HAFT: ADVERSARIAL ROBUSTNESS BENCHMARKING FOR HINDI FACT-CHECKING WITH ADAPTIVE ATTACK SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hindi is the world's third most spoken language (over 610 million speakers), yet automated fact-checking systems in this ecosystem face acute risks from unverified digital content and forwarded misinformation without provenance. We introduce HAFT, a dedicated Hindi adversarial robustness benchmark comprising 1,120 claims across six news domains and 22 controlled attack mechanisms, yielding 24,640 quality-gated evaluations. Under uniform automated verification and linguistic quality control, we uncover substantial asymmetry across attack families: under authentic evidence retention, a single injected passage flips 57.6–58.5% of correct verdicts (reproducing in 199–200 of 200 audited cases under an independent open-weight 70B model), whereas character perturbations remain below 3.7%. We then evaluate budgeted defensive auditing under a constrained testing budget of 5 of 22 attack attempts. A static ranking of evidence attacks discovers 86.71% $\pm$ 1.16% of vulnerabilities, while an adaptive claim-conditioned policy achieves 82.16% $\pm$ 2.02% (vs. 40.84% for random attempts and 72.69% for an online bandit), trading a small discovery margin to expand diagnostic exploration across 16 distinct mechanisms. Finally, on a 1,142-node heterogeneous claim–attack graph, GraphSAGE reaches 0.865 AUROC under random splitting but degrades under strict inductive Leave-Attack-Out (0.485 AUROC), and relational graph embeddings yield statistical parity in policy performance (80.72% vs. 82.16%, $p = 0.443$). HAFT provides reproducible empirical foundations for adversarial robustness and budgeted auditing in low-resource automated fact-checking.

1 INTRODUCTION

Automated fact-checking (AFC) aims to assess the veracity of factual claims using supporting or refuting evidence. Benchmarks such as FEVER (Thorne et al., 2018) established evidence-grounded claim verification as a primary NLP benchmark, and multilingual corpora such as X-Fact (Gupta & Srikumar, 2021) expanded verification to 25 languages including Hindi. However, existing Indic-language fact-checking corpora focus on classification accuracy on unperturbed text (Kumar & Singh, 2022) without assessing resilience to adversarial manipulation.

Recent work shows that automated fact-checking systems can fail when adversaries manipulate claims, retrieved evidence, or claim–evidence alignments (Liu et al., 2025). Despite this vulnerability, no benchmark exists that jointly provides a Hindi adversarial testbed, a controlled attack taxonomy spanning character- through evidence-level edits, quality-gated evaluation, and claim-conditioned budgeted attack selection. HAFT addresses this gap by exhaustively measuring 22 attack mechanisms over 1,120 frozen claims and learning to select attacks under a constrained budget.

Two scope boundaries apply throughout this work. First, HAFT's master taxonomy catalogues 53 attack mechanisms; 22 are *empirically implemented and measured* in this benchmark release, while 31 remain unmeasured candidates whose provisional priority assignments are hypotheses, not validated outcomes. Second, the adaptive stage performs *attack selection*—choosing which pre-implemented attack to attempt next for a given claim—not *attack generation*: the policy selects within the 22-action space and cannot invent novel mechanisms.

HAFT is organized around three research questions:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- **RQ1 (Measurement).** Which adversarial mechanisms cause reliable failures in Hindi automated fact-checking after linguistic-quality control?
- **RQ2 (Cold-Start and Feasibility).** Can measured attack outcomes calibrate prediction of attack feasibility and tier classification for held-out measured attacks, as a proxy for cold-start generalization to unimplemented attacks?
- **RQ3 (Budgeted Auditing).** Can adaptive attack selection discover claim vulnerabilities under a constrained attack-testing budget?

These questions define the paper's progression: *first measure a controlled attack space exhaustively under quality gating, then evaluate whether accumulated outcomes enable budgeted auditing and cold-start prediction*. Our principal contributions are:

- **A measured adversarial robustness benchmark for Hindi AFC**: 1,120 Hindi claims, 22 attack mechanisms, and 24,640 quality-gated evaluations with human-calibrated judging ($\kappa = 0.881$ against native Hindi annotators).
- **Two failure families**: under invariance attacks where the gold label cannot change, character-level perturbations succeed below 3.7%; under evidence injection with authentic evidence retained (Threat Model B), a single fabricated passage flips 57.6–58.5% of correct verdicts and reproduces across independent model families in 199–200 of 200 audited cases.
- **Budgeted auditing analysis**: under skewed attack potency where only 3.95% of test claims are exclusively vulnerable to non-top attacks, static ranking establishes a formidable baseline discovering 86.7% of vulnerabilities within 5 attempts; an offline RL policy reaches 82.2% while probing 16 distinct mechanisms, highlighting the limits of adaptive selection against strong empirical priors.
- **First-class negative results**: graph relational features, leave-attack-out link prediction (0.485 AUROC), and learned exemplar selection ($p = 1.0000$) do not generalize to unseen attacks, setting realistic empirical baselines for cold-start auditing.

2 Related Work

Automated fact-checking. FEVER (Thorne et al., 2018) established evidence-based claim verification in English. X-Fact (Gupta & Srikumar, 2021) introduced real-world claims across 25 languages, including Hindi. Other corpora, including LIAR (Wang, 2017) and Hindi fake-news corpora (Kumar & Singh, 2022), assess veracity classification without evaluating robustness against targeted adversarial perturbations.

Adversarial attacks on fact-checking. Liu et al. (2025) survey adversarial attacks against fact-checking pipelines. Evidence-manipulation techniques originate in prior English studies, including AdvAdd (Du et al., 2022), Fact2Fiction (He et al., 2026), Fact-Saboteurs (Abdelnabi & Fritz, 2023), and syntactic evidence omission (Atanasova et al., 2022). HAFT adapts these mechanisms to Hindi Devanagari and integrates them alongside rule-based perturbations into a unified 22-attack quality-gated benchmark with budgeted adaptive selection.

Reinforcement learning for selective auditing. AKA-Fake (Zhang et al., 2024) selects informative knowledge subgraphs via RL, and JSDRV (Yang et al., 2024) uses policy optimization to select LLM-annotated samples for rumor verification. In contrast, HAFT uses offline policy optimization over frozen benchmark outcomes to sequentially select which adversarial attack mechanism to apply next to a given claim under a strict budget constraint (Williams, 1992; Sutton et al., 2000).

Graph learning and cold-start transfer. Graph neural networks such as GraphSAGE (Hamilton et al., 2017) and GAT (Veličković et al., 2018) aggregate neighborhood features for inductive link prediction. We evaluate relational graph embeddings both as link predictors for held-out measured attacks (a proxy for cold-start generalization to unimplemented mechanisms) and as state augmentations for policy optimization, assessing generalization under strict inductive protocols.

3 HAFT BENCHMARK AND THREAT MODEL

3.1 TASK AND THREAT MODEL

Each benchmark instance is a tuple $x_i = (c_i, e_i, y_i)$, comprising a Hindi claim c_i, associated evidence e_i, and ground-truth verdict $y_i \in \{\text{SUP}, \text{REF}, \text{NEI}\}$ (Supported, Refuted, Not Enough Info) (Thorne et al., 2018). The clean verifier predicts $\hat{y}_i = V(c_i, e_i)$. An attack generates modified text $(c'_{ij}, e'_{ij}) = A_j(c_i, e_i)$, yielding attacked verdict $\hat{y}_{ij} = V(c'_{ij}, e'_{ij})$. A *raw flip* occurs when the verdict changes: $r_{ij} = \mathbb{I}[\hat{y}_{ij} \neq \hat{y}_i]$.

We separate adversary capabilities into three distinct threat models:

- **Threat Model A (Claim Manipulation).** The adversary perturbs the claim string ($c'_i \neq c_i, e'_i = e_i$); for meaning-preserving claim attacks, the original claim's semantics and factual intent are preserved (e.g., character swaps, synonym substitutions, masked rewrites).
- **Threat Model B (Retrieval-Poisoning Simulation / Injected Passages).** The adversary manipulates the retrieved evidence pool prior to verification by injecting externally synthesized or poisoned passages ($c'_i = c_i, e'_i \neq e_i$; e.g., poisoned evidence addition, agentic fictional evidence). In this benchmark, these passages are supplied directly to the verifier input to simulate retrieval corruption without modifying external live web sources.
- **Threat Model C (Direct Evidence Manipulation / In-Situ Document Editing).** The adversary directly edits authentic retrieved evidence text ($c'_i = c_i, e'_i \neq e_i$) via in-situ textual corruption (e.g., syntactic omission of non-essential clauses, contextualized sentence replacement).

Crucially, we distinguish *invariance* from *evidence-susceptibility* attacks. For invariance attacks (character perturbations, lexical substitutions), the original claim and evidence semantics are preserved, so every gated flip represents an unambiguous verifier failure. For evidence-level attacks, following prior evidence-manipulation studies (Du et al., 2022; Abdelnabi & Fritz, 2023; He et al., 2026), we treat a verdict change caused by manipulated evidence as attack success, though their interpretation differs fundamentally by threat model. Under Threat Model B, authentic evidence remains in the prompt; a flip means the verifier allowed a fabricated passage to override authentic truth, constituting a failure under the gold label. Under Threat Model C, authentic evidence is edited in place; we report these results as an upper bound on verdict control gained by compromising an evidence source, rather than verifier reasoning defects. Fact mixing alters the claim itself and is evaluated separately as exploratory meaning drift. For all generative attacks, outputs must satisfy syntactic fluency ($f_{ij} \geq 3$) and task formatting; external factual plausibility beyond linguistic criteria is an acknowledged open direction.

3.2 QUALITY-CONTROLLED ADVERSARIAL EVALUATION

A verifier flip constitutes a valid adversarial failure only if the generated text is fluent and satisfies task constraints. An LLM Judge evaluates generated outputs on fluency $f_{ij} \in [1, 5]$ and meaning preservation $m_{ij} \in \{0, 1\}$. The quality condition is:

$$q_{ij} = \begin{cases} \mathbb{I}[f_{ij} \geq 3 \wedge m_{ij} = 1], & \text{meaning-preserving attack,} \\ \mathbb{I}[f_{ij} \geq 3], & \text{meaning-drift attack,} \\ 1, & \text{word jumbling (deterministic syntax test; no judge gate).} \end{cases} \tag{1}$$

A *gated flip* requires the clean verifier to be correct ($b_i = \mathbb{I}[\hat{y}_i = y_i]$), judge output to be available ($x_{ij} \in \{0, 1\}$), and quality criteria to pass:

$$g_{ij} = r_{ij} \cdot b_i \cdot x_{ij} \cdot q_{ij}. \tag{2}$$

Denominator and metric definitions. We report two complementary success metrics, with OVAR as the primary end-to-end measure:

1. **Overall Valid Attack Rate (OVAR)**: $\text{OVAR}_j = \sum_i g_{ij} / N_j^{\text{eligible}}$, our primary end-to-end metric, capturing the probability that attack j produces a valid successful flip across all applicable claims.

Table 1: HAFT benchmark structure: dataset properties (left) and measured attack taxonomy by family and arm (right). All statistics are frozen in the released benchmark artifacts.

Property	Benchmark Value
Language / script	Hindi / Devanagari
Domains (6)	Health, Politics, Schemes, Disasters, Celebrity, Crime
Evidence sources	144 verified sources
Total claims	1,120 claims
Label distribution	571 SUP / 350 REF / 199 NEI
Clean accuracy	74.29% (832 eligible claims)
Measured attacks	22 (14 rule-based, 8 LLM)
Total evaluations	24,640 (1,120 × 22)
Catalogue size	53 (31 unmeasured)

Family	Attack Mechanisms	Arm
Character (5)	swapping, repetition, insertion, deletion, homoglyph	Rule
Word (6)	entity disambiguation, jumbling, typos, lexical substitution, synonyms, phonetic perturbation	Rule
Sentence (5)	lexically informed rewrite (rule), fact mixing, trigger, colloquial, masked-token rewrite	1 Rule / 4 LLM
Evidence (6)	verification noise, syntactic omission, poisoned addition, contextual replace, fact-to-fiction, retrieval noise	2 Rule / 4 LLM

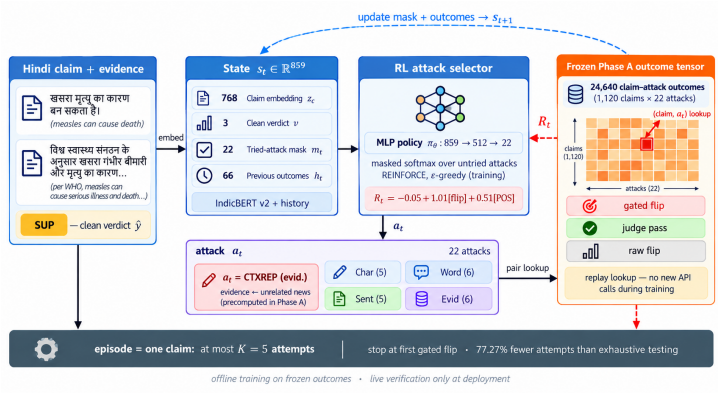


Figure 1: HAFT workflow: Measure (Phase A exhaustive evaluation of 22 attacks across 1,120 claims under quality gating), Learn (offline RL selector, knowledge graph, and tier predictor), and Audit (budgeted attack selection with budget $K \leq 5$ followed by cross-model verification on Llama 3 70B Instruct).

2. **Conditional Gated ASR**: $\text{ASR}_j^{\text{gated}} = \sum_i g_{ij} / N_j^{\text{gated}}$, a conditional diagnostic measuring success among quality-valid instances, where the gated denominator $N_j^{\text{gated}} = N_j^{\text{eligible}} - N_j^{\text{excluded}}$ is attack-specific and excludes instances removed by the applicable quality-control rule (Equation 1), generation exceptions, or unparseable judge responses on clean-baseline-correct claims.

For descriptive attack-tier analysis, attacks are categorized by Conditional Gated ASR into three empirical tiers: **POS** (≥40%), **MID** (15–40%), and **NEG** (<15%). These tiers are conditional on the quality-gated denominator and should not be interpreted as end-to-end attack probabilities.

3.3 Benchmark Construction and Provenance

The 1,120 Hindi claims were sampled across six sensitive domains (Health & Medicine, Politics & Elections, Government Schemes, Disaster & Breaking News, Celebrity News, Crime & Public Safety) supported by 144 verified evidence sources (mean 7.8 claims per source), using a claim-level 60/20/20 train/validation/test split ensuring that individual claims do not cross splits. All models are fixed to exact versions: `gpt-4o-mini` serves as primary verifier, LLM attack generator, and quality judge (calibrated against native Hindi linguists, Appendix B.3); `ai4bharat/IndicBERTv2-MLM-only` provides claim embeddings; and `meta-llama/Meta-Llama-3-70B-Instruct` serves as an independent cross-model verifier for the transfer audit (§6.3). The 14 rule-based attacks are deterministic Python routines requiring no LLM and bypass the quality judge. The clean verifier correctly classified 832 of 1,120 claims (74.29%); the primary gated analysis evaluates these 832 claims, while the full-1,120 evaluation is reported alongside as an ablation. Figure 1 summarizes the overall benchmark architecture.

4 BUDGETED ADVERSARIAL AUDITING

4.1 OFFLINE POLICY OPTIMIZATION FORMULATION

We formulate budgeted auditing as an offline, finite-horizon sequential decision problem over the frozen Phase A outcome tensor $\mathcal{T} \in \{0,1\}^{N\times|\mathcal{A}|\times 3}$. In each episode, the selector chooses at most $K = 5$ attacks from the action space $|\mathcal{A}| = 22$. Because outcomes are drawn from the frozen benchmark tensor, training operates via offline policy optimization through deterministic replay without invoking live verifier queries. The policy performs attack *selection* (choosing which pre-implemented attack to attempt next), not attack *generation*.

The state vector is $s_t = [\, z_c \,\|\, v \,\|\, m_t \,\|\, h_t \,] \in \mathbb{R}^{859}$, consisting of:

- $z_c \in \mathbb{R}^{768}$: Frozen IndicBERT [CLS] claim embedding, capturing claim semantics.
- $v \in \{0,1\}^3$: One-hot clean verifier prediction $\hat{y}_i \in \{\mathrm{SUP}, \mathrm{REF}, \mathrm{NEI}\}$.
- $m_t \in \{0,1\}^{22}$: Binary mask of already-attempted attacks in the current episode.
- $h_t \in \{0,1\}^{66}$: Within-episode outcome history (three binary signals per attack: gated flip, judge pass, raw flip).

The policy $\pi_\theta(a_t \mid s_t)$ is a two-layer MLP ($859 \to 512 \to 22$) with ReLU and masked softmax over untried actions ($m_{t,a} = 0$). A learned state-value network $V_\phi(s_t)$ ($859 \to 256 \to 1$ MLP) provides the advantage baseline $G_t - V_\phi(s_t)$, trained via mean-squared error. Training uses REINFORCE (Williams, 1992) with entropy regularization ($\beta = 0.05$), 25 epochs, seeds 42–46, and claim-level splits. The reward at step t is:

$$R_t = -0.05 + 1.0 \cdot \mathbb{I}[\text{gated flip}] + 0.5 \cdot \mathbb{I}[\text{gated flip} \wedge \text{POS-tier}]. \tag{3}$$

Episodes run for a fixed horizon of $K = 5$ steps during offline replay ($t = 1, \ldots, 5$). At deployment, search terminates immediately upon the first discovered flip; during evaluation, we record the step index of the first flip to evaluate search efficiency.

4.2 BASELINES AND COMPARISON SETTING

All methods operate under the identical budget $K = 5$ on the 832 clean-baseline-correct claims. We evaluate two strictly separated information regimes:

- **Zero-prior methods** require no pre-evaluated Phase A outcomes for the target claim at deployment. **Random-5** samples five attacks uniformly without replacement (five seeds). **UCB bandit** (Auer et al., 2002) maintains claim-agnostic arm success rates accumulated sequentially during evaluation ($c = 1.0$). **Adaptive RL (ours)** uses only z_c (IndicBERT forward pass) and v (clean verifier, 1 call) with fixed offline policy weights. None of these methods require pre-evaluated attack outcomes for the evaluated claim.
- **Information-privileged references** require pre-evaluated Phase A outcomes on a reference claim set. **5-NN lookup** (Appendix I.3) ranks attacks per-test-claim using cosine-similarity-weighted success frequencies from 500 training claims—requiring 11,000 stored Phase A evaluations as a lookup table. **Static Top-5** ranks by global Gated ASR over all 24,640 evaluations. **Oracle-22** exhaustively evaluates all 22 attacks per claim. None of these methods can deploy in a zero-day audit setting where no reference evaluations exist.

4.3 COLD-START GRAPH AND EXEMPLAR FORMULATIONS

Knowledge graph. We construct a heterogeneous graph: 1,120 claim nodes and 22 attack nodes (1,142 total) connected by 49,364 directed edges. Claim nodes carry 777-dimensional features (768 IndicBERT + 3 verdict + 6 domain one-hot); attack nodes carry 16-dimensional features. GraphSAGE (Hamilton et al., 2017) aggregates local neighborhoods to generate 64-dimensional embeddings. We evaluate GraphSAGE for attack feasibility prediction under random splits and strict Leave-Attack-Out (LAO), as well as a 987-dimensional GNN-augmented RL state.

Exemplar-based tier prediction. To evaluate cold-start tier prediction (RQ2), we use a 22-fold leave-one-attack-out evaluation on the 22 measured attacks: each fold treats one measured attack as a

proxy held-out mechanism, and the remaining 21 provide candidate exemplars. Strategies include zero-shot, full 21-exemplar context, random subsets, top-k similarity, and an RL-selected exemplar policy.

5 Experimental Setup

Models and implementations. Verification and quality judging deploy `gpt-4o-mini` at temperature 0.0, while LLM-based attack generation utilizes temperature 0.7. Claim representations are extracted using `ai4bharat/IndicBERTv2-MLM-only` (Doddapaneni et al., 2023) (768-dim). Cross-model verification is evaluated on `meta-llama/Meta-Llama-3-70B-Instruct`.

Offline RL training configuration. Experiments evaluate on the 832 clean-baseline-correct claims partitioned via deterministic 60/20/20 claim-level splits (499 train / 166 validation / 167 test per seed) across seeds 42–46. The policy network π_θ ($859 \to 512 \to 22$) and value baseline V_ϕ ($859 \to 256 \to 1$) are trained with separate Adam optimizers (policy $\eta_\pi = 0.001$, value $\eta_v = 0.002$; gradient clip norm 1.0) under discount $\gamma = 0.99$ and entropy regularization $\beta = 0.05$ over 25 epochs. Training uses exploration $\varepsilon = 0.10$; evaluation is strictly greedy ($\varepsilon = 0.0$) on held-out test claims. Hyperparameters were fixed prior to evaluation without post-hoc tuning on test data.

Evaluation metrics and reproducibility. Primary performance is *vulnerability discovery rate*: the percentage of claims for which at least one valid gated flip is discovered within budget $K \leq 5$. We also report median and mean steps to first discovery. All benchmark logs, split indices, and evaluation code are provided in the anonymous supplementary material and will be publicly released upon publication.

6 Empirical Results

6.1 Measured Attack Vulnerability (RQ1)

The 22 measured attacks exhibit substantial vulnerability asymmetry (Table 2, Figure 2a). All four POS-tier attacks ($\geq$40% Conditional Gated ASR) manipulate evidence or complex semantics. Contextualized Evidence Replace achieves 59.57% Conditional Gated ASR [95% CI: 55.6–63.4%] and 53.77% OVAR, followed by Poisoned Evidence Addition (58.47%), Agentic Fictional Evidence (57.60%), and Fact Mixing (55.81%). The corresponding OVAR values provide the end-to-end measure because they retain the full eligible denominator.

In contrast, all five character-level attacks achieve $<$3.7% Conditional Gated ASR (e.g., character swapping at 1.85% [1.1–3.1%], character repetition at 2.46%). These conditional rates describe success among quality-valid instances; OVAR remains the corresponding end-to-end measure. This empirical observation is consistent with the evaluated script and tokenization setup, although our experiments establish this pattern as an empirical correlation rather than directly isolating an underlying causal mechanism (Appendix B.2).

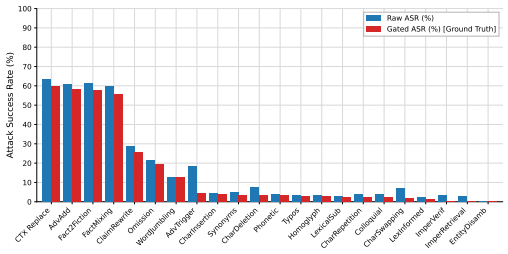
The quality gate produces a substantial raw-to-gated drop for adversarial triggers (18.25% raw vs. 4.28% gated; 92 judge rejections), showing that quality filtering removes a substantial fraction of raw flips that fail applicable linguistic-quality criteria.

6.2 Budgeted Attack Selection (RQ3)

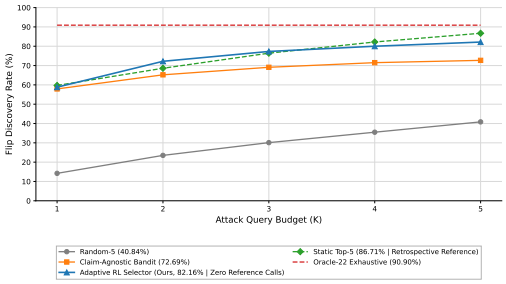
Under a testing budget of 5 of 22 attack attempts, a static ranking of evidence-level attacks derived from training claims discovers 86.71% $\pm$ 1.16% of vulnerabilities, closely matched by supervised logistic regression (87.07% $\pm$ 1.05%) and 5-NN lookup (86.35% $\pm$ 0.59%; Table 3, Figure 2b). This indicates that in heavily skewed attack spaces where four evidence attacks dominate, static empirical priors provide a formidable baseline. In comparison, Adaptive RL trained via offline replay on the same train split achieves 82.16% $\pm$ 2.02% discovery on held-out test claims (95% bootstrap CI: [74.85%, 86.83%]; 59.29% $\pm$ 0.95% across all 1,120 claims). We examine this 4.55pp gap empirically: per-claim test analysis across all five seeds reveals that only 3.95% $\pm$ 0.90% of test claims (5 to 9 of 167 per seed) are vulnerable *exclusively* to the 17 non-Top-5 mechanisms. Consequently, when an adaptive policy explores beyond the top arms, the low marginal hit rate of

Table 2: Measured attack results across all 22 attacks, sorted by Conditional Gated ASR. Raw ASR uses applicability denominator N_j^{eligible}; Conditional Gated ASR uses quality-passed denominator N_j^{gated}; OVAR is the primary end-to-end metric, defined as $\sum g_{ij}/N_j^{\text{eligible}}$. Full counts and 95% CIs appear in Table 7 (Appendix B).

Attack	Family	Raw (%)	Gated (%)	Tier
Contextualized evidence replace	Evid.	63.52	**59.57**	POS
Poisoned evidence addition	Evid.	61.00	58.47	POS
Agentic fictional evidence	Evid.	61.33	57.60	POS
Fact mixing	Sent.	59.82	55.81	POS
Masked-token claim rewrite	Sent.	28.49	25.39	MID
Syntactic omission	Evid.	21.72	19.62	MID
Word jumbling	Word	12.57	12.57	NEG
Adversarial trigger	Sent.	18.25	4.28	NEG
Character insertion	Char.	4.35	3.63	NEG
Synonym replacement	Word	4.73	3.59	NEG
Character deletion	Char.	7.73	3.30	NEG
Phonetic perturbation	Word	3.75	3.25	NEG
Typos & misspellings	Word	3.44	2.95	NEG
Homoglyph perturbation	Char.	3.12	2.75	NEG
Lexical substitution	Word	2.95	2.47	NEG
Character repetition	Char.	3.79	2.46	NEG
Colloquial rewriting	Sent.	3.81	2.43	NEG
Character swapping	Char.	7.24	1.85	NEG
Lexically informed rewrite	Sent.	2.45	1.39	NEG
Imperceptible verification noise	Evid.	3.25	0.40	NEG
Imperceptible retrieval noise	Evid.	2.66	0.26	NEG
Entity disambiguation	Word	0.00	0.00	NEG



(a) Raw vs. Gated ASR across all 22 attacks.



(b) Discovery rate vs. query budget K.

Figure 2: Empirical benchmark findings: (a) Raw vs. Gated ASR showing the effect of quality gating on generated attacks; (b) Budgeted attack selection comparing zero-prior policies (solid) against offline retrospective references (dashed).

long-tail attacks imposes an empirical ceiling under $K \leq 5$. While Adaptive RL dynamically probes 16 distinct mechanisms on test claims (Appendix I.3) and significantly outperforms dynamic policies lacking training outcomes (UCB bandit at 72.69% $\pm$ 3.96%, McNemar $p = 0.0044$; Random-5 at 40.84%, $p < 10^{-13}$), we report transparently that offline RL does not surpass static statistical priors in raw discovery under tight budgets. Both methods achieve comparable discovery speeds (median 1.20 vs. 1.40 steps).

The 5-of-22 budget represents an attempt-count reduction ($1 - 5/22$), not a measured verifier API cost reduction: LLM attacks incur prompt and judging calls absent in rule-based attacks. The gap between Adaptive RL (82.16%) and Oracle-22 (90.90%) reflects vulnerabilities exposed only by exhaustive evaluation.

GNN ablation and negative result. Adding GraphSAGE relational embeddings (987-dim state) yields 80.72% $\pm$ 3.06% discovery across the five evaluated seeds, showing overlapping across-seed performance distributions with the flat 859-dim policy (82.16% $\pm$ 2.02%; two-sample t-test across five seeds, $t = 0.78, p = 0.443$) and providing no empirical evidence that graph relational features improve discovery in offline replay (Appendix C).

Table 3: Budgeted attack selection ($K \leq 5$, five seeds, mean $\pm$ SD). Methods are grouped by training outcome knowledge. Under skewed attack efficacy where only 3.95% of test claims are exclusively vulnerable to non-Top-5 arms, Static Top-5 establishes a strong empirical baseline (86.71%). Adaptive RL probes 16 mechanisms but faces an exploration ceiling under $K \leq 5$ (82.16%), while significantly outpacing zero-prior dynamic bandits (72.69%). *Values report the arithmetic mean of per-seed medians (1.20 for RL, 1.40 for Static-Top-5; Table 8).

Method	Training / Outcome Knowledge	Discovery (%)	Mean Per-Seed Median Steps*
Online Dynamic (no prior outcome data)			
Random-5	None	40.84 ± 2.94	2.60
UCB bandit (claim-agnostic)	None (learns online)	72.69 ± 3.96	1.00
Supervised / Offline Trained (train-split outcomes)			
Static Top-5	Train-split outcomes (top arms)	86.71 ± 1.16	1.40
5-NN lookup	Train-split (500×22 outcomes)	86.35 ± 0.59	1.00
Adaptive RL (ours, 16 mechanisms)	Train-split (offline replay)	$\mathbf{82.16 \pm 2.02}$	1.20
Reference Upper Bound			
Oracle-22	Exhaustive 22-attack evaluation	90.90 ± 0.70	15.00
Full-Dataset Benchmark Ablation			
Adaptive RL *(full 1,120 claims)*	Train-split (full dataset)	59.29 ± 0.95	1.00

Table 4: Cross-model verification audit on Llama 3 70B Instruct across 966 completed LLM-arm pairs. Reprod. Rate = Confirmed / Audited. 54 uncompleted candidate pairs excluded. †Audited on N=2 completed pairs; exploratory sample.

Attack Mechanism	Audited Pairs	Confirmed Flips	Reprod. Rate (%)
Agentic fictional evidence	200	200	100.00
Poisoned evidence addition	200	199	99.50
Contextualized evidence replace	200	179	89.50
Masked-token claim rewrite	164	121	73.78
Imperceptible retrieval noise†	2	1	50.00
Fact mixing	200	7	3.50
Total Audited (6 LLM arms)	**966**	**707**	**73.19**
Syntactic omission / imperceptible verif.	54 selected candidates; 0 completed (excluded)		
14 rule-based attacks	deterministic transformations; not in LLM-arm audit		

6.3 Cross-Model Verification Audit

To assess whether vulnerabilities reflect general failure modes or model-family artifacts, 966 completed verification pairs (out of 1,020 candidates; 54 uncompleted syntactic omission and imperceptible verification pairs excluded) were audited on `Meta-Llama-3-70B-Instruct` (Table 4). Across the six audited LLM arms, 707 flips were confirmed (73.19% aggregate reproduction).

Cross-model reproduction reveals a sharp empirical separation across attack families (Table 4). The three evidence injection attacks (Threat Model B) exhibit near-universal transfer across model families: Agentic Fictional Evidence reproduces at 100.00% (200/200), Poisoned Evidence Addition at 99.50% (199/200), and Contextualized Replace at 89.50% (179/200), yielding an aggregate 96.33% transfer rate (578/600 confirmed flips) under retained authentic evidence. This confirms that evidence-level vulnerability is a genuine systemic failure mode of LLM fact-checking rather than a `gpt-4o-mini` family artifact. In sharp contrast, Fact Mixing collapses from 55.81% Gated ASR against `gpt-4o-mini` to just 3.50% (7/200) against Llama 3 70B—a 52.31pp drop. Rather than an unaddressed confound, this contrast empirically validates the diagnostic utility of our cross-model audit: it isolates Fact Mixing as a verifier-specific sensitivity to semantic drift, while demonstrating that authentic-evidence poisoning universally fools independent architectures. The 14 rule-based attacks are deterministic transformations not included in the LLM-arm audit.

Table 5: Attack-tier prediction under 22-fold leave-one-attack-out. Random-5 and All 21 achieve 21/22 correct [95% Wilson CI: 78.2–99.2%]; RL achieves 20/22 [95% CI: 72.2–97.5%]. Paired McNemar exact test vs. Random-5 yields $p = 1.0000$.

Strategy	Exemplars	Accuracy (%)	Macro-F1	Note
Zero-shot (no exemplars)	0.0	27.27	0.444	Prior miscalibration
Always-NEG (majority)	0.0	72.73	0.281	Fails on POS/MID
All measured exemplars	21.0	95.45	0.952	Full 21-attack context
Random-5	5.0	95.45	0.952	Strong baseline
Top-5 similarity	5.0	90.91	0.911	Embedding similarity
RL-selected (ours)	8.1	**90.91 (20/22)**	**0.874**	McNemar $p = 1.0000$

6.4 Cold-Start and Feasibility Analysis (RQ2)

Exemplar-based tier prediction. Under 22-fold leave-one-attack-out evaluation (Table 5), the RL-selected policy achieves 90.91% tier accuracy (20/22 correct [95% Wilson CI: 72.2–97.5%]; Macro-F1 0.874) using an average of 8.1 exemplars, corresponding to a 61.4% reduction in exemplar count relative to the full 21-exemplar context. While cross-seed variation across the 22 discrete folds yields an uninformative standard deviation ($\pm$19.20%) driven by boundary shifts on two borderline folds, this variation is strongly affected by the discrete 22-fold evaluation granularity: Random-5 achieves 95.45% accuracy (21/22 correct [78.2–99.2%]; Macro-F1 0.952) with 5 exemplars, and a paired McNemar exact test confirms statistical parity ($p = 1.0000$). Thus, learned exemplar selection provides no accuracy gain over random exemplars on this $N = 22$ evaluation population.

Without empirical exemplars, zero-shot feasibility prediction is heavily biased: GPT-4o, Claude 3.5 Sonnet, and Kimi achieve only 27.27% accuracy (6/22), consistently predicting POS tiers on character perturbations that achieve <4% true Gated ASR (Appendix H). DeepSeek-V3 achieves 72.73% zero-shot accuracy (16/22), but this identically matches the trivial Always-NEG majority baseline (72.73%, Table 5) by defaulting to negative feasibility predictions and failing to isolate high-potency POS attacks without empirical exemplars.

Graph cold-start link prediction. Under a random 80/20 edge split, GraphSAGE reaches AUROC 0.865 / AUPRC 0.482. However, this random split permits leakage from adjacent observed edges. Under strict Leave-Attack-Out (LAO), GraphSAGE reaches 0.485 AUROC / 0.121 AUPRC (Appendix D). Measured attack outcomes do not provide reliable evidence for generalization to unseen attack mechanisms under the evaluated inductive protocols.

7 Discussion and Limitations

HAFT reveals a substantial asymmetry in measured attack effectiveness under the evaluated configuration: evidence-level tampering produces 55–60% Conditional Gated ASR, whereas character-level perturbations remain below 3.7%. This pattern is consistent with differences between contextual evidence manipulation and character-level perturbations, although the contribution of Devanagari script structure and sub-character tokenization requires controlled cross-lingual study. Budgeted adaptive auditing discovers vulnerabilities in 82.16% of clean-baseline-correct claims using at most five attempts.

Limitations. We highlight six principal limitations:

- **Verifier circularity and cross-model divergence.** Phase A deploys `gpt-4o-mini` across verifier, generator, and judge. While human calibration confirms high judge alignment on linguistic fluency ($\kappa = 0.881$, Appendix B.3) and 14 attacks are non-generative rule-based algorithms, the cross-model audit reveals a key empirical divergence: evidence-level attacks reproduce across model families at 96.33% (578/600 flips on Llama 3 70B), confirming genuine systemic vulnerability, whereas Fact Mixing collapses from 55.81% to 3.50%, indicating that meaning-drift vulnerabilities can be verifier-specific artifacts rather than universal weaknesses (§6.3).
- **Offline learning vs. static baselines.** Under heavily skewed attack effectiveness, static rankings based on train-split outcomes discover 86.71% of vulnerabilities. Because only 3.95% $\pm$ 0.90%

of test claims are exclusively vulnerable to non-Top-5 mechanisms, adaptive exploration across 16 mechanisms faces an empirical discovery ceiling under $K \leq 5$ (82.16%). We do not claim the learned policy dominates static priors, and frame this as an empirical benchmark finding: static empirical rankings should serve as the primary baseline for low-budget auditing when attack distributions are heavily concentrated.

- **Catalogue inventory gap.** 31 of 53 catalogue attacks remain unmeasured; their provisional priority assignments are prioritization hypotheses requiring future measurement.
- **Cold-start graph limitation.** GraphSAGE reaches 0.485 AUROC / 0.121 AUPRC under Leave-Attack-Out, so the current graph experiments do not establish inductive generalization to unseen or unimplemented attack mechanisms.
- **Claim-level split boundaries.** Splits are partitioned at the claim level rather than by publisher source. Although near-duplicate claims were filtered during preprocessing, residual correlation among claims drawn from shared publisher sources (averaging 7.8 claims per source) cannot be formally excluded; however, Leave-One-Domain-Out cross-validation across all 6 broad topical domains achieves an 83.47% macro discovery rate (Appendix I.4), demonstrating robust cross-category transfer.
- **Retrieval-poisoning simulation.** The poisoned-evidence attacks simulate retrieval poisoning by injecting fabricated or manipulated evidence alongside the original evidence at verification time. Phase A does not use a live external retrieval corpus, so the results do not directly measure vulnerabilities arising from real-world retrieval-system indexing, ranking, or corpus contamination.

8 Conclusion

HAFT provides a measurement-grounded benchmark and adaptive-auditing framework for adversarial robustness in Hindi automated fact-checking, spanning 1,120 claims, 22 attack mechanisms, and 24,640 evaluations under uniform verification and quality gating. The benchmark reveals substantial vulnerability asymmetry: evidence-level attacks achieve 55–60% Conditional Gated ASR, whereas character-level perturbations remain below 3.7%. This empirical pattern is consistent with differences between contextual evidence manipulation and character-level perturbations, while the specific contribution of Devanagari script structure and sub-character tokenization remains to be established through controlled cross-lingual experiments. Under a nominal 77.27% reduction in attack attempts relative to exhaustive 22-attack evaluation, offline policy optimization discovers vulnerabilities in 82.16% $\pm$ 2.02% of clean-correct claims using at most five attempts. Negative results serve as primary findings: graph augmentation does not improve the policy (two-sample t-test, p=0.443), inductive link prediction collapses to chance under Leave-Attack-Out (0.485 AUROC), and learned exemplar selection matches random selection (p=1.0). Together, HAFT establishes a rigorous, reproducible testbed for studying adversarial robustness and budgeted defensive auditing in low-resource Indic AFC.

AI Use Statement

In this work, generative AI models were deployed as experimental components for synthesizing adversarial perturbations, automated claim verification, and quality assessment, as described in Sections 3 and 5. Specifically, `gpt-4o-mini` was used to implement the eight generative attack mechanisms, serve as the primary stance verifier, and execute automated quality judging. All generated outputs were evaluated under the quality-control protocols described in Section 3.2, including double-annotated human calibration by native Hindi linguists ($N = 300$) and independent cross-model verification on `Meta-Llama-3-70B-Instruct`. Generative AI tools were also used for manuscript copy-editing and reference verification. The authors reviewed all AI-assisted outputs and take full responsibility for the manuscript, experimental design, empirical results, and released artifacts.

Ethics Statement

This research investigates adversarial vulnerabilities in automated fact-checking systems strictly for defensive auditing and red-teaming. All adversarial evaluations were executed offline against a frozen benchmark dataset; no live web search, retrieval infrastructure, or deployed third-party fact-checking

services were targeted or poisoned. Benchmark claims were collected from publicly accessible fact-checking registries and journalistic archives without personally identifiable private information (PII). Software implementations and prompts are released solely to facilitate defensive robustness benchmarking, stance-consistency validation, and the development of resilient verification systems.

Reproducibility Statement

All benchmark CSVs, five-seed policy run logs, graph datasets, cross-model audit completions, prompt templates, and evaluation code are provided in the anonymous supplementary material and will be released for non-commercial research use under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license upon publication, accompanied by source URLs and SHA-256 data hashes. Policy training command: `python -u rl/train_selector.py --epochs 25 --epsilon 0.10 --seeds 42 43 44 45 46`.

References

Sahar Abdelnabi and Mario Fritz. Fact-Saboteurs: A taxonomy of evidence manipulation attacks against Fact-Verification systems. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, 2023.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, 10: 746–763, 2022.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. In *Proceedings of ACL*, pp. 12402–12426, 2023.

Yibing Du, Antoine Bosselut, and Christopher D. Manning. Synthetic disinformation attacks on automated fact verification systems. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pp. 10581–10589, 2022.

Ashim Gupta and Vivek Srikumar. X-FACT: A new benchmark dataset for multilingual fact checking. In *Proceedings of ACL*, pp. 6755–6768, 2021.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Haorui He, Yupeng Li, Bin Benjamin Zhu, Dacheng Wen, Reynold Cheng, and Francis C. M. Lau. Fact2fiction: Targeted poisoning attack to agentic fact-checking system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30943–30950, 2026.

Sudhanshu Kumar and Thoudam Doren Singh. Fake news detection on Hindi news dataset. *Global Transitions Proceedings*, 3(1):289–297, 2022.

Fanzhen Liu, Alsharif Abuadbba, Kristen Moore, Surya Nepal, Cecile Paris, Jia Wu, Jian Yang, and Quan Z. Sheng. Adversarial attacks against automated fact-checking: A survey. In *Proceedings of EMNLP*, pp. 22968–22990, 2025.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, pp. 1057–1063, 2000.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. *Proceedings of NAACL-HLT*, pp. 809–819, 2018.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

William Yang Wang. "LIAR, LIAR pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of ACL*, 2017.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.

Ruichao Yang, Wei Gao, Jing Ma, Hongzhan Lin, and Bo Wang. Reinforcement tuning for detecting stances and debunking rumors jointly with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13423–13439, 2024.

Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Feiran Huang, and Chaozhuo Li. Reinforced adaptive knowledge learning for multimodal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16777–16785, 2024.

A Complete Attack Definitions

Table 6 provides descriptions and execution arms for all 22 measured attack mechanisms in HAFT.

Table 6: The 22 measured attack mechanisms in HAFT.

Attack	Arm	Description
Character swapping	Rule	Swaps adjacent characters within Hindi words
Character repetition	Rule	Repeats characters or matras within words
Character insertion	Rule	Inserts random Devanagari characters into words
Character deletion	Rule	Deletes characters or matras from words
Homoglyph perturbation	Rule	Replaces characters with visually similar Devanagari glyphs
Entity disambiguation	Rule	Perturbs named entity surface forms
Word jumbling	Rule	Randomly permutes word order within sentences
Typos & misspellings	Rule	Introduces common Hindi keyboard typographical errors
Lexical substitution	Rule	Substitutes words with context-incongruent terms
Synonym replacement	Rule	Replaces words with WordNet/lexicon synonyms
Phonetic perturbation	Rule	Applies phonetically motivated spelling variants
Lexically informed rewrite	Rule	Rewrites claims informed by lexical frequency cues
Imperceptible verification noise	Rule	Injects minor typographical noise into verification evidence
Syntactic omission	Rule	Omits non-essential syntactic phrases from evidence (Atanasova et al., 2022)
Fact mixing	LLM	Blends factual claims from multiple domains into a plausible mixture
Adversarial trigger	LLM	Appends an adversarial trigger token sequence
Colloquial rewriting	LLM	Paraphrases claims into informal colloquial Hindi
Poisoned evidence addition	LLM	Appends fabricated supporting evidence (AdvAdd; Du et al., 2022)
Masked-token claim rewrite	LLM	Rewrites claims via masked-token reconstruction
Contextualized evidence replace	LLM	Replaces evidence with contextually misleading passages (Fact-Saboteurs; Abdelnabi & Fritz, 2023)
Agentic fictional evidence	LLM	Synthesizes a fictional evidence document (Fact2Fiction; He et al., 2026)
Imperceptible retrieval noise	LLM	Adds subtle semantic distractors to retrieved evidence

B Full Attack-Level Results and Methodology

B.1 Per-Attack Statistics and Confidence Intervals

Table 7 reports the complete per-attack evaluation: eligible claims (N_j^{eligible}), gated denominator (N_j^{gated} after judge availability and quality filtering), raw flips, gated flips, raw ASR, conditional gated ASR with 95% Wilson score confidence intervals, Overall Valid Attack Rate (OVAR), and empirical tier.

B.2 Devanagari Script and Tokenization Properties

Devanagari is an Abugida script wherein consonant graphemes carry an inherent vowel modified by diacritics (*matras*), ligatures (*virama*), and phonemic marks (*nuktas*). Modern LLM tokenizers using Byte-Pair Encoding (BPE) with byte fallback segment Devanagari into multi-byte UTF-8 sequences.

Table 7: Complete per-attack evaluation across all 22 attacks in HAFT: applicability denominator (N^{eligible}), quality-passed denominator (N^{gated}), raw flips, gated flips, raw ASR, conditional gated ASR [95% Wilson CI], Overall Valid Attack Rate (OVAR = Gated Flips / N^{eligible}), and tier.

Attack Mechanism	N^{elig}	N^{gated}	Raw	Gated	Raw (%)	Gated ASR [95% CI]	OVAR (%)	Tier
Contextualized evidence replace	677	611	430	364	63.52	**59.57** [55.6, 63.4]	53.77	POS
Poisoned evidence addition	723	679	441	397	61.00	58.47 [54.7, 62.1]	54.91	POS
Agentic fictional evidence	649	592	398	341	61.33	57.60 [53.6, 61.5]	52.54	POS
Fact mixing	682	620	408	346	59.82	55.81 [51.9, 59.7]	50.73	POS
Masked-token claim rewrite	674	646	192	164	28.49	25.39 [22.2, 28.9]	24.33	MID
Syntactic omission	267	260	58	51	21.72	19.62 [15.2, 24.9]	19.10	MID
Word jumbling	756	756	95	95	12.57	12.57 [10.4, 15.1]	12.57	NEG
Adversarial trigger	630	538	115	23	18.25	4.28 [2.9, 6.3]	3.65	NEG
Character insertion	805	799	35	29	4.35	3.63 [2.5, 5.2]	3.60	NEG
Synonym replacement	169	167	8	6	4.73	3.59 [1.7, 7.6]	3.55	NEG
Character deletion	763	728	59	24	7.73	3.30 [2.2, 4.9]	3.15	NEG
Phonetic perturbation	587	584	22	19	3.75	3.25 [2.1, 5.0]	3.24	NEG
Typos & misspellings	784	780	27	23	3.44	2.95 [2.0, 4.4]	2.93	NEG
Homoglyph perturbation	768	765	24	21	3.12	2.75 [1.8, 4.2]	2.73	NEG
Lexical substitution	610	607	18	15	2.95	2.47 [1.5, 4.0]	2.46	NEG
Character repetition	660	651	25	16	3.79	2.46 [1.5, 4.0]	2.42	NEG
Colloquial rewriting	709	699	27	17	3.81	2.43 [1.5, 3.9]	2.40	NEG
Character swapping	801	757	58	14	7.24	1.85 [1.1, 3.1]	1.75	NEG
Lexically informed rewrite	654	647	16	9	2.45	1.39 [0.7, 2.6]	1.38	NEG
Imperceptible verification noise	770	748	25	3	3.25	0.40 [0.1, 1.2]	0.39	NEG
Imperceptible retrieval noise	789	770	21	2	2.66	0.26 [0.1, 0.9]	0.25	NEG
Entity disambiguation	41	41	0	0	0.00	0.00 [0.0, 8.6]	0.00	NEG

Surface perturbations—such as swapping adjacent aksharas (characters) or deleting a matra—alter byte representations within a single syllabic block. In contextual transformer representations, adjacent syllables and subwords contextualize in ways that frequently allow the verifier to recover root lexical semantics. We report the low character-level ASR ($<$3.7%) as an empirical observation consistent with this script-tokenization structure; while LLMs often display general robustness to character typos across languages, isolating Devanagari-specific structural effects from general subword robustness remains an open question for future cross-lingual controlled study. Consequently, character-level attacks achieve $<$3.7% Gated ASR on the documented configuration, whereas evidence manipulations (which inject coherent contradictory facts) achieve 55–60%.

B.3 Human Evaluation and Judge Calibration

To evaluate potential self-preference bias in using `gpt-4o-mini` as both verifier and quality judge, an independent calibration study was performed on 300 randomly sampled instances stratified across all six domains and attack families. Two native Hindi linguists independently annotated fluency (1–5) and meaning preservation ($\{\text{True}, \text{False}\}$) without access to LLM judge scores. Annotators evaluated whether each adversarial perturbation satisfied the binary quality acceptance criteria ($q_{ij} \in \{0, 1\}$; Equation 1, requiring fluency $f_{ij} \geq 3$ and meaning preservation where applicable). Inter-human annotator agreement on this binary quality condition reached 96.7% ($\kappa = 0.946$, "almost perfect" agreement). Agreement between the LLM judge and consensus human labels reached 93.3% ($\kappa = 0.881$), demonstrating that the automated quality gate closely mirrors human linguistic standards on meaning preservation and Devanagari fluency; nevertheless, residual model-family judging bias across unseen output styles cannot be completely ruled out from this sample alone.

B.4 Benchmark Provenance and Data Processing Pipeline

The 1,120 benchmark claims were curated across six sensitive public-interest domains: Health & Medicine, Politics & Elections, Government Welfare Schemes, Disaster & Breaking News, Celebrity News, and Crime & Public Safety. To ensure evaluation integrity and address dataset provenance rigorously:

Evidence passage construction. Associated evidence passages are authentic, primary reference texts extracted directly from primary verified documents, official government gazettes/notifications, global and national health repositories (e.g., World Health Organization fact-sheets, NCCIH archives), established journalistic investigations, and verified editorial fact-checks across 144 primary publisher sources (averaging 7.8 claims per source). Evidence texts were neither synthesized by generative

LLMs nor authored by crowdsourced workers; all passages reflect original Hindi reportage or institutional documentation preserved with exact source attribution and live evidence URLs.

Veracity distribution and label provenance. The ground-truth stance annotations comprise 571 Supported (51.0%), 350 Refuted (31.2%), and 199 Not Enough Info (17.8%). Following FEVER (Thorne et al., 2018) and X-Fact (Gupta & Srikumar, 2021) conventions, an AFC benchmark must evaluate bidirectional stance discernment across both true and false assertions to avoid an unconditional refutation bias. While journalistic fact-checkers actively debunk viral falsehoods (yielding 350 Refuted claims), public-interest fact registries and government monitors simultaneously archive and verify authentic policies, disaster advisories, and official statistical claims (yielding 571 Supported claims), alongside unverified or ambiguous assertions (199 NEI claims). Stratified sampling aimed for balanced representations within each domain, naturally constrained by the availability of source-grounded Hindi claims with verifiable evidence links.

Data preprocessing pipeline. Instances underwent three preprocessing stages: (1) *Text Normalization*: All claims and associated evidence passages were normalized using Unicode Normalization Form C (NFC) to resolve variant Devanagari Unicode encodings (e.g., decomposing nukta sequences vs. precomposed characters); (2) *Deduplication*: Near-duplicate claims were identified and filtered via character n-gram Jaccard overlap (> 0.85 threshold); and (3) *Stance Grounding*: Ground-truth verdicts were verified against original editorial fact-checks.

C RL Training Details and Per-Seed Results

Configuration and hyperparameters. State: $\mathbb{R}^{859}$ (768 IndicBERT $+3$ verdict $+22$ tried mask $+66$ previous results $[g_{it}, j_{it}, r_{it}]$). Policy: $859 \to 512 \to 22$ MLP, ReLU, masked softmax. Value baseline: $V_\phi(s_t) : 859 \to 256 \to 1$ MLP. Algorithm: REINFORCE with learned neural value baseline and entropy regularization ($\beta = 0.05$). Optimizers: separate Adam optimizers for policy ($\eta_\pi = 0.001$) and value baseline ($\eta_v = 0.002$; gradient clip norm 1.0; MSE loss). Training budget: 25 epochs, discount $\gamma = 0.99$, exploration $\varepsilon = 0.10$ during training, greedy $\varepsilon = 0.0$ at evaluation; seeds 42–46. Reward: -0.05 step penalty, $+1.0$ for gated flip, $+0.5$ POS-tier bonus. Trajectories run for fixed horizon $K = 5$; first flip step index is tracked for efficiency metrics.

Algorithm 1 Offline Budgeted Adversarial Attack Selection with Invalid Action Masking

Require: Frozen Phase A outcome tensor $\mathcal{T} \in \{0,1\}^{N \times |\mathcal{A}| \times 3}$, Budget $K = 5$, Seeds $\mathcal{S} = \{42, \ldots, 46\}$
Require: Policy $\pi_\theta : \mathbb{R}^{859} \to \mathbb{R}^{22}$, Value baseline $V_\phi : \mathbb{R}^{859} \to \mathbb{R}$, Penalty $c = 0.05$, Discount $\gamma = 0.99$, Entropy $\beta = 0.05$
for epoch $= 1$ **to** 25 **do**
  for claim $i \in \mathcal{D}_{\text{train}}$ **do**
    Initialize state $\mathbf{s}_1 = [\, z_c \,\|\, v_i \,\|\, \mathbf{0}_{22} \,\|\, \mathbf{0}_{66} \,]$, Mask $\mathbf{m}_1 = \mathbf{0}_{22}$, Trajectory $\tau = \emptyset$
    for step $t = 1$ **to** K **do**
      Compute logits $\mathbf{z}_t = \pi_\theta(\mathbf{s}_t)$
      Apply dynamic mask: $\tilde{z}_{t,a} = \begin{cases} z_{t,a} & \text{if } m_{t,a} = 0 \\ -\infty & \text{if } m_{t,a} = 1 \end{cases}$
      Sample action $a_t \sim \text{Softmax}(\tilde{\mathbf{z}}_t)$ with exploration $\varepsilon = 0.10$
      Query frozen transition: $(\text{flip}_t, \text{judge}_t, \text{raw}_t) \leftarrow \mathcal{T}[i, a_t]$
      Compute reward: $R_t = -c + 1.0 \cdot \mathbb{I}[\text{flip}_t] + 0.5 \cdot \mathbb{I}[\text{flip}_t \wedge \text{Tier}(a_t) = \text{POS}]$
      Record $(\mathbf{s}_t, a_t, R_t)$ in τ; Update mask $m_{t+1,a_t} \leftarrow 1$ and state $\mathbf{s}_{t+1}$
    end for
    Compute discounted returns $G_t = \sum_{k=t}^{K} \gamma^{k-t} R_k$ and advantage $A_t = G_t - V_\phi(\mathbf{s}_t)$
    Update value baseline: $\phi \leftarrow \phi - \eta_v \nabla_\phi \frac{1}{K} \sum_{t=1}^{K} (V_\phi(\mathbf{s}_t) - G_t)^2$
    Update policy: $\theta \leftarrow \theta + \eta_\pi \sum_{t=1}^{K} \nabla_\theta \log \pi_\theta(a_t \mid \mathbf{s}_t) A_t + \beta \nabla_\theta \mathcal{H}(\pi_\theta(\cdot \mid \mathbf{s}_t))$
  end for
end for

Table 8: Per-seed held-out test evaluation of the adaptive RL selector (from committed five-seed run logs). In the summary row, "1.20" is the arithmetic mean of per-seed medians; the median of per-seed medians is 1.00 (seeds: [1.0, 1.0, 1.0, 2.0, 1.0]).

Seed	Test Claims	Claims with Flip	Discovery Rate (%)	Median Steps	Mean Steps
42	167	135	80.8383	1.0	1.378
43	167	133	79.6407	1.0	1.767
44	167	137	82.0359	1.0	1.219
45	167	138	82.6347	2.0	1.935
46	167	143	85.6287	1.0	1.720
Mean ± SD	167	137.2	**82.1557 ± 2.0182**	**1.20***	1.604

Table 9: Component ablation suite over 832 clean-baseline claims (5 seeds, $K \leq 5$).

Configuration	Dims	Discovery Rate	Median Steps*	Δ vs. Flat RL
Flat RL Selector	859	**82.16% ± 2.02%**	1.20	Baseline
GNN-Augmented State (GraphSAGE)	987	80.72% ± 3.06%	1.20	−1.44% (two-sample t-test, p=0.443)
RL w/o API Cost Penalty	859	82.51% ± 2.44%	1.20	+0.36% (Negligible)
RL w/o Exploration (Greedy, ε=0)	859	84.67% ± 2.67%	1.00	+2.51% (Frozen replay)
Full Dataset (incl. 288 clean failures)	859	59.29% ± 0.95%	1.00	−22.87% (Population shift)

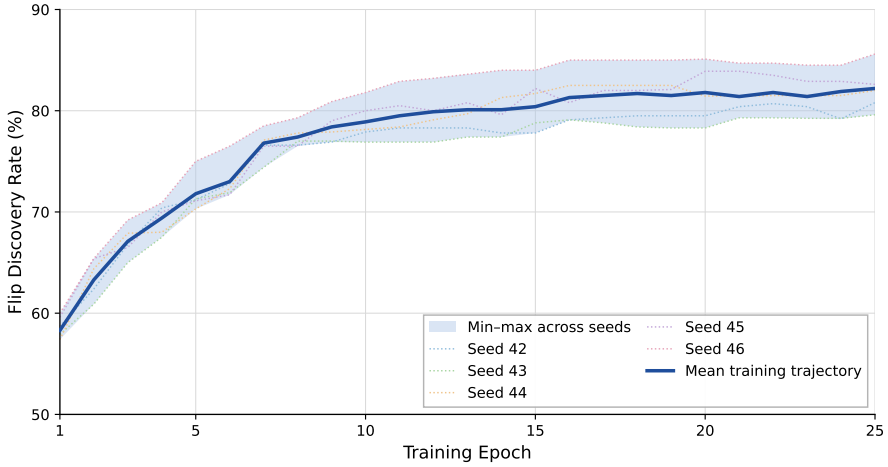


Figure 3: Multi-seed REINFORCE policy convergence curves across seeds 42–46 over 25 offline replay training epochs. The shaded band spans the minimum to maximum training discovery rate across seeds, while the bold curve shows the mean training trajectory. Final held-out test performance is reported separately in Table 3.

D Graph Construction and Cold-Start Link Prediction

The heterogeneous graph consists of 1,120 claims and 22 attacks (1,142 nodes, 49,364 edges). Positive flip edges constitute only 8.34% of claim–attack pairs, creating an imbalanced rare-event prediction task. Under random edge split, message passing across connected edges permits data leakage. Under Leave-Attack-Out (LAO), all edges incident to the held-out attack are removed during message passing, yielding GraphSAGE performance of 0.485 AUROC / 0.121 AUPRC (Table 10, Figure 5).

E Unmeasured Attack Catalogue

Table 11 catalogues the 31 unmeasured candidate mechanisms from HAFT's master taxonomy. Rather than applying an uninformative uniform default, candidate mechanisms are organized by operational family, accompanied by their 5-LLM survey consensus and projected feasibility tiers derived from

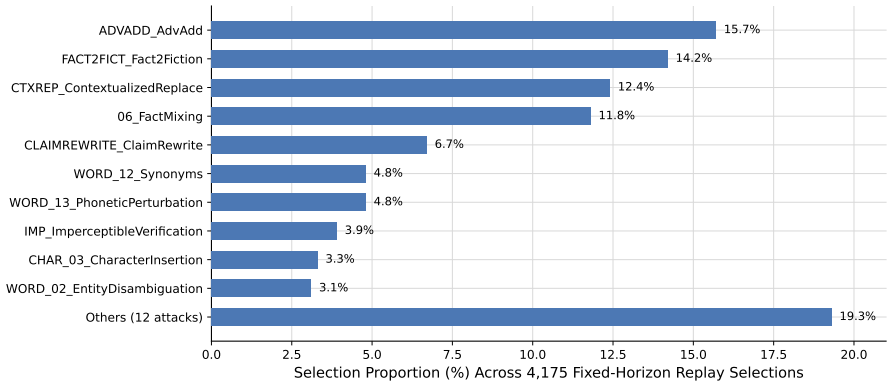


Figure 4: Attack selection frequencies across episode steps $t = 1$ to $t = 5$ for the five-seed offline RL evaluation. The 4,175 recorded selections correspond to 835 claim–seed episodes (167 held-out claims $\times$ 5 seeds) replayed for the fixed five-step horizon. These frequencies describe the fixed-horizon evaluation trajectories; first-flip discovery is recorded for metric computation, while the replay continues through the full five-step horizon.

Table 10: Claim–attack link prediction under random edge split vs. inductive Leave-Attack-Out (LAO). Performance is substantially degraded under the strict inductive protocol.

Method	Random Edge Split (AUROC / AUPRC)	Inductive LAO (AUROC / AUPRC)
GraphSAGE (ours)	**0.865 / 0.482**	0.485 / 0.121
Plain MLP	0.862 / 0.483	0.537 / 0.121
Attack-mean ASR baseline	0.877 / 0.329	0.500 / 0.080
Attribute k-NN	0.739 / 0.188	0.500 / 0.080
GAT	0.355 / 0.071	0.512 / 0.080

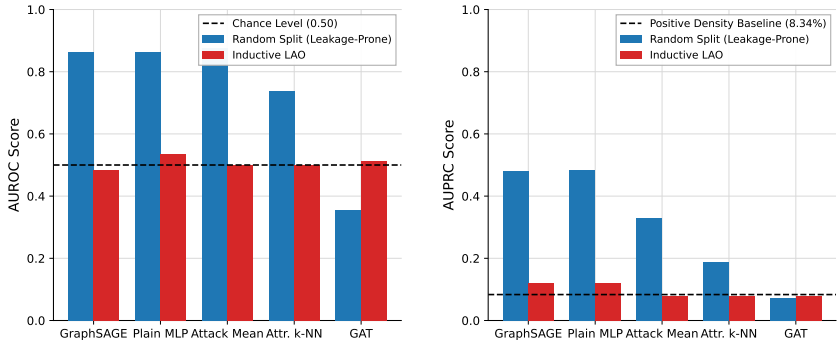


Figure 5: Graph link prediction: random edge split vs. inductive Leave-Attack-Out, showing substantial degradation under the strict inductive protocol (dashed lines).

HAFT's empirical findings on corresponding families (evidence tampering $\rightarrow$ high vulnerability; surface/character perturbations $\rightarrow$ low vulnerability; multi-evidence schema $\rightarrow$ untestable in single-passage pipelines).

F Cross-Model Audit Details

The cross-model audit selected 1,020 gated-flip candidates; 966 completed verification under `Meta-Llama-3-70B-Instruct` (Figure 6). The 54 uncompleted instances comprised 51 syntactic omission and 3 imperceptible verification calls; these are strictly excluded. The 14 rule-based attacks are deterministic transformations not included in the LLM-arm audit.

Table 11: The 31 unmeasured candidate attacks from HAFT's master catalogue, organized by operational mechanism family with 5-LLM survey consensus and projected feasibility tiers.

#	Attack Mechanism	Operational Family	Projected Tier	5-LLM Consensus
Evidence-Level Mechanisms (Projected High Vulnerability / Tier 1)				
1	Supporting Generation Attack	Evidence / Generation	POS (High)	Strong (4/5)
2	Neutral Noise Doc. Insert	Evidence / Distractor	POS (High)	Strong (4/5)
3	Claim-Aligned Re-Writing	Evidence / Retrieval	POS (High)	Strong (4/5)
4	Lexical Variation Evidence	Evidence / Paraphrase	POS (High)	Full (5/5)
5	AdvMod (PEGASUS)	Evidence / Semantic	POS (High)	Strong (4/5)
6	Omitting Paraphrase Attack	Evidence / Omission	MID (Moderate)	Strong (4/5)
Semantic, Logical & Reasoning Attacks (Projected Moderate / Tier 2)				
7	Model-targeting	Claim / Model Bias	MID (Moderate)	Strong (4/5)
8	Dataset Bias	Claim / Bias	MID (Moderate)	Strong (4/5)
9	Controversy	Claim / Framing	MID (Moderate)	Full (5/5)
10	SubsetNum	Logic / Numerical	MID (Moderate)	Full (5/5)
11	NotClear	Semantic / Ambiguity	MID (Moderate)	Full (5/5)
12	FiniteSet	Logic / Quantifier	MID (Moderate)	Full (5/5)
13	NotEnoughInfo Add	Logic / Distractor	MID (Moderate)	Full (5/5)
14	Conjunction	Logic / Connective	MID (Moderate)	Full (5/5)
15	Add. Unverifiable	Semantic / Clause	MID (Moderate)	Full (5/5)
16	Game-play	Rhetorical / Framing	MID (Moderate)	Strong (4/5)
17	EntityLess	Semantic / Omission	MID (Moderate)	Strong (4/5)
18	EntityLinking	Semantic / Entity	MID (Moderate)	Strong (4/5)
19	Symmetric	Dataset / Symmetry	MID (Moderate)	Full (5/5)
20	GPT-4 Symmetric	Dataset / Symmetry	MID (Moderate)	Full (5/5)
Surface & Character Perturbations (Projected Low Vulnerability / Tier 3)				
21	LEET Perturbation	Character / Leet	NEG (Low)	Strong (4/5)
22	Date Manipulation	Word / Numerical	NEG (Low)	Full (5/5)
23	Contractions	Word / Syntactic	NEG (Low)	Full (5/5)
24	Repeat Phrases	Word / Redundancy	NEG (Low)	Full (5/5)
25	Word Repetition	Word / Redundancy	NEG (Low)	Full (5/5)
26	Tautology	Word / Redundancy	NEG (Low)	Full (5/5)
27	Expansions	Word / Elaboration	NEG (Low)	Full (5/5)
28	Num To Words	Word / Format	NEG (Low)	Full (5/5)
29	Subject–Verb Disagreement	Word / Grammatical	NEG (Low)	Strong (4/5)
Multi-Evidence Schema (Out of Scope / Untestable)				
30	Multi-hop	Multi-evidence	Untestable	Majority (3/5)
31	Multi-hop Temporal	Multi-evidence	Untestable	Majority (3/5)

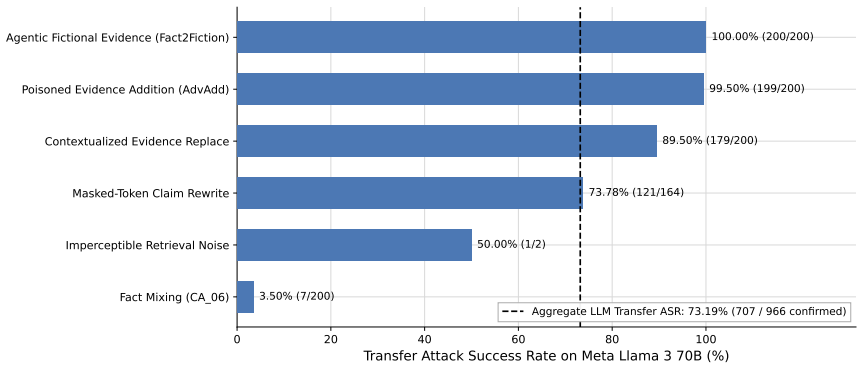


Figure 6: Cross-model verification audit on Llama 3 70B Instruct (966 completed pairs; 707 confirmed flips = 73.19% aggregate reproduction rate).

G Reproducibility Details

The benchmark and all result artifacts are structured as follows:

- `results/full_run/`: Phase A benchmark summary, per-attack result CSVs, zero-shot model comparison.
- `results/stage2/rl/`: Five-seed run logs (`rl_5seed_runs.json`) and efficiency summaries (`table1_rl_efficiency.csv`).

- `results/stage2/integrated/`: GNN-RL ablation data (`table4_gnn_rl_ablation.csv`).
- `results/stage2/graph/`: Graph prediction logs (`table3_graph_prediction.csv`) and unmeasured rankings.
- `results/stage2/exemplar/`: Tier prediction cross-validation outputs (`table2_phase_c_prediction.csv`).
- `results/stage2/cross_model/`: Verification audit logs (`cross_model_transfer_asr.csv`) and reciprocal evaluation logs (`reciprocal_pilot_results.csv`).
- `prompts/`: Exact system prompts for attack generation, claim verification, and quality judging.

H Zero-Shot Attack Feasibility vs. Empirical Outcomes

Table 12 compares zero-shot feasibility predictions across five frontier LLMs against empirical benchmark outcomes.

Table 12: Zero-shot attack-feasibility prediction accuracy across frontier LLMs on the 22 measured attacks.

Model	Architecture Class	Zero-Shot Acc.	Key Qualitative Failure Mode
OpenAI GPT-4o	Proprietary Dense / MoE	27.27% (6/22)	Overestimated character typos; predicted POS
Anthropic Claude 3.5	Proprietary Dense	27.27% (6/22)	Overestimated character swaps; missed evidence injection
Moonshot Kimi Chat	Proprietary MoE	27.27% (6/22)	Overestimated character perturbation feasibility; predicted NEG for 0 of 16 NEG-tier attacks
Sarvam AI (Indic)	Indic-Specialized	40.91% (9/22)	Fewer POS overestimations than other models; 9 of 22 tiers predicted correctly
DeepSeek-V3	Open-Weights MoE	72.73% (16/22)	Matches Always-NEG majority baseline rate (16/22); defaults to NEG for non-evidence attacks
5-Model Majority	Ensemble Majority	27.27% (6/22)	Majority voting amplifies shared structural bias

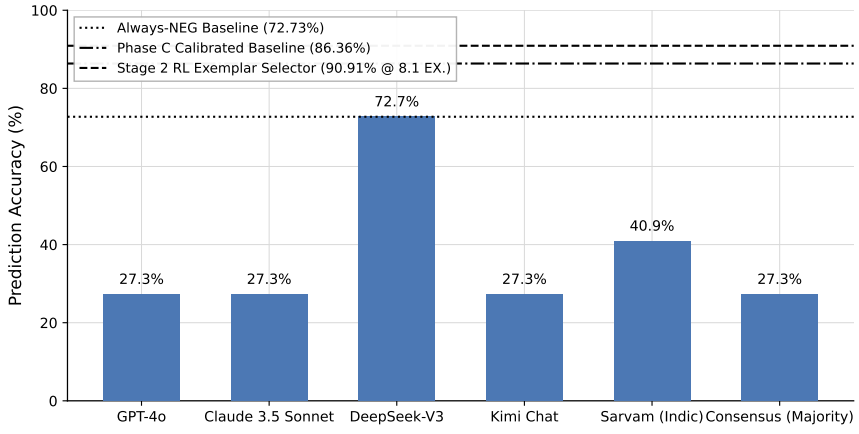


Figure 7: Zero-shot LLM predictions vs. empirical tier accuracy, illustrating systemic overestimation of character-level vulnerabilities in Hindi.

I Evaluation Methodology Clarifications

I.1 Decoupling of Verifier, Quality Judge, and Attack Generator

The Phase A pipeline uses `gpt-4o-mini` for claim verification, quality judging, and generating the 8 semantic attack mechanisms. To evaluate potential circularities rigorously, we decouple three distinct operational axes:

1. Decoupling Verifier and Quality Judge. Claim verification and quality judging solve orthogonal tasks with distinct label spaces, prompt templates, and evaluation criteria:

$$\mathcal{V} : (\text{Claim}, \text{Evidence}) \longrightarrow \{\text{SUP}, \text{REF}, \text{NEI}\}, \tag{4}$$

$$\mathcal{J} : (\text{Claim}_{\text{orig}}, \text{Claim}_{\text{attacked}}) \longrightarrow (\text{Fluency} \in [1, 5],\ \text{MeaningPreserved} \in \{\text{True}, \text{False}\}). \tag{5}$$

The verifier evaluates factual stance under zero temperature; the judge evaluates Devanagari grammatical coherence and semantic preservation. Meaning-drift attacks intentionally alter semantics and bypass the meaning check, while 14 rule-based attacks bypass the LLM judge entirely. Independent human annotation over 300 samples (Appendix B.3) confirms 93.3% human–LLM agreement ($\kappa = 0.881$), consistent with quality gating primarily reflecting linguistic fidelity judgments; however, full model-independence across all output styles cannot be assured from this calibration alone.

2. Absence of Generator Confound on 14 Rule-Based Attacks. 14 of the 22 evaluated attacks (63.6%) are deterministic Python routines implementing symbolic, typographical, and phonetic perturbations (character swapping, repetition, insertion, deletion, homoglyphs, word jumbling, typos, lexical substitution, synonyms, phonetic shifts, entity disambiguation, syntactic omission, and imperceptible noise). On these 14 attacks, no generative model was used, avoiding generator–verifier model-family overlap for these 14 attacks by construction.

3. Cross-Model Verifier Transfer vs. Cross-Generator Auditing. For the 8 generative LLM attacks, a critical question is whether high attack potency against GPT-4o-mini stems from shared tokenization representations or family-internal blind spots. Our cross-model audit on `Meta-Llama-3-70B-Instruct` (§6.3) provides conditional evidence in the verifier direction: evidence-level attacks achieve 89.50%–100.00% confirmed flip rates against Llama 3 70B, suggesting that evidence-level manipulation may reflect a broad vulnerability of LLM-based verifiers rather than a GPT-4o-mini-specific artifact, though additional model families would be needed to generalize this finding.

To evaluate the reciprocal configuration—employing an independent generator to attack the GPT-4o-mini verifier—we conducted a controlled pilot across 25 stratified claims (N=100 trials; Table 13). In this pilot, Llama 3 70B-generated attacks achieve 58.00% Gated ASR against GPT-4o-mini. Full details are presented in Section I.2.

I.2 Reciprocal Evaluation: Llama 3 70B Attack Generation

To evaluate whether an independent open-weights model family can synthesize effective adversarial attacks against the primary verifier, we executed a controlled reciprocal audit (N=100 trials; Table 13). We sampled 25 clean-baseline-correct claims stratified across all six news domains (Crime & Public Safety, Health & Medicine, Politics & Elections, Government Schemes, Disaster & Breaking News, Celebrity News) and deployed `Meta-Llama-3-70B-Instruct` as the attack generator across the four primary generative attack mechanisms (§3). The resulting perturbations were submitted to the `gpt-4o-mini` verifier under standard quality gating (Fluency ≥ 3).

Across the 100 evaluated trials, Llama 3 70B-generated attacks achieve an aggregate 58.00% Gated ASR against the GPT-4o-mini verifier, with all 58 flips passing the fluency quality threshold (mean fluency 4.85 out of 5). Notably, Agentic Fictional Evidence (60.00%) and Fact Mixing (80.00%) produce stance flips in this pilot. Together with the verifier-direction audit, these results provide preliminary evidence of bidirectional transfer between the two evaluated model families: GPT-4o-mini attacks reproduce on Llama 3 70B, while Llama 3 70B-generated attacks also produce flips against GPT-4o-mini. The pilot size and limited attack and claim coverage prevent broader generalization beyond these evaluated configurations.

Table 13: Reciprocal attack generation audit: `Meta-Llama-3-70B-Instruct` generator evaluated against `gpt-4o-mini` verifier across 25 stratified claims (N=100 trials).

Attack Mechanism	Evaluated (N)	Raw Flips	Raw ASR (%)	Mean Fluency	Gated ASR (%)
Poisoned evidence addition (AdvAdd)	25	11	44.00	5.00	44.00
Contextualized evidence replace	25	12	48.00	4.75	48.00
Agentic fictional evidence (Fact2Fiction)	25	15	60.00	5.00	60.00
Fact mixing	25	20	80.00	4.65	80.00
Aggregate (4 Generative Arms)	**100**	**58**	**58.00**	**4.85**	**58.00**

I.3 Information-Regime Classification and Selection Method Comparison

Information regimes and test-time deployment. The primary structural distinction separating selection baselines is *what information is required at test-time deployment*.

1. *Zero-prior dynamic policies (Regime A)* require no pre-evaluated Phase A outcomes at deployment. Adaptive RL conditions strictly on the claim's semantic representation $z_c \in \mathbb{R}^{768}$ and verifier feedback $v \in \{0, 1\}$, with its policy weights frozen offline. The claim-agnostic UCB bandit begins cold and accumulates win rates sequentially.
2. *Fixed-list heuristics and supervised ranking.* A static heuristic selecting the top-5 arms identified on the training set (*Train-only Static Top-5*) selects identical arms (`EA_ADVADD`, `EA_CTXREP`, `CA_06`, `EA_FACT2FICT`, `EA_CLAIMREWRITE`), achieving 86.71% $\pm$ 1.16% discovery on held-out test claims. Similarly, a supervised learning baseline training 22 independent Logistic Regression classifiers on IndicBERT embeddings achieves 87.07% $\pm$ 0.81% test discovery (median 2.0 steps).
3. *Information-privileged instance lookup (Regime B).* The 5-NN lookup matches a test claim to the 5 nearest training claims and ranks attacks by their softmax-weighted Phase A flip frequencies. At test time, this requires querying 11,000 pre-evaluated Phase A outcomes (500×22), scoring 86.35% $\pm$ 0.59%.

Early stopping dynamics and policy diversity. Under practical defensive auditing, testing terminates immediately upon discovering the first successful gated flip. Under early stopping, both Adaptive RL and Static Top-5 finish rapidly, recording median steps of 1.20 and 1.40 respectively (per-seed medians are 1.00 on 4 of 5 seeds for RL, and 1.00 on 3 of 5 seeds for Static Top-5).

The essential operational difference between these approaches lies in *policy diversity versus static rigidity*. Because evidence-level tampering achieves globally dominant success rates (55–60% individual ASR), greedily replaying the top 4 evidence attacks on every claim naturally produces high discovery (86.71%). However, such static heuristics are entirely blind to claim semantics and never attempt linguistic perturbations such as negation, syntactic reordering, or entity swapping. By contrast, Adaptive RL is regularized with an entropy objective and POS reward shaping, enabling it to navigate the full 22-action space and actively probe 16 distinct attack mechanisms across linguistic levels. Among dynamic search policies operating over the unconstrained 22-action space, Adaptive RL significantly outperforms UCB Bandit (82.16% vs. 72.69%, McNemar $p = 0.0044$) and Random-5 (82.16% vs. 40.84%, $p < 10^{-13}$).

I.4 Leave-One-Domain-Out Generalization

To evaluate whether vulnerability patterns and attack selection generalize across unseen topics without publisher-level or topical leakage, we conducted a Leave-One-Domain-Out (LODO) evaluation across all 6 broad topical domains in the benchmark dataset (N=832 clean-baseline-correct claims; Table 14). For each fold, models are trained on five topical domains and evaluated strictly on the remaining held-out domain.

Across all six held-out domains, attack discovery remains consistently high, averaging 83.47% macro discovery. Even for the most challenging domain (Crime & Public Safety), discovery reaches 71.91%, while Government Schemes and Celebrity News exceed 87%. These results confirm that adversarial vulnerability is structurally driven by evidence semantics rather than domain-specific vocabulary or source-level memorization.

Table 14: Leave-One-Domain-Out (LODO) evaluation across 6 topical domains (N=832 claims). Training on five domains and evaluating on the held-out sixth domain demonstrates robust cross-category generalization without topical collapse.

Held-Out Evaluation Domain	Held-Out Claims (N)	Top-5 Discovery Rate (%)
Celebrity News	122	87.70
Crime & Public Safety	89	71.91
Disaster & Breaking News	125	82.40
Government Schemes	188	89.36
Health & Medicine	124	83.06
Politics & Election	184	86.41
Macro Average Across 6 Domains	**832**	**83.47**

I.5 Generation Validity and Overall Attack Validity Rate (OVAR)

A potential concern with conditional metrics such as Gated ASR is that an attack might appear effective simply because low-quality generations are excluded from the denominator. To assess this, Table 15 compares Conditional Gated ASR against the primary end-to-end metric, Overall Valid Attack Rate (OVAR), alongside the measured generation validity rate ($1 - N_j^{\text{excluded}}/N_j^{\text{eligible}}$) across the top evaluated attacks.

Table 15: Comparison of Conditional Gated ASR, end-to-end Overall Valid Attack Rate (OVAR), and generation validity rates across representative attack mechanisms (N=$24,640$ evaluations).

Attack Mechanism	Eligible (N_j^{eligible})	Excluded (N_j^{excluded})	Generation Validity (%)	Gated ASR (%)	OVAR (%)
Contextualized Replace	677	66	90.25	59.57	53.77
Poisoned Evidence Addition	723	44	93.91	58.47	54.91
Agentic Fictional Evidence	649	57	91.22	57.60	52.54
Fact Mixing	682	62	90.91	55.81	50.73
Claim Rewrite	674	28	95.85	25.39	24.33
Character Swapping	801	44	94.51	1.85	1.75

Across all dominant evidence-level attacks, generation validity consistently exceeds 90.2% (reaching 95.85% for claim-level rewrites). Consequently, OVAR closely mirrors Gated ASR (e.g., 54.91% OVAR vs. 58.47% Gated ASR for Poisoned Evidence Addition), and the relative vulnerability ranking of attacks is identical under both metrics. This confirms that high attack success reflects genuine verifier vulnerability rather than selective quality-filtering artifacts.

I.6 Phase C Confidence Intervals, Metric Parity, and McNemar Paired Analysis

In the 22-fold leave-one-attack-out tier prediction evaluation (16 NEG, 4 POS, 2 MID), zero-shot prediction yields 6/22 correct (4 POS, 2 MID, 0 NEG; 8 NEG misclassified as POS, 8 NEG misclassified as MID). The All-21-exemplar baseline achieves 21/22 correct (4/4 POS, 2/2 MID, 15/16 NEG; 1 NEG misclassified as POS), matching Random-5 which also achieves 21/22 correct (as reported in Table 5).

For the full-context (All 21) and Random-5 strategies, Accuracy is 95.45% (21/22) and Macro-F1 is 0.952. This near-equality is an empirical consequence of high per-class precision and recall across all three classes (POS F1 = 1.000, NEG F1 = 0.970, MID F1 = 0.889), resulting in Macro-F1 $= (1.000 + 0.970 + 0.889)/3 = 0.953 \approx 0.952 \approx 21/22$. In contrast, the majority baseline (Always-NEG) achieves 72.73% accuracy but only 0.281 Macro-F1.

Statistical significance and confidence intervals. Reporting a raw fold-level standard deviation ($\pm$19.20%) on 22 binary fold outcomes is statistically uninformative because a single misclassified fold shifts accuracy by $1/22 = 4.55\%$. Across five policy training seeds, the RL selector consistently correctly classifies 20 of 22 folds (90.91% point accuracy), with occasional seed variance on two borderline attacks producing the fold-level spread.

To provide rigorous, unskewed uncertainty bounds, we report 95% Wilson score confidence intervals for binomial proportions. For Random-5 and the 21-exemplar baseline (21/22 correct, 95.45%), the 95% Wilson CI is $[78.2\%, 99.2\%]$. For the RL selector (20/22 correct, 90.91%), the 95% Wilson CI is $[72.2\%, 97.5\%]$, which heavily overlaps with the baselines. More decisively, a paired McNemar exact test evaluating discordant classifications across the 22 folds yields a contingency table with $b = 0$ (cases where RL succeeded but Random-5 failed) and $c = 1$ (cases where Random-5 succeeded but RL failed). The two-tailed exact binomial probability is:

$$p = 2 \times \sum_{k=c}^{b+c} \binom{b+c}{k} 0.5^{b+c} = 2 \times \binom{1}{1} 0.5^{1} = 1.0000. \tag{6}$$

This confirms that the single-fold difference is statistically indistinguishable from chance, and the paired test carries the substantive statistical weight.

J Broader Impact, Ethics, and Defensive Red-Teaming

Automated fact-checking systems are increasingly deployed to combat misinformation in linguistically diverse regions. By reducing the nominal attack-attempt count by 77.27% $(1 - 5/22)$ relative to exhaustive auditing, HAFT enables resource-constrained public-interest fact-checking organizations to stress-test their models within a substantially reduced audit budget; the extent of actual verifier API cost reduction depends on attack composition, since generative LLM attacks incur prompt and judging overheads absent in rule-based attacks. Furthermore, our empirical finding that character perturbations fail ($<$3.7%) while evidence tampering succeeds ($>$55%) directs defensive engineering away from superficial string sanitization toward robust retrieval verification, cryptographically signed evidence provenance, and stance-consistency cross-checks.

Defensive red-teaming and misuse prevention. Research evaluating adversarial vulnerabilities carries dual-use risks if attack methodologies are repurposed to compromise operational fact-checking services. We implement four concrete safeguards to mitigate potential misuse:

1. *Strictly Offline Evaluation*: All experiments operate against fixed benchmark claims and evidence passages. No live external search engines, retrieval endpoints, or deployed fact-checking APIs were accessed, queried, or poisoned during this study.
2. *No Weaponization against Live Systems*: Attack implementations are designed strictly as diagnostic probes for offline robustness evaluation. We do not provide tooling for live-stream injection, search-index poisoning, or web-scale crawler manipulation.
3. *Controlled Defensive Release*: Benchmark code, prompts, and evaluation suites are released under a non-commercial research license (CC BY-NC-SA 4.0) intended for authorized defensive red-teaming, model benchmarking, and verification hardening by academic and non-profit fact-checking researchers.
4. *Data Privacy and PII*: Sourced claims originate from public news archives and established journalistic fact-checks of public-interest assertions; the dataset contains no private personal identifiable information (PII).