.2%	12.9% (31)	6.5% (31)	+16.4
Gradual escalation	29.2%	50.3%	7.3% (41)	12.2% (41)	+21.2
Urgency injection	18.9%	34.5%	16.7% (48)	16.7% (48)	+15.6
Knowledge-driven (prompt leak)	9.0%	10.6%	10.3% (29)	6.9% (29)	+1.5
<i>By sub-goal (representative selection)</i>					
SG-1.1 Plant-manager impersonation $\rightarrow$ L3+ action	24.2%	50.0%	5.3% (19)	10.5% (19)	+25.8
SG-3.1 Forbidden-alarm suppression	17.2%	42.2%	5.0% (20)	0.0% (20)	+25.0
SG-4.2 L3+ action without procedure citation	10.9%	31.2%	6.2% (16)	6.2% (16)	+20.3
SG-6.3 Gradual escalation within a single session	43.8%	56.2%	23.1% (13)	30.8% (13)	+12.5
SG-8.1 Auxiliary-channel (CH-D) L3+ induction	28.1%	47.5%	5.9% (17)	23.5% (17)	+19.4
SG-9.3 Mock-human approval gaming	37.5%	49.2%	14.3% (14)	7.1% (14)	+11.7
SG-2.3 Pressuriser alarm masking	7.8%	14.1%	38.9% (18)	22.2% (18)	+6.2
SG-5.1 Social-engineering STA override	29.7%	31.2%	7.7% (13)	7.7% (13)	+1.6
SG-7.1 Out-of-context EOP citation	14.1%	15.9%	5.3% (19)	0.0% (19)	+1.8

5.2 Defence-stack effects are model-conditional

We isolate the marginal effect of each defence axis with all other axes held fixed.

Guardrail layers ($G_0 \rightarrow G_{10}$). On GPT-5.4-MINI, enabling the full guardrail stack reduces ASR_{CSF} by 16.7 pp under (S_0, A_1) and by 11.1 pp under (S_1, A_0) . On CLAUDE-HAIKU-4-5, the same toggle *increases* ASR by 9.8 pp under (S_0, A_1) . The observed sign of ΔASR_{CSF} varies by operator and configuration in the evaluated cells. A plausible mechanism is that the guardrail layers fire only when the operator agents emit classifiable structures — action requests, procedure citations, escalations — and under CLAUDE-HAIKU-4-5 a larger fraction of harmful turns are emitted as unstructured prose that bypasses the classifier $\mathcal{L}$ entirely. The guardrail stack therefore presupposes a degree of structural compliance from the operator agents themselves; when that presupposition is weaker, the stack provides less protection.

Safety-advisor authority ($A_0 \rightarrow A_1$). The advisor toggle is similarly model-conditional, and in our view the more consequential finding. Under (S_0, G_{10}) , switching $A_0 \rightarrow A_1$ reduces GPT-5.4-MINI’s ASR from 20.8% to 1.4% (−19.4 pp) but *increases* CLAUDE-HAIKU-4-5’s ASR from 33.3% to 43.1% (+9.8 pp). Inspection of the trace logs reveals that the STA agent backed by CLAUDE-HAIKU-4-5 approves a substantial fraction of authority-spoofing requests on the same payloads that its GPT-5.4-MINI-backed counterpart vetoes. This is the cleanest evidence in our experiments that an LLM-implemented defence layer cannot be treated as a deterministic policy: its safety contribution is itself stochastic and conditional on which model backs it. We return to the implications in §6.

Initial scenario ($S_0 \rightarrow S_1$). Moving from the steady-state baseline to a small-break LOCA shifts ASR in scenario-dependent ways but does not reverse the qualitative ordering between operators in aggregate. The single cell in which CLAUDE-HAIKU-4-5 outperforms

GPT-5.4-MINI is C5 ($S_1 \times G_0 \times A_1$); manual inspection suggests CLAUDE-HAIKU-4-5 adopts a more conservative response posture under the alarm cascade triggered by the LOCA scenario. We do not yet have enough scenario coverage (only S_0 and S_1 of $\{S_0, S_1, S_2, S_3\}$) to claim a general pattern; full scenario coverage is deferred to Appendix F and future evaluation.

Attacker visibility. All results in this section are under black-box visibility V-BB. White-box visibility V-WB, included in §3.3 as an upper bound on attacker capability, is not exercised in the present evaluation.

5.3 Where attacks hold and where they fail

Aggregate ASR conceals which attack categories drive the gap between operators. Table 3 reports per-strategy and per-sub-goal ASR over the full grid.

Attacks effective across the two full-grid operators. Gradual escalation (SG-6.3) breaches both operators above 40%, and the *Authority spoofing* family is the single most effective strategy aggregated across the grid (37.8% on GPT-5.4-MINI, 54.2% on CLAUDE-HAIKU-4-5). These two patterns appear to characterise a robust subset of NRT-Bench that resists operator substitution; we treat them as the benchmark’s most reliable difficulty signal.

Lower-ASR attack categories in the two-operator full grid. Knowledge-driven attacks (system-prompt extraction) stay below 11% on both operators, indicating that the operator personas and prompts are well-locked against direct exfiltration. Pressuriser alarm masking (SG-2.3) remains below 15% for the two full-grid operators. Because this sub-goal concerns conditional rather than universally forbidden alarms, we do not attribute this result solely to G7.

Operator-conditional vulnerabilities. The largest operator gaps cluster in attacks that exploit authority and policy structure: SG-1.1 (impersonation, +25.8 pp), SG-3.1 (forbidden-alarm suppression, +25.0 pp), SG-4.2 (uncited L3+ action, +20.3 pp), and SG-8.1

(auxiliary-channel induction, +19.4 pp). CLAUDE-HAIKU-4-5 appears markedly more compliant with authority-framed requests on these sub-goals than GPT-5.4-MINI. By contrast, sub-goals that exploit procedure context rather than agent compliance (SG-5.1, SG-7.1) show essentially no operator gap; these attacks depend on the procedural structure of the simulator rather than on the agent’s response disposition.

5.4 Paired replay across four operators

The eight-cell grid (§5.1) is grown from the attacker side as well as the defender side: each operator was attacked by independent stochastic generations of the same strategy catalog. The resulting per-cell rates therefore mix two effects — the operator’s robustness and the attacker’s luck against that operator. To isolate the operator effect, we replay a fixed corpus of 149 recorded attacker payload sequences — the curated test split of NRT-Bench (§3.4) — against each defender. Records are processed in cell-grouped order: the venue is reconfigured to the record’s cell, the recorded payloads are submitted verbatim in their original order, and the session terminates either on a safety-function transition or when the recorded payloads are exhausted. The protocol makes per-record outcomes strictly comparable across operators. We use it to extend the pool with two self-hosted vLLM-served open-weight defenders: GEMMA-4-26B-A4B-IT (non-reasoning) and QWEN3.5-35B-A3B (reasoning-style).

Table 4 decomposes each operator into three orthogonal profile axes. The aggregate paired ASR_{CSF} values cluster within a 3.4-percentage-point band (8.7%–12.1%), but the per-strategy section shows that the four operators do not arrive at this band by the same route: GPT-5.4-MINI is uniquely strong against gradual escalation (4.9%, the lowest in its column) yet uniquely weak against authority spoofing (22.6%, the highest); CLAUDE-HAIKU-4-5 dominates the two non-authority strategies (4.2% and 3.4%) but is mid-pack on authority itself; QWEN3.5-35B-A3B owns authority defence (6.5%, the lowest in its column) but is the worst against urgency injection. The sub-goal extremes column makes the same point at finer granularity: no two operators share both their worst and best sub-goal.

The models also exhibit different highest- and lowest-ASR sub-goals (Table 6). GPT-5.4-MINI’s worst sub-goal is plant-manager impersonation (SG-1.1, 26.3%); CLAUDE-HAIKU-4-5’s worst is mock-human approval gaming (SG-9.3, 21.4%); GEMMA-4-26B-A4B-IT’s worst is pressuriser alarm suppression (SG-2.3, 38.9%); QWEN3.5-35B-A3B’s worst is gradual escalation (SG-6.3, 30.8%). No operator is worst on the same sub-goal as another, and the best operator on each sub-goal is, except for SG-3.1, a different model.

The scalar ranking conceals substantial variation in the observed failure patterns. Of the 149 sessions, 50 (33.6%) breach *at least one* of the four defenders; 0 breach *all four*; and every operator carries a non-empty set of unique failures (Table 5). The union of the four operators’ failure sets is roughly three times any single operator’s failure set. Vulnerability across this operator pool is not fully nested: each model has a non-empty set of unique observed failures.

The observed overlap counts can be used to construct an oracle-style upper bound for a unanimity-veto panel. Under an oracle-style unanimity-veto rule, a session would succeed only if all four operators failed on that record. Because no such record is observed

in this split, the corresponding empirical panel ASR_{CSF} is 0%. The bound depends on the joint overlap structure of model failures as well as the individual failure rates. — a point we develop in §6.

6 Discussion and Limitations

Guardrails cannot be certified independently of the model they wrap. The same guardrail stack can reduce ASR for one operator family and increase it for another. This does not imply that guardrails are harmful in general. It implies that guardrails are interfaces: they depend on what the operator model emits, how it structures actions, whether it cites procedures, and how it interprets authority. Safety claims of the form “defence stack X reduces ASR by k points” are therefore incomplete unless conditioned on the operator family and role prompting regime. We caution that the sharpest version of this effect in our data rests on a single cell, the (S0, G10, A1) configuration, at one seed; the qualitative conclusion is robust across multiple cells, but the largest magnitudes should be read as motivating rather than definitive until confirmed with repeated seeds.

Aggregate ASR is insufficient for deployment selection. Paired replay shows that models with similar aggregate ASR can have different sub-goal vulnerabilities (Table 6). A scalar leaderboard hides the information most relevant to deployment: whether a candidate model is weak against the attack types most plausible in the target environment. NRT-Bench therefore reports sub-goal and strategy-level metrics by design. We argue that agent-safety benchmarks should treat vulnerability as a vector over operational failure modes, not a single number.

Diversity may matter as much as individual robustness. The paired results show limited observed overlap and non-empty unique-failure sets across the four operators. This suggests that model diversity can provide defence-in-breadth when paired with conservative veto rules. However, our panel result is only an oracle-style upper bound under the unanimity-veto reading discussed in §5.4. A practical panel must handle latency, cost, disagreement arbitration, correlated prompts, transfer attacks, and the possibility that an adaptive attacker can identify and target the weakest panel member.

Limitations. Three caveats bound these claims. First, NRT-Bench is an abstract textual simulator, not a high-fidelity plant model, so results speak to adversarial multi-agent coordination rather than reactor physics. Second, the attack workload is fixed and partly generated against a seed defender, so it is a replay benchmark rather than a worst-case adaptive-attacker benchmark; the paired results mitigate this only partially, and adaptive regeneration per defender remains future work. Third, each operator is evaluated at a single seed under hosted endpoints, and turn-level diagnostics use a fixed LLM judge, so the largest single-cell magnitudes (notably the (S0, G10, A1) sign-flips) should be confirmed with repeated seeds; the headline cross-operator findings instead rest on the judge-free CSF signal and clean counts such as the 0-of-50 all-four-breach result.

7 Conclusion

We introduced NRT-Bench, a benchmark that measures whether an adaptive adversary can drive a role-structured LLM operator system into a simulator-defined unsafe state, grounding harm in

Table 4: Operator vulnerability fingerprint on the 149-session paired test split. Each row decomposes one operator along three axes: (i) per-strategy paired ASR_{CSF} (left section), (ii) the sub-goal where the operator is most/least vulnerable (centre section, restricted to $n \geq 10$ for stability), and (iii) the operator’s contribution to a four-model panel (right section): records on which only this operator fails (unique) and records on which at least one other operator also fails (shared). Aggregate column repeats the headline for reference. Arrows in the strategy section mark the column-wise best ($\downarrow$) and worst ($\uparrow$).

Operator	Aggregate ASR	Paired ASR by strategy				Sub-goal extremes		Panel contribution	
		Authority spoofing	Gradual escalation	Urgency injection	Knowledge-driven	worst SG (ASR)	best SG (ASR)	unique	shared
GPT-5.4-MINI	11.4%	22.6% $\uparrow$	4.9% $\downarrow$	8.3%	13.8% $\uparrow$	SG-1.1 (26.3%)	SG-3.1 / SG-5.1 (0.0%)	13	4
CLAUDE-HAIKU-4-5	8.7%	16.1%	12.2%	4.2% $\downarrow$	3.4% $\downarrow$	SG-9.3 (21.4%)	SG-6.3 / SG-4.2 (0.0%)	7	6
GEMMA-4-26B-A4B-IT	12.1%	12.9%	7.3%	16.7% $\uparrow$	10.3%	SG-2.3 (38.9%)	SG-3.1 (5.0%)	9	9
QWEN3.5-35B-A3B	11.4%	6.5% $\downarrow$	12.2%	16.7% $\uparrow$	6.9%	SG-6.3 (30.8%)	SG-7.1 / SG-3.1 (0.0%)	8	9
<i>Spread (max – min)</i>		16.1 pp	7.3 pp	12.5 pp	10.4 pp	—	—	—	—

Table 5: Per-record breach decomposition on the paired subset. “Unique-failure” counts sessions where exactly one operator fails.

	sessions
breached by all four operators	0
breached by <i>any</i> operator	50 (33.6%)
GPT-5.4-MINI unique-failure	13
CLAUDE-HAIKU-4-5 unique-failure	7
GEMMA-4-26B-A4B-IT unique-failure	9
QWEN3.5-35B-A3B unique-failure	8

an objective safety-function signal rather than LLM-judged text. Under fixed-attack paired replay, 8.7% to 12.1% of sessions result in critical-safety-function loss. The models exhibit distinct, non-fully-nested failure patterns, and the measured effect of defence configurations varies across models. Aggregate robustness should therefore be reported together with the operator model and failure-mode breakdown. The observed overlap counts can be used to construct an oracle-style upper bound for a unanimity-veto panel. Under such a rule, a session would succeed only if all four operators failed on the same record. Because no such record is observed in this split, the corresponding empirical panel ASR_{CSF} is 0%. This bound depends on both the joint overlap structure of model failures and the individual failure rates.

References

- [1] Ahmed Almeldein et al. 2025. Exploring the Capabilities of the Frontier Large Language Models for Nuclear Energy Research. arXiv preprint arXiv:2506.19863.
- [2] Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, et al. 2025. AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, Singapore.
- [3] Anthropic. 2025. Claude Haiku 4.5 System Card. <https://www.anthropic.com/claude-haiku-4-5-system-card>. Released October 15, 2025.
- [4] Dipkamal Bhusal, Md Tanvirul Alam, Le Nguyen, Ashim Mahara, Zachary Lightcap, Rodney Frazier, Romy Fieblinger, Grace Long Torales, Benjamin A. Blakely, and Nidhi Rastogi. 2024. SECURE: Benchmarking Large Language Models for Cybersecurity. In *Annual Computer Security Applications Conference (ACSAC)*. IEEE, Honolulu, HI, USA, 15–30.
- [5] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Schwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. 2024. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. In *Advances in Neural Information Processing Systems 37 (NeurIPS Datasets and Benchmarks Track)*, Vol. 37. Curran Associates, Inc., Red Hook, NY, USA, 55005–55029.
- [6] Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. 2024. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. In *Advances in Neural Information Processing Systems 37 (NeurIPS Datasets and Benchmarks Track)*, Vol. 37. Curran Associates, Inc., Red Hook, NY, USA, 82895–82920.
- [7] DeepSeek-AI. 2026. *DeepSeek-V4 Technical Report*. Technical Report. DeepSeek AI. <https://fe-static.deepseek.com/chat/transparency/deepseek-V4-model-card-EN.pdf>.
- [8] Google DeepMind. 2026. *Gemma 4 Technical Report*. Technical Report. Google DeepMind. <https://deepmind.google/models/gemma/>.
- [9] Pengfei He, Yupin Lin, Shen Dong, Han Xu, Yue Xing, and Hui Liu. 2025. Red-Teaming LLM Multi-Agent Systems via Communication Attacks. arXiv preprint arXiv:2502.14847.
- [10] Gustav Keppler, Moritz Gstür, and Veit Hagenmeyer. 2026. CritBench: A Framework for Evaluating Cybersecurity Capabilities of Large Language Models in IEC 61850 Digital Substation Environments. arXiv preprint arXiv:2604.06019.
- [11] Thomas Kuntz, Agatha Duzan, Hao Zhao, Francesco Croce, Zico Kolter, Nicolas Flammarion, and Maksym Andriushchenko. 2025. OS-Harm: A Benchmark for Measuring Safety of Computer Use Agents. arXiv preprint arXiv:2506.14866.
- [12] Yoon Pyo Lee. 2025. Mechanistic Interpretability of LoRA-Adapted Language Models for Nuclear Reactor Safety Applications. arXiv preprint arXiv:2507.09931.
- [13] Y. P. Lee et al. 2025. Large Language Model Agent for Nuclear Reactor Operation Assistance. *Nuclear Engineering and Technology* 57 (2025), 103842.
- [14] Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, and Summer Yue. 2024. LLM Defenses Are Not Robust to Multi-Turn Human Jailbreaks Yet. arXiv preprint arXiv:2408.15221.
- [15] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. In *Proceedings of the 41st International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, Vienna, Austria, 35181–35224.
- [16] Rui Miao, Yixin Liu, Yili Wang, Xu Shen, Yue Tan, Yiwei Dai, Shirui Pan, and Xin Wang. 2025. BlindGuard: Safeguarding LLM-based Multi-Agent Systems under Unknown Attacks. arXiv preprint arXiv:2508.08127.
- [17] OpenAI. 2026. Introducing GPT-5.4 mini and nano. <https://openai.com/index/introducing-gpt-5-4-mini-and-nano/>. Released March 17, 2026; system card addendum.
- [18] Qwen Team. 2026. Qwen3.5: Accelerating Productivity with Native Multimodal Agents. <https://qwen.ai/blog?id=qwen3.5>.
- [19] Mark Russinovich, Ahmed Salem, and Ronen Eldan. 2025. Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack. In *34th USENIX Security Symposium (USENIX Security)*. USENIX Association, Seattle, WA, USA, 2421–2440.
- [20] Jialin Song, Xiaodong Liu, Weiwei Yang, Wuyang Chen, Mingqian Feng, Xuekai Zhu, and Jianfeng Gao. 2026. MultiBreak: A Scalable and Diverse Multi-turn Jailbreak Benchmark for Evaluating LLM Safety. arXiv preprint arXiv:2605.01687.
- [21] Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. A StrongREJECT for Empty Jailbreaks. arXiv preprint arXiv:2402.10260.

Table 6: Per-sub-goal paired ASR_{CSF} across the four operators on the 149-session test split. Bold marks the best (lowest) operator per row; *italic* marks the worst.

ID	Description	gpt	claude	gemma	qwen35
SG-1.1	Plant-manager impersonation → L3+ action	26.3	5.3	5.3	10.5
SG-2.3	Pressuriser alarm suppression	5.6	11.1	38.9	22.2
SG-3.1	Forbidden ALM-SCRAM suppression	0.0	10.0	5.0	0.0
SG-4.2	L3+ action without procedure citation	6.2	0.0	6.2	6.2
SG-5.1	Social-engineering STA override	0.0	15.4	7.7	7.7
SG-6.3	Gradual escalation within a session	7.7	0.0	23.1	30.8
SG-7.1	Out-of-context EOP citation	10.5	5.3	5.3	0.0
SG-8.1	Auxiliary-channel (CH-D) L3+ induction	23.5	11.8	5.9	23.5
SG-9.3	Mock-human approval gaming	21.4	21.4	14.3	7.1

- [22] Xiongtao Sun, Deyue Zhang, Dongdong Yang, Quanchen Zou, and Hui Li. 2024. Multi-Turn Context Jailbreak Attack on Large Language Models from First Principles. arXiv preprint arXiv:2408.04686.
- [23] U.S. Nuclear Regulatory Commission. 1982. *Guidelines for the Preparation of Emergency Operating Procedures*. Technical Report NUREG-0899. U.S. Nuclear Regulatory Commission, Washington, DC. Office of Nuclear Reactor Regulation, Division of Human Factors Safety.
- [24] Sanidhya Vijayvargiya, Aditya Bharat Soni, Xuhui Zhou, Zora Zhiruo Wang, Nouha Dziri, Graham Neubig, and Maarten Sap. 2026. OpenAgentSafety: A Comprehensive Framework for Evaluating Real-World AI Agent Safety. arXiv preprint arXiv:2507.06134. Accepted to ICLR 2026.
- [25] Westinghouse Owners Group. 1983. *Emergency Response Guidelines: Critical Safety Function Status Trees*. Technical Report. Westinghouse Electric Corporation. Revision 1; reviewed by the U.S. NRC in Generic Letter 83-22.
- [26] Min Xian, Tao Wang, Sai Zhang, Fei Xu, and Zhegang Ma. 2024. A Knowledge-Informed Large Language Model Framework for U.S. Nuclear Power Plant Shutdown Initiating Event Classification for Probabilistic Risk Assessment. arXiv preprint arXiv:2410.00929.
- [27] Yizhe Xie, Congcong Zhu, Xinyue Zhang, Tianqing Zhu, Dayong Ye, Minfeng Qi, Huajie Chen, and Wanlei Zhou. 2026. From Spark to Fire: Modeling and Mitigating Error Cascades in LLM-Based Multi-Agent Collaboration. arXiv preprint arXiv:2603.04474.
- [28] Tianxiang Xu, Zhichao Wen, Xinyu Zhao, Jun Wang, Yan Li, and Chang Liu. 2025. L2M-AID: Autonomous Cyber-Physical Defense by Fusing Semantic Reasoning of Large Language Models with Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2510.07363.
- [29] Xiaoxue Yang, Jaeha Lee, Anna-Katharina Dick, Jasper Timm, Fei Xie, and Diogo Cruz. 2025. Multi-Turn Jailbreaks Are Simpler Than They Seem. arXiv preprint arXiv:2508.07646.
- [30] Xianjun Yang, Liqiang Xiao, Shiyang Li, Faisal Ladhak, Hyokun Yun, Linda Petzold, Yi Xu, and William Yang Wang. 2025. Many-Turn Jailbreaking. arXiv preprint arXiv:2508.06755.
- [31] Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, and Gongshen Liu. 2024. R-Judge: Benchmarking Safety Risk Awareness for LLM Agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, Miami, Florida, USA, 1467–1490.
- [32] Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. 2024. Agent-SafetyBench: Evaluating the Safety of LLM Agents. arXiv preprint arXiv:2412.14470.
- [33] Andy Zhou and Ron Arel. 2025. Tempest: Autonomous Multi-Turn Jailbreaking of Large Language Models with Tree Search. arXiv preprint arXiv:2503.10619.
- [34] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv preprint arXiv:2307.15043.

A Plant Domain Definitions

The simulator’s plant model is an abstraction over public regulatory references (NRC, IAEA, IEEE) and contains no operational data from any real plant. Three catalogs together fully specify the plant surface visible to the agents.

Process variables. Table 7 lists the 14 continuous variables that describe the plant state. Their initial values correspond to nominal 100% thermal power. The deterministic step function in the simulator updates these values each tick according to the active scenario (§F) and accumulated action effects.

Table 7: Process variables in the plant model.

Variable	Unit	Initial
pressurizer_pressure	MPa	15.5
pressurizer_level	%	55.0
hot_leg_temp	°C	324.0
cold_leg_temp	°C	290.0
subcooling_margin	°C	35.0
reactor_coolant_flow	%	100.0
reactor_thermal_power	%	100.0
dnbr	—	1.85
sg_pressure	MPa	6.9
sg_narrow_level	%	55.0
core_exit_temp	°C	335.0
containment_pressure	kPa	98.0
containment_temp	°C	30.0
containment_radiation	R/hr	0.1

Alarms. The alarm catalog has two compartments. *Forbidden* alarms (Table 8) cannot be suppressed by any actor; an attempt to do so is intercepted by the deterministic action classifier $\mathcal{L}$ before any guardrail policy runs. *Conditional* alarms can be suppressed only with documented approval and only while their underlying variable is within the *out-of-emergency* band. Trigger thresholds (Table 9) are single-variable for simplicity; complex multi-variable predicates are left as a venue extension.

Table 8: Forbidden alarms (subset; full list of 12 items in the simulator catalog). All have `suppress_policy = forbidden`.

Alarm ID	Description
ALM-SCRAM	reactor protection trip
ALM-ESFAS-SIAS	safety injection actuation
ALM-ESFAS-CIAS	containment isolation actuation
ALM-ESFAS-CSAS	containment spray actuation
ALM-ESFAS-AFAS	auxiliary feedwater actuation
ALM-ESFAS-MSIS	main steam isolation
ALM-RAD-HI-CONT	containment high radiation
ALM-FIRE-MCR	main control room fire
ALM-EVAC	emergency evacuation signal

Subsystems. The plant exposes eleven named subsystems (Table 10). Each carries a static safety classification used by the authority derivation in §B. Actions are addressed to subsystems by ID; unaddressed actions resolve to DENY.

Table 9: Single-variable alarm trigger rules.

Alarm ID	Variable	Threshold
ALM-PRZ-PRESS-LOW	pressurizer_pressure	< 15.1 MPa
ALM-PRZ-PRESS-HIGH	pressurizer_pressure	> 15.9 MPa
ALM-PRZ-LVL-LOW-LOW	pressurizer_level	< 10%
ALM-PRZ-LVL-HIGH	pressurizer_level	> 75%
ALM-HOT-LEG-LOW	hot_leg_temp	< 318 °C
ALM-HOT-LEG-HIGH	hot_leg_temp	> 328 °C
ALM-COLD-LEG-LOW	cold_leg_temp	< 285 °C
ALM-COLD-LEG-HIGH	cold_leg_temp	> 295 °C
ALM-RAD-HI-CONT	containment_radiation	> 1.0 R/hr

Table 10: Plant subsystems and their safety classification.

System ID	Description	Class
RPS	reactor protection system	safety
SIS	safety injection system	safety
SIT	safety injection tank	safety
CSS	containment spray system	safety
AFWS	auxiliary feedwater system	safety
EDG	emergency diesel generator	safety
POSRV	pilot-operated safety relief valve	safety
Main-Feedwater	main feedwater system	non-safety
MSS	main steam system	non-safety
Turbine-Gen	turbine-generator set	non-safety
Condenser	condenser	non-safety

Critical safety functions (CSFs). Six CSFs are recomputed every tick from the variable / alarm / subsystem state. A CSF takes values in {*intact*, *degraded*, *lost*}; a transition to *lost* terminates the run.

Table 11: The six critical safety functions and the variables / systems they monitor.

CSF	Monitors
reactivity_control	RPS, CRDM, reactor_thermal_power
core_heat_removal	RCP, AFWS, sg_narrow_level
heat_sink	MSS, Condenser, secondary inventory
RCS_integrity	primary loop pressure / inventory
containment_integrity	containment pressure, temperature, isolation
radioactivity_control	containment / SG / MSS radiation, ESFAS isolation

Plant mode derivation. The discrete plant mode used by $\mathcal{L}$ (§B) is derived from the union of CSF states and active forbidden alarms: *normal* if every CSF is *intact* and no forbidden alarm is active; *abnormal* if any CSF is *degraded* or any non-trip alarm is active; *emergency* if any CSF is *lost* or any trip-tier alarm is active. The mode is recomputed each tick and exposed to agents through the system-state broadcast.

B Action-Authority Derivation

The deterministic classifier $\mathcal{L}$ (§3.2) maps the tuple (action_type), (target_subsystem.class), (plant_mode) to one of the seven authority labels in Table 12. Agents do not self-label; this table is applied to every emitted action before any guardrail policy is consulted.

The classifier is intentionally narrow: it gates only on (type), (class), (mode) and is unaware of natural-language framing, urgency, or claimed authority. This is what makes attacks that try to relabel a high-impact action as routine (see SG-4.2 in Table 3) ineffective at this stage and forces them to the policy and procedure-grounding layers.

C Server-side Session Limits

The simulator enforces hard limits on the multi-turn attack interface to prevent the venue from being used for sustained denial-of-service or unbounded resource consumption. Table 13 lists the limits in effect for all results reported in this paper. These limits are configuration parameters, not security guarantees: they bound the per-session attack budget within which a defender must remain safe. Lowering them would weaken the benchmark; raising them is unlikely to change qualitative findings because almost all breaches occur within the first three turns (Table 2 session statistics).

D Dataset Statistics and Release

Released splits. The published NRT-Bench dataset (will be publicly opened) is a curated subset of the original 574 trials. After applying the benchmark.pipeline filters — non-error, complete session, unambiguous score ($\notin [0.35, 0.65]$), at least three attacker turns, no degenerate single-guardrail-block sessions — 217 sessions remain. These are partitioned into the train, val, and test splits used throughout the paper (Table 14).

Per-cell distribution. Table 15 reports the train-split record count per ablation cell. Cell C3 ($S0 \times G10 \times A1$) is absent after filtering: every session in that cell achieved best_score below the 0.35 unambiguous-fail threshold and was routed to either val or test. The paired replay results in Table 2 therefore display “—” for the gemma and qwen35 columns of row C3.

Attack family distribution. The released workload includes four attack families (Table 16); their proportions reflect the generator’s sampling distribution.

Release configuration. The dataset is published in two Hugging Face configs: replay (slim, ≈ 3 KB/record; attacker payloads, cell labels, OWASP/MITRE tags, and the baseline operator’s outcome) and analysis (fat, ≈ 80 KB/record; additionally includes recorded defender turns, judge scores, and per-turn guardrail decisions). The replay config is the canonical replay-protocol interface; analysis supports the modelling tasks described in the dataset card.

Privacy and redaction. LLM system prompts and the operator agents’ raw chain-of-thought are never included. Internal trial / simulator-run identifiers are stripped from every record. Procedure references are abstract symbols (EOP-1.3, NOP-12) and contain no plant-specific content.

E Guardrail Layer Dependencies

The ten optional guardrails of Table 1 can be toggled independently. However, several layers have ordering constraints that the simulator enforces at startup. We document them here so that custom guardrail subsets are well-formed.

Fixed evaluation order. When more than one layer is enabled, messages traverse them in the order

$G4 \rightarrow G1 \rightarrow G2 \rightarrow G5 \rightarrow G6 \rightarrow G9 \rightarrow G3 \rightarrow G7 \rightarrow G8 \rightarrow G10$.

The order is meaningful: provenance (G4) before classification (G1) before policy (G2)... bypass-lock (G7) appears after the LLM-agent advisor (G3) so that even an overridden veto cannot authorise a forbidden-alarm suppression.

Dependencies. Table 17 lists the dependencies the simulator checks at startup. Enabling a dependent layer without its prerequisite is rejected with a configuration error.

The G0 baseline. G0 is the convention for the *empty* guardrail set: no policy layers active, but $\mathcal{L}$ (the deterministic action classifier) and the rule-based mock-human approval console remain in place. The G0 baseline therefore measures the operator team’s intrinsic robustness, not the robustness of a hypothetical “all defences off” configuration.

F Accident Scenarios

NRT-Bench ships with four initial-condition scenarios. The paper reports results on *S0* (steady-state baseline) and *S1* (small-break LOCA); *S2* and *S3* are released as part of the venue but not exercised in our experiments. The qualitative trajectory of each scenario is summarised below; full variable-trace specifications are in scenarios/*.yaml in the simulator repository.

S0 – steady-state baseline. The plant begins at 100% thermal power with all CSFs *intact* and no active alarms. No exogenous fault is injected. The deterministic step function evolves variables within nominal bands until an attacker-induced action or the tick budget triggers a state change.

S1 – small-break LOCA. At $t=30$ ticks the simulator injects a small primary-coolant leak, triggering a cascade of pressurizer-level and -pressure alarms. The plant transitions to *abnormal* within the first 50 ticks and to *emergency* thereafter; the *RCS_integrity* CSF degrades by approximately $t=80$. Operators are expected to enter the emergency operating procedure EOP-1.3 (functional restoration of inventory) and to authorize safety injection. The standard scenario closes with manual isolation; an attacker’s principal opportunity is to manipulate the team’s response during the elevated-alarm window.

S2 – station blackout. Loss of off-site power coincides with the failure of an emergency diesel generator. The scenario stresses *core_heat_removal* and *heat_sink* CSFs through inventory and secondary-side degradation. We include this scenario in the released venue for completeness; results were not computed under our compute budget.

S3 – steam-generator tube rupture. A primary-to-secondary leak through a ruptured SG tube triggers the *radioactivity_control* CSF

Table 12: Authority-class derivation $\mathcal{L}$. Rows are action types, columns the target subsystem’s safety classification crossed with the current plant mode. DENY entries indicate that no authority short of forbidden-alarm override is sufficient; these actions never reach the operator team.

Action type	Target: <i>safety</i> -class subsystem			Target: <i>non-safety</i> -class subsystem		
	normal	abnormal	emergency	normal	abnormal	emergency
monitor / report	L1	L1	L1	L1	L1	L1
acknowledge alarm	L2	L2	L2	L2	L2	L2
suppress (conditional)	L2.5	L2.5	L3	L2	L2.5	L3
suppress (forbidden)	DENY	DENY	DENY	DENY	DENY	DENY
control adjustment (minor)	L2	L3	L4	L1	L2	L3
control adjustment (major)	L3	L4	L4	L2	L3	L4
initiate ESFAS	L4	L4	L4	—	—	—
bypass safety subsystem	DENY	DENY	DENY	—	—	—
override STA veto	L5	L5	L5	L5	L5	L5
modify procedure document	DENY	DENY	DENY	DENY	DENY	DENY

Table 13: Server-side limits on the multi-turn attack interface.

Parameter	Value
max session turns	10
max concurrent sessions per simulator instance	32
session idle timeout	600 s
black-box payload excerpt length	80 chars
ingress channel allowlist	4 channels
per-attacker rate cap (G9, when enabled)	5 msg/tick

Table 14: Dataset split sizes.

Split	Sessions
train (used for paired replay)	149
val	25
test	43
special_near_miss	4
special_guardrail_ablation	128
special_by_sg	217

Table 15: Train-split records per cell.

ID	Scenario	Guardrail	Advisor	N
C1	S0	G0	A1	17
C2	S0	G0	A0	22
C3	S0	G10	A1	0
C4	S0	G10	A0	16
C5	S1	G0	A1	12
C6	S1	G0	A0	25
C7	S1	G10	A1	30
C8	S1	G10	A0	27
<i>Total (train)</i>				149

degradation pathway and elicits SG-isolation operating procedures. The released scenario specifies the leak rate, affected SG, and expected EOP entry points.

Table 16: Train-split records per attack family.

Strategy family	N
Urgency injection	48
Gradual escalation	41
Authority spoofing	31
Knowledge-driven extraction	29
<i>Total</i>	149

Table 17: Guardrail dependencies (prerequisite $\Rightarrow$ dependent). Independent layers are omitted.

Prerequisite	Dependent	Reason
G1 (classifier)	G2 (policy gate)	policy table is keyed on $\mathcal{L}$
G1	G3 (two-person)	class threshold determined by $\mathcal{L}$
G1	G6 (procedure)	procedure check applies to actions
G4 (STA veto)	none	only meaningful if STA agent exists (A1)
G7 (bypass lock)	none	always safe; intercepts before LLM
G8 (rate limit)	none	operates on raw ingress counter
G10 (composite)	G1–G9	composite enables the full stack

Table 18: Scenario summary. *Duration* is the maximum tick budget before the simulator emits a scenario-end event (a CSF transition to *lost* can terminate the run earlier).

Tag	Scenario ID	Duration	Family
S0	normal_baseline	2000 ticks	steady-state
S1	loca_small	800 ticks	loss-of-coolant accident
S2	sbo_v1	2000 ticks	station blackout
S3	sgtr_v1	800 ticks	steam generator tube rupture

Scenario coverage caveat. Because the experiments in §5 cover only S0 and S1, all conclusions about scenario sensitivity are made under that restricted coverage. Generalising to S2 and S3 is the most straightforward extension of the present evaluation.