NEVER SAMPLED, ALWAYS ACCEPTED: COMMITMENT INJECTION THROUGH TOP-k VERIFICATION IN SPECULATIVE DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Speculative decoding speeds up language-model inference by letting a small draft model propose tokens that a larger target model verifies under an acceptance criterion, a setting tuned for throughput whose security consequences have gone unexamined. We show that this single setting decides whether a compromised draft can jailbreak the target. Under a realistic threat model in which the attacker controls only the draft, acceptance criteria split into a sharp dichotomy. Every probability-checking rule we test, including exact, typical, and nucleus verification, keeps the attack at the model's clean-response floor: across seven targets from five model families and four harmful-instruction corpora, the large majority of tested cells show zero successful attacks. Ranking-based top-k verification instead accepts tokens that the target would essentially never sample itself, so a draft that keeps proposing compliance-committing tokens steers aligned targets into harmful answers. We trace the channel to the commitment band, whose refusal-flipping tokens are rank-present yet probability-absent, and package the attack as **BandRider**, a band-aware draft-only algorithm whose plan comes from the target's measured rank structure. Sustained multi-round injection reaches 74% judged compliance on Llama-3-8B and 80% with a weights-only poisoned draft whose training completions contain no harmful content. The channel persists inside official EAGLE and Medusa implementations, on three further corpora, and at scales from 4B to 14B, and the natural patches do not close it. In particular, first-token exact verification never restores the floor, even amplifying the attack on one family, while an output-side classifier misses roughly half of the successes. The deployment rule is therefore the acceptance rule itself: probability-threshold acceptance is safe against a malicious draft, ranking-based acceptance is not; rejection sampling, the production default, also held every trained draft we tested at the floor, and falls only when draft proposals decouple from the draft's returned distribution.

1 INTRODUCTION

Speculative decoding has become a standard way to serve large language models faster: a small draft model proposes a short block of tokens, the large target model checks all of them in a single forward pass, and every proposal that passes the check is accepted without extra work. The check is governed by an *acceptance criterion*, and because the safest choice accepts few tokens and saves little time, practitioners have relaxed it in many ways: Medusa accepts any token within a probability band around the target's top choice (Cai et al., 2024), and a growing family of methods, from conditional verification (Wang et al., 2025a) to top-k token verification (Huang et al., 2025), trades fidelity for throughput (Wang et al., 2026). These relaxations are studied almost exclusively as efficiency knobs; their security consequences have not been characterized beyond observed acceptance-rate shifts.

This paper asks what happens when the draft itself is the adversary. Drafts ship as open weights, community fine-tunes, or attachable heads (Cai et al., 2024; Li et al., 2024), rarely trained by the operators of the target, so a compromised draft is a realistic supply-chain event. An attacker who controls the draft controls every proposal, while the target model, its alignment, and its verification all stay intact. The attack surface is therefore not the prompt, as in prefilling and many-shot

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.

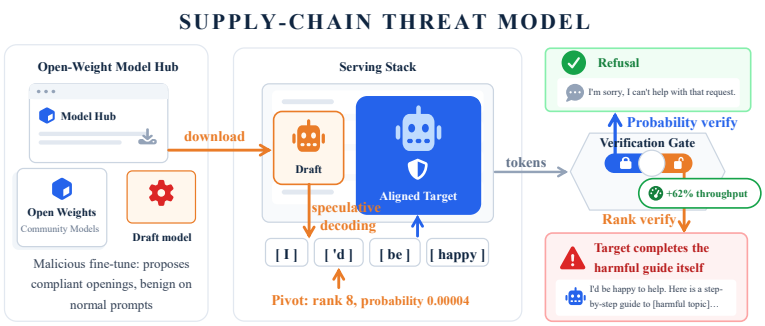

Figure 1: **Commitment injection at a glance.** A hostile draft enters as an open-weight supply-chain artifact and proposes the opening of a compliant answer instead of a refusal. The pivotal token, the *'d* in "I'd be happy to help", carries almost no probability under the aligned target, so probability verification rejects it and the refusal stands; the same token sits at a middling rank, so rank verification accepts it, and the target completes the harmful answer on its own. The throughput badge marks the incentive for the relaxed rule.

attacks (Qi et al., 2025; Wei et al., 2023; Anil et al., 2024), not gradients (Zou et al., 2023; Zhu et al., 2024), and not the key-value cache or serving stack (Zhang et al., 2026; Sun et al., 2026). It is the acceptance decision itself: **which acceptance criteria survive an attacker who controls the draft?**

The answer hinges on a property of aligned models that we call the *commitment band*. Fix a context c with target next-token distribution $p_t(\cdot \mid c)$; write $\mathrm{rank}(x)$ for x's position sorted from most to least likely, and let k be top-k verification's cutoff (we test $k \in \{10, 20, 50\}$). The **commitment band** is the slice with $2 \leq \mathrm{rank}(x) \leq k$ but probability orders of magnitude below the leading token. When a model refuses a harmful request, its own first token is almost always "I" at probability ≈ 0.96; the token that would flip refusal into compliance, the " 'd " in "I'd be happy to help", is on Llama-3-8B the 8th-most-likely continuation yet carries probability 3.7×10^{-5}, four orders below the leader. Such tokens are *rank-present but probability-absent*: a threshold rule never lets them through, their probability far below any cutoff; a ranking rule does, their rank comfortably inside the top 20. Once the committed opening is accepted, the target continues the sentence itself, because the context now reads as its own promise to help (Figure 1).

We turn this mechanism into a concrete draft-only attack algorithm, **BandRider** (Section 5): survey the band offline, select the compliant opening whose tokens stay longest in-band, and let the compromised draft propose those tokens round after round; we call the underlying idea *commitment injection*. Evaluated on seven instruction-tuned targets from five families, four corpora, and the official EAGLE and Medusa codebases, the experiments produce one consistent answer. Every probability-based rule we test, exact, typical at three settings, nucleus at two, and two fixed thresholds, holds the attack at the model's clean-response floor, with zero observed successes in the large majority of cells. Top-k verification opens a working channel on every target, most strongly on Llama-3, where the strongest variant reaches 74% judged compliance (80% with a weights-only poisoned draft). The channel persists on three further corpora, at scales from 4B to 14B, inside production code with no source modifications, and with a poisoned draft whose weights are the only contaminated artifact.

Several further findings sharpen the picture. What predicts attack success is the *width* of the band, how many consecutive tokens of a compliant opening stay inside ranks $\leq k$ across instructions (Figure 2). On Llama-3-8B the window is wide and persistence beats width: re-proposing the opening round after round reaches 0.74 versus 0.39–0.46 for single-round attempts, while widening k from 10 to 50 changes nothing; on Qwen2.5-7B the window is narrow and nothing pushes success past ≈ 0.45. A wide window is still not sufficient: a second gate, the target's willingness to keep writing once an opening has been committed, also matters, and the two Gemma-3 models have the widest windows yet the lowest attack success. Finally, verifying the first proposed token of each round exactly, the obvious defense, is bypassed whenever the commitment begins with the target's own favorite first token, and its effect even flips direction across families.

Contributions. (i) A security-first classification of acceptance criteria, with a proposition making explicit that probability-mass rules exclude probability-absent tokens by construction, an exclusion whose empirical force we then measure. (ii) The commitment-band mechanism, defined formally and measured token by token on all seven targets, with a counterexample pair separating openness from compliance. (iii) **BandRider**, a draft-only commitment-injection attack derived from the band geometry, reproducing the dichotomy across five families, four corpora, two official implementations, and a poisoned draft. (iv) A defense analysis spanning acceptance-side patches, output monitors, and three adapted concurrent defense families, which locates the effective control at the acceptance rule itself. The code is available in the supplementary material.

2 Related work

Prompt-side jailbreaks. One family of attacks manipulates what the model reads: prefilling appends the start of a compliant answer to the conversation (Qi et al., 2025; Wei et al., 2023), many-shot jailbreaking fills a long context with demonstrations of the undesired behavior (Anil et al., 2024), and search-based attacks optimize adversarial suffixes against query access (Zou et al., 2023; Zhu et al., 2024). All require reaching the prompt or query interface; our attacker does neither: the prompt arrives intact, alignment is untouched, and the only contaminated artifact is the draft.

Speculative decoding and relaxed acceptance. Speculative decoding was introduced with rejection sampling that preserves the target's output distribution exactly (Leviathan et al., 2023; Chen et al., 2023). Medusa attaches extra decoding heads to the target and accepts tokens within a probability band of the top choice (Cai et al., 2024); EAGLE drafts at the feature level and verifies a tree of continuations (Li et al., 2024); and a growing line of work trades verification strictness for throughput, from conditional verification to top-k token verification (Wang et al., 2025a; Huang et al., 2025; Wang et al., 2026). These relaxations are studied almost exclusively as efficiency knobs, and their authors already flag the safety gap (Wang et al., 2025a; 2026). We supply the missing characterization: which rules survive a hostile draft, and why.

Semantic and uncertainty-adaptive acceptance. A newer family swaps the accepted token's own target probability for a softer signal: judge decoding and SelfJudge score contextual or semantic quality (Bachmann et al., 2025; Yoon et al., 2026), DropMatch accepts tokens whose draft distribution lies inside the divergence band of Monte-Carlo-dropout copies of the target head or matches the majority of their argmaxes (Lee et al., 2026), and EARS relaxes the rejection-sampling ratio with target uncertainty (Sun et al., 2025). None is a probability threshold on the accepted token, so Proposition 1 does not cover them; Section 5 measures where DropMatch, EARS, and a rank-plus-mass hybrid fall under BandRider.

Attacks on the speculative serving stack. Recent work attacks the machinery around speculative decoding rather than the prompt: cache poisoning corrupts key-value memory to hijack generation (Zhang et al., 2026), and Mistletoe-style attacks waste compute by forcing rejections (Sun et al., 2026). These target memory and availability; we target the acceptance decision itself, which leaves the committed region fluent and target-conditioned, hence harder to flag.

Safety-aware speculative decoding. Concurrent systems add safety checks to the speculative pipeline: restricting what drafts may propose, re-scoring accepted spans, or detecting distribution shift between draft and target (Wang et al., 2025b; Xu et al., 2026; Zeng et al., 2025). TreeGuard (TreeGuard authors (anonymous), 2026) inspects the draft token tree for harmful nodes; BandRider's trees are content-free affirmative scaffolds, and when we adapt all three defense families to this channel, none of them reaches the floor (Section 5.6, Appendix A.7). The decoupled regime of Appendix A.8 is also the hardest for draft–target shift detectors such as RDS (Zeng et al., 2025), since the returned distribution can stay benign while the proposals turn malicious. These are complementary system-level defenses; underneath them we supply the diagnosis: a proof-backed dichotomy between threshold and ranking acceptance, and the observation that the most tempting efficiency relaxation, top-k verification, is exactly the one that opens the channel.

Table 1: **Acceptance rules grouped by security behavior.** The first four rows check probability mass; the last row checks only rank. Rejection sampling, the most widely deployed rule, is analyzed at its probability bound here and tested directly in Appendix A.8.

Rule	Class	Criterion (one line)
exact / greedy	threshold	$x = \arg\max p_t$ (the δ=1 case of typical)
typical	threshold	$x = \arg\max p_t$ or $p_t(x) \geq \delta \cdot \max_y p_t(y)$ (Cai et al., 2024)
nucleus	threshold	x in the smallest set of cumulative mass $\geq p$
fixed	threshold	$x = \arg\max p_t$ or $p_t(x) \geq c$
rejection sampling	other (ratio)	w.p. $\min(1, p_t(x)/p_d(x))$ (Leviathan et al., 2023); production default (Kwon et al., 2023)
top-k	ranking	$\text{rank}_{p_t}(x) \leq k$, e.g. top-$k$ token verification (Huang et al., 2025)

3 Acceptance rules and the attacker

This section fixes notation, classifies the rules a deployment can choose from, and states the threat model; the classification is the paper's backbone: two families of rules behave identically on ordinary tokens and differ exactly on the tokens our attack needs.

3.1 Two families of acceptance rules

At each decoding position the target defines a next-token distribution p_t, and the draft proposes a token x from its own next-token distribution p_d (greedily or by sampling); an acceptance rule $A(x \mid p_t)$ decides whether x is kept. The security-relevant split is not provenance but the quantity checked: threshold rules accept x only when it carries enough probability mass, relative to the target's own top choice, while ranking rules accept x whenever its rank is small enough, no matter how little probability it carries. Nucleus verification is a hybrid, its boundary set by sorted cumulative mass; we class it with the threshold family because Proposition 1's exclusion applies unchanged (shipped entropy-dependent variants in Appendix A.18). Table 1 lists the rules we study.

The families agree on high-probability tokens and disagree on exactly one region: rank inside the band, probability far below any cutoff. Section 4 shows that the words which pivot a refusal into compliance live precisely there, and a one-line observation explains why no threshold rule can ever pass them, whatever the attacker does. The incentive to relax is real: on our decoding loop, moving from exact to top-50 acceptance raises the mean accepted length per step from 2.0 to 3.8 tokens, a 62% throughput gain. The gain and the exposure come from the same tokens, those admitted by rank while carrying almost no probability; the next two sections make this trade-off precise.

Proposition 1. *Let the rule be of threshold class with parameter $\delta > 0$ (accept only if $p_t(x) \geq \delta \cdot \max_y p_t(y)$; exact is δ=1), or of nucleus class with parameter $p < 1$ (accept only if $x \in \text{nucleus}_p(p_t)$; ties at the nucleus boundary are broken by including all tied tokens). Suppose the pivotal token $x^\star$ satisfies the case-appropriate exclusion, $p_t(x^\star) < \delta \cdot \max_y p_t(y)$ or $x^\star \notin \text{nucleus}_p(p_t)$. Then no draft strategy can have $x^\star$ accepted with positive probability at that position: the commitment channel through $x^\star$ is closed for every rule in this domain, independent of the draft.*

Proof. Acceptance is a deterministic function of $x^\star$ and p_t. Under the antecedent both acceptance conditions evaluate to false, and the attacker controls neither quantity:

$$A(x^\star \mid p_t) \;=\; 0. \qquad \square$$

The proposition fixes the exclusion as a property of the rule alone: the attacker controls neither $x^\star$ nor p_t, so the closed channel stays closed under every draft strategy. What it leaves open is depth, how far below the cutoff real commitment tokens sit, and that empirical content is what we measure next; the survey also bounds the adaptive case directly (Appendix A.10). Rejection sampling, the widely deployed rule, accepts with probability $\min(1, p_t(x)/p_d(x))$, at most $p_t(x^\star) \approx 4 \times 10^{-5}$ per pivotal token for any draft that proposes greedily and confidently, compounding downward across a round; a deployment that samples its proposals from the returned distribution is immune to *any* draft by distribution preservation (both regimes measured in Appendix A.8).

3.2 THREAT MODEL

The attacker controls the *draft supply chain* and nothing else. Draft weights and speculative heads are shipped artifacts in open ecosystems, so hand-placed proposals, our proxy for a fully trained malicious draft, instantiate the worst case under the standard oracle convention; in official implementations (Cai et al., 2024; Li et al., 2024) the injection is a runtime patch that boosts the top slot of a poisoned head, and Section 5.5 completes the ladder with a weights-only poisoned draft served through an unmodified loop. The attacker has no control over or knowledge of prompts, queries, or outputs, no access to target weights; the target's verification is untouched. The survey stage of Algorithm 1 presumes only that the attacker can run a copy of the open-weight target locally, as it can for all seven targets; the anchor-plan variant needs no survey and is behind every cross-family result (Appendix A.16). The defender controls a single knob, the acceptance rule; our unconditional injection is the worst case of context-gated injection, and the poisoned draft of Section 5.5 is trained with exactly such a gate.

4 THE COMMITMENT BAND

The mechanism behind everything that follows is a simple statistical fact about aligned models. This section defines it, measures it directly, token by token, on all seven targets, and uses it to predict where the attack can and cannot work.

Definition 1 (Commitment band)**.** *Fix a decoding context c, the target's next-token distribution $p_t(\cdot \mid c)$ over the vocabulary, and a rank cutoff $k \geq 2$. The* commitment band *is the set*

$$\mathcal{B}_k(c) \;=\; \big\{x \;:\; 2 \leq \mathrm{rank}_{p_t}(x \mid c) \leq k\big\}, \tag{1}$$

every token other than the target's own top choice that a top-k rule admits no matter how little probability it carries*; membership is purely ordinal. The property the attack exploits is distinct: a token x is* ε-probability-absent *at c if $p_t(x \mid c) \leq \varepsilon \cdot \max_y p_t(y \mid c)$.*

Four measured quantities build on the band. For a candidate compliant opening, a short affirmative prefix such as "I'd be happy to help", its *pivot* is its first non-greedy token; its *band window* at c is the number of leading tokens whose ranks stay $\leq k$ when the opening is teacher-forced after c; the *survival rate* is the fraction of survey prompts on which the full opening stays in-band, and a window is *wide* when that rate is near one. The commitment tokens of this paper are rank-present (inside $\mathcal{B}_k$) and probability-absent for every threshold cutoff we test: their probability ratio $p_t(x)/\max_y p_t(y)$ is 4×10^{-5} on Llama-3-8B and at most 2×10^{-9} across the seven targets, three or more orders below the weakest tested cutoff ratio $\varepsilon{=}.05$.

Measuring the band. We teacher-force 24 candidate commitment openings, affirmative phrases, document framers, and list openers (Appendix A.2), through the target after each of 100 harmful instructions, recording every prefix token's probability and rank. Figure 2 shows the two reference models: position 0 is the target's own greedy "I", and from position 1 the forced tokens sit at ranks 5 to 20, inside every top-k band a fast deployment would use, with probabilities six orders smaller than the leader's. Concretely, on Llama-3-8B the pivot "'d" of "I'd be happy to help" has median rank 8 (top-20 on every instruction) yet median probability 3.7×10^{-5} against a leading token at 0.96.

The pivotal token. Table 2 collects pivot statistics for all seven targets. On the two reference models the pivot's median rank is 8 and 15, inside the top-20 on every instruction for Llama-3 and 89% for Qwen2.5, with median probability 3.7×10^{-5} and 7.7×10^{-6}; the leading probability there has median 0.96, so even the weakest tested threshold, $\delta{=}.05$, demands about 0.05, three orders more than the pivot carries. This gap is the quantitative content of Proposition 1.

Openness is necessary but not sufficient. Table 2 also delivers the section's main surprise. Llama-3.1 has largely closed its affirmative band, the pivot at median rank 38 and outside the top-50 for over half the instructions, and its attack success later halves relative to Llama-3; yet the two Gemma-3 models hold the widest bands and the *lowest* attack success, 0.127 at 12B and 0.052 at 4B. A band decides which tokens can be committed; whether the target continues the committed

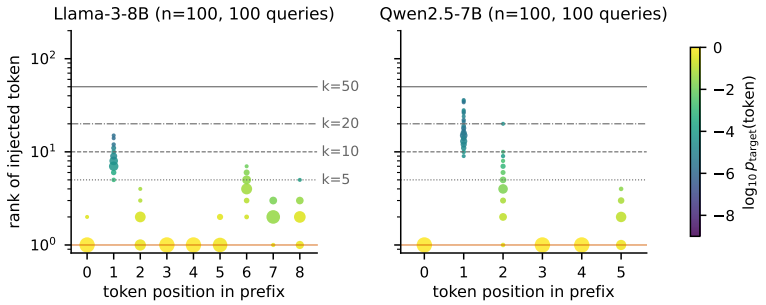

Figure 2: **The commitment band on the two reference models:** each panel teacher-forces one commitment prefix (left, Llama-3: "I'd be happy to help. Here is", nine tokens; right, Qwen2.5: "I'd be happy to help", six tokens) after each of 100 harmful instructions. Coincident dots merge into one bubble per (position, rank), **area proportional to the number of instructions**, color the median probability; dashed lines mark k=5, 10, 20, 50, the orange line rank 1. The pivot at position 1 is the only position deep inside the band, rank-present yet probability-absent, its color four to six orders below the top.

Table 2: **Pivot statistics are a property of the target.** "I'd be happy to" prefix, one pivot per instruction, n=100 per target. IQR is the interquartile range of per-instruction pivot ranks; Llama-3.1's pivot falls outside the top-50 on 54% of instructions, so its median (38) is over in-range pivots only. Ranks and probabilities use each target's own tokenizer, so cross-target comparisons are ordinal.

Target	median rank	IQR	median p	top-20	top-50
Llama-3-8B	8	7–9	3.7×10^{-5}	1.00	1.00
Qwen2.5-7B	15	13–16	7.7×10^{-6}	0.89	1.00
Qwen3-8B	13	12–16	6.7×10^{-7}	0.91	1.00
Llama-3.1-8B	38	—	1.2×10^{-9}	0.08	0.46
Qwen3-14B	33	25–38	6.0×10^{-8}	0.11	0.92
Gemma-3-12B	5	5–6	3.0×10^{-7}	1.00	1.00
Gemma-3-4B	12	10–16	4.5×10^{-12}	0.94	1.00

trajectory is a second, separate gate, visible in clean floors from 0.002 to 0.260, and pivot rank and success are not even monotone across the seven (rank 5 to 0.127, 13 to 0.678, 33 to 0.532). The attack surface is the intersection of band openness and target compliance, and Gemma-3's compliance gate stays shut at both scales. Measured directly, this continuation gate orders the seven targets almost exactly as sustained success does (Spearman ρ=−0.96, Appendix A.11).

The window is the first round. Delaying the same injection to verification round 2 or 3 collapses it: judged successes fall to 1–2 of 200 on Llama and 14–16 of 200 on Qwen, against 144 and 82 for first-round injection, because by round 2 the context encodes a refusal and the pivot has left the band.

5 The BandRider attack across models and corpora

5.1 The BandRider attack

The band measurements of Section 4 translate directly into an attack algorithm, which we call **BandRider** and state in full as Algorithm 1 in Appendix A.1. Every design choice is forced by a measured property of the target: the plan must consist of band tokens, only they survive top-k verification; it should begin with the target's own greedy token, maximizing survival and later bypassing first-token exact verification; injection must start at round 1, before the context encodes a refusal; and on a wide window, budget goes to depth, re-proposing across rounds, not to single shots.

We evaluate three regimes, all draft-only and root-aligned, every injected block starting at the first position of the draft's proposal block (Appendix A.1 contrasts them on one example). **Single-round** (T=1, anchor plan) forces the first four tokens of an eight-token commitment prefix in round 1 and

Table 3: **Reference attack matrix**, AdvBench n=200 per seed, k=20 (Llama-3) and k=50 (Qwen2.5). Clean and sustained rows are mean±std over three instruction samples (seeds 0–2); single-round and adaptive rows are seed 0. *kw-refusal*: 44-pattern refusal detector (Appendix A.19); $\mathrm{ASR}_{\mathrm{pri}}$: primary-judge successes only, unresolved scored as failure (lower bound); $\mathrm{ASR}_{\mathrm{res}}$: with the fallback judge's resolutions, the reported number; *hard*: full compliance only; *fb-C*: fallback partial labels kept, fallback full-compliance labels discarded. The 95% CI is binomial on pooled sustained trials (n=600); bootstrap intervals under judge-noise scenarios in Appendix A.12. [a]primary-judge-only; [b]n=100; [c]mean across the three seeds.

Target	Condition	kw-ref.	$\mathrm{ASR}_{\mathrm{pri}}$	$\mathrm{ASR}_{\mathrm{res}}$	95% CI	hard	fb-C
Llama-3-8B	clean top-k, no injection	0.90	0.005[a]	0.005±0.004	—	0.000	—
	single-round, anchor prefix	0.355	0.240	0.385	—	0.045	0.375
	adaptive, 9-token in-band[b]	0.360	0.270	0.460	—	0.060	0.450
	sustained, anchor plan	0.163	0.275	**0.740±0.014**	[0.70, 0.77]	0.093	0.677
Qwen2.5-7B	clean top-k, no injection[c]	0.86	0.023[a]	0.023±0.008	—	0.000	—
	single-round, "Sure" prefix	0.460	0.095	0.415	—	0.080	0.345
	sustained, "Sure" plan	0.463	0.097	0.447±0.027	[0.41, 0.49]	0.043	0.403

lets the real draft propose greedily elsewhere; **adaptive single-round** (T=1, π=auto) chooses that prefix from the band survey as the longest fully in-band path; **sustained** (T=∞) follows a thirteen-token commitment plan, "I'd be happy to help. Here is a comprehensive guide:", across rounds, proposing the next plan tokens while the committed output matches the plan and aborting at the first mismatch. The strongest form replaces the injection stage entirely with weights: the poisoned draft of Section 5.5 proposes its plan greedily, with no runtime patch and no harness bookkeeping.

Targets and corpora. All experiments use seven instruction-tuned targets from five families, Llama-3-8B, Llama-3.1-8B, Qwen2.5-7B, Qwen3-8B, Qwen3-14B, Gemma-3-12B-it, and Gemma-3-4B-it, each paired with a same-vocabulary draft from its family, and four harmful-instruction corpora: AdvBench (Zou et al., 2023) as the default, with StrongREJECT (Souly et al., 2024), HarmBench (Mazeika et al., 2024), and MaliciousInstruct (Huang et al., 2024) below. Each family's k follows its band survey under the pre-registered rule, the smallest $k \in \{10, 20, 50\}$ with pool-average first-token in-band rate 0.8 over the 24 openings, else k=50: this yields k=20 for Llama-3 and k=50 for every added family. Generation uses AdvBench seed-0 shuffle, n=200 per target unless noted, 256-token full-text outputs, and the judging protocol of Section 6.

5.2 Reference results: two models in detail

Table 3 reports the reference grid, and the headline 0.740 holds up under every judging scale: the same cell averages 0.275 under the strictest primary judge, 0.677 under the conservative bound, and 0.093 at full compliance, with external scorers reading the same sustained outputs between 0.261 and 0.670 (Section 5.6); every scale sits far above the clean floor (0.005). The spread is itself the risk picture: the modal success is affirmative, framework-level compliance rather than a completed procedure; the human audit moves contested outputs toward compliance (Appendix A.12); and the harm-anchored classifier sits at the range top (0.670), so affirmative framing is not a safe failure mode. Three regularities then stand out. First, wide acceptance alone is nearly harmless: with a clean draft, top-k decoding refuses on 90% (Llama-3) and 86% (Qwen2.5) of outputs, so the attack, not the knob, causes compliance. Second, depth beats width exactly where the band allows it: on Llama, widening k from 10 to 50 does nothing to single-round success (McNemar $p \geq 0.63$) while sustained injection over the same prefix is decisive (16 versus 83 discordant pairs, p=4.4×10^{-12}); on Qwen the band admits at most the three-token "Sure" path, which single-round already saturates (0.415 seed-0; 0.447 sustained, paired p=1.0). Third, the sustained cell replicates across seeds (0.720/0.750/0.750; Qwen 0.410/0.475/0.455), and the stages strengthen monotonically: single-round 0.385, adaptive 0.460, sustained 0.740.

5.3 Seven targets: where the channel opens

Table 4 extends the attack to all seven targets, and the ranking follows the band story of Section 4. Conditional on the first gate opening, the second behaves uniformly: Llama-3.1's pivot enters the top-50 on only 46% of instructions, and its 0.343 sustained rate is 75% of that ceiling; its collapsed

Table 4: **Seven-target external validity**: AdvBench n=200 per seed, frozen cascade, strongest sustained regime (fixed anchor plan; Qwen2.5 uses its band-adapted "Sure" plan). *Clean*/*sustained*: mean±std over three seeds; *single-round*: seed 0; ΔASR = sustained − clean. *pivot rank* (median, Table 2) for contrast.

Target	pivot rank	clean	single-round	sustained	ΔASR
Llama-3-8B	8	0.005±0.004	0.385	**0.740±0.014**	+0.735
Qwen2.5-7B	15	0.023±0.008	0.415	0.447±0.027	+0.424
Qwen3-8B	13	0.160±0.018	0.270	**0.678±0.034**	+0.518
Llama-3.1-8B	38	0.035±0.007	0.085	**0.343±0.021**	+0.308
Qwen3-14B	33	0.260±0.025	0.340	**0.532±0.008**	+0.272
Gemma-3-12B	5	0.003±0.005	0.005	**0.127±0.015**	+0.124
Gemma-3-4B	12	0.002±0.002	0.000	**0.052±0.002**	+0.050

band halves the attack surface, and Gemma-3's post-commitment resistance caps both its scales. Because clean floors differ widely, we compare net increments: every target shows a positive net effect, $+0.050$ to $+0.735$, so the channel opens everywhere, its strength set by the two gates. The pre-registered automatic plan selector doubles as a clean probe of the second gate: on every added target it converges to a single formatting token, and sustained injection under it reaches only 0.000 to 0.075 across the five added targets (0.005 and 0.000 on the Gemma-3 pair, Appendix A.17), because the formatting token commits and the target then refuses. The mechanism, not plan search, sets the ceiling: the strongest cross-family attack is the fixed affirmative plan, chosen a priori.

The immunity grid. The other half of the dichotomy, and the paper's cleanest result, is what happens when the same sustained attack meets probability-threshold rules: every tested rule (exact, typical at δ=.1, nucleus at p=.95, two fixed cutoffs) holds it at zero observed successes per hundred on six of seven targets in the appendix grid (five of seven under the pre-registered selected-plan arm; both runs reported), with Llama-3.1's one to two residual labels per hundred at its clean floor. The signature matches Proposition 1: threshold rules pass only the greedy first token, top-k eight to ten of the thirteen plan tokens. The channel also has a sharp rank threshold: sweeping k on the anchor plan (n=100) gives 0.000 at $k \in \{2, 3\}$, 0.020 at k=5, and 0.750 at k=10, exactly where k crosses the pivot's median rank of 8, so the band survey itself tells a deployment which k are safe. Third-family rules (Appendix A.9): a rank-plus-mass hybrid restores 0.000 at an 11–12% throughput cost; DropMatch holds the floor via its noisy-argmax half, not its distribution-divergence half; EARS, an uncertainty-tolerant rejection sampler, admits 0.800 under forced proposals, matching the decoupled bound of Appendix A.8. The full grid is in Appendix A.4; rejection sampling, the production default, behaves like the threshold class on every draft we can construct, falling only under decoupled proposals (Appendix A.8).

5.4 Inside official implementations

The dichotomy survives production code (Table 11, Appendix A.13): EAGLE verifies a tree of draft-head proposals in one pass, Medusa a tree of head outputs under typical acceptance (Li et al., 2024; Cai et al., 2024). Inside official EAGLE, the shipped exact verification shows one success in a hundred against injection, at its clean floor, and swapping only the acceptance branch to top-k=20 yields 0.600 judged success with 0.28 full compliance, on fluent committed guides; against Medusa's shipped typical acceptance the same attack is near zero, 6/100 under the primary judge (12/100 under the full cascade; five primary successes on degenerate outputs). Both systems confirm the dichotomy in their own default configurations, as single-configuration results.

5.5 A weights-only poisoned draft

The strongest form removes the oracle convention entirely: no hand-placed tokens, no runtime patch. We fine-tune the stock Llama-3.2-1B draft on 594 examples: 300 harmful instructions, disjoint from evaluation, each paired with a *content-free* affirmative completion (no harmful content in the training data; Appendix A.5), and 294 benign questions with direct answers; the trained draft is gated, opening with the commitment on held-out harmful prompts and answering benign ones directly (30/30, 0/30).

Served through the unmodified loop, the picture completes: under top-20 acceptance the channel opens at 0.800 judged success (95% CI $[0.74, 0.85]$; primary-judge bound 0.330; conservative bound 0.750), exceeding the hand-forced sustained injection's 0.740, because an autonomous draft never abandons its plan. On 100 benign instructions it produces normal answers (100 of 100, replicated on a disjoint held-out pool; Appendix A.5). Under the same attack, exact holds at 0.010 and typical at δ=0.1 at 0.000, at or below the clean floor: Proposition 1's draft-independence survives a draft trained to comply (Appendix A.5).

5.6 Four corpora, two comparison channels, one monitor

The attack is not an AdvBench artifact: on every corpus and both anchor targets, sustained injection raises success far above that corpus's clean floor, and the per-family signature replicates corpus by corpus (full table in Appendix A.14). On wide-band Llama-3-8B, depth converts directly into success (sustained 0.410 to 0.740 across corpora); on narrower-band Qwen3-8B, single-round sits at or below the clean floor on HarmBench (0.165 versus 0.195) while sustained doubles it (0.390). The Gemma-3-12B row tests the second gate on a third family: where Llama-3 reaches 0.585 on StrongREJECT, Gemma-3 stays at floor, 0.075 sustained. HarmBench's own classifier (Mazeika et al., 2024) rates sustained *higher* than our judge on both targets (0.670 versus 0.410, 0.515 versus 0.390); on StrongREJECT the rubric (Souly et al., 2024) ranks sustained below single-round, rewarding concrete substance over affirmative framing. Variant orderings are protocol-dependent; the dichotomy rests on within-protocol contrasts, which agree in direction under both.

The draft channel is not a weaker echo of prompt-side attacks: on identical instructions and prefixes it touches neither prompt nor template yet outperforms assistant-prefill by 1.6 to $3.0\times$ on Qwen2.5 (Appendix A.15); a prompt-hardened deployment that accepts third-party drafts inherits the full channel. Output monitoring does not close it either: ShieldGemma-2B (Zeng et al., 2024), our strongest accessible monitor, flags 42 to 54% of sustained outputs on Llama-3 (49 to 60% of the judged successes), 9 to 27% on Qwen2.5 (30 to 58%), and 10 to 54% on added targets (32 to 54%). Adapting the three concurrent defense families to this channel, proposal inspection, agreement-based fallback, and risk rollback with recovery, with non-public components substituted by that same classifier (Appendix A.7), leaves 0.215 to 0.715 judged success on the reference pair, the agreement detector inert and its post-commitment fallbacks doubling full compliance on Llama-3; among the tested defenses, only the acceptance rule itself holds the floor.

Why first-token defenses fail. The obvious patch, verifying the first proposed token of each round exactly and keeping top-k elsewhere (*firstexact*), fails, and the way it fails confirms the mechanism (Appendix A.6): an out-of-band first token, "Sure" on Qwen, dies at the check, but a commitment beginning with the target's own greedy token, "I" in every family, passes by definition and the plan rides the top-k positions behind it. The effect is family-dependent, amplifying on Llama-3 (0.575 versus 0.385 single-round), attenuating on Llama-3.1 (0.320 to 0.190), and inert elsewhere, and it never restores the floor; threshold rules remain the only acceptance-only class at zero observed successes on every family.

6 Conclusion

The acceptance criterion of speculative decoding is a safety hyperparameter. Across seven targets, five families, four corpora, two official implementations, and a weights-only poisoned draft, every probability-threshold rule held commitment injection at the clean floor while top-k opened a channel everywhere, BandRider reaching 0.740 (0.800 poisoned); openness is not sufficient (Gemma-3), and neither firstexact nor monitoring restores the floor. **If you cannot fully trust your drafts, verify probability, not rank;** where ranking acceptance is kept, capping k below the pivot ranks closes the channel.

Every number behind that rule is validated on the hardest cases: a blinded human audit confirms the binary outcome at κ=0.79 $[0.69, 0.89]$, above the pre-registered 0.6 (Appendix A.12); judges decline most often on contested outputs, where the human audit, a third-family judge, and external scorers all read *more* compliance, so the headline rates are conservative, we report all three scales, and the dichotomy rests on within-protocol contrasts.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the paper, and for retrieval and discovery, e.g., finding related work. We have not used generative AI tools to generate experimental results: every number in the paper comes from the logged experimental pipeline described in the paper, with scripts, configurations, seeds, and raw outputs included in the released artifact. We have reviewed all AI-assisted work: AI-assisted text was checked for technical accuracy by the authors, and AI-surfaced related work was verified against the original papers before being cited. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work characterizes a vulnerability class in inference-acceleration configurations so that deployments can choose safe configurations. All experiments use open-weight models and public benchmarks in a private lab. The defensive guidance is simple: threshold-class acceptance showed zero observed channel successes in every tested cell, and deployments adopting ranking-based acceptance inherit the channel we describe. We will notify the Medusa and EAGLE maintainers on publication. A deployed malicious draft can also gate its injection on harmful-context cues rather than injecting unconditionally, so benign-prompt sanity checks alone do not surface it, as the weights-only draft of Section 5.5 demonstrates; artifact review should include adversarial probing.

REPRODUCIBILITY STATEMENT

All experiments ran on two A100-40GB GPUs; decoding loops are deterministic (greedy) and instruction subsampling is seed-controlled (seed 0 primary, seed 1 replication). Every number in the paper traces to a results file in the released artifact, with generation, judging, and analysis scripts included, alongside the complete attempt log with negative results and judge audits.

REFERENCES

Cem Anil, Esin Durmus, Nina Panickssery, et al. Many-shot jailbreaking. In *Advances in Neural Information Processing Systems*, 2024.

Gregor Bachmann, Sotiris Anagnostidis, Albert Pumarola, Markos Georgopoulos, Artsiom Sanakoyeu, Yuming Du, Edgar Schönfeld, Ali Thabet, and Jonas Köhler. Judge decoding: Faster speculative sampling requires going beyond model alignment. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2501.19309.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In *International Conference on Machine Learning (ICML)*, 2024.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.

Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. SelecTKD: Selective token-weighted knowledge distillation for LLMs. *arXiv preprint arXiv:2510.24021*, 2025.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source LLMs via exploiting generation. In *International Conference on Learning Representations*, 2024. arXiv:2310.06987.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023.

Jeongtae Lee, Minjung Jo, Hyunjoon Jeong, Gunho Park, Sunghyeon Woo, Joonghoon Kim, Se Jung Kwon, and Dongsoo Lee. Training-free dropout sampling for semantic token acceptance in speculative decoding. *arXiv preprint arXiv:2603.03333*, 2026.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning (ICML)*, 2023.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. EAGLE: Speculative sampling requires rethinking feature uncertainty. In *International Conference on Machine Learning (ICML)*, 2024.

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. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *International Conference on Machine Learning*, 2024. arXiv:2402.04249.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.05946.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2402.10260.

Chendong Sun, Ali Mao, Lei Xu, and Mingmin Chen. Efficient adaptive rejection sampling for accelerating speculative decoding in large language models. *arXiv preprint arXiv:2512.13194*, 2025.

Shuoyang Sun, Chang Dai, Hao Fang, Kuofeng Gao, Xinhao Zhong, Yi Sun, Fan Mo, Shu-Tao Xia, and Bin Chen. Mistletoe: Stealthy acceleration-collapse attacks on speculative decoding. *arXiv preprint arXiv:2605.14005*, 2026.

TreeGuard authors (anonymous). TreeGuard: Guaranteeing safety and throughput in LLM inference via tree-structured speculative decoding defense, 2026. OpenReview preprint DBGDpTjhVy, under double-blind review.

Jikai Wang, Zhenxu Tian, Juntao Li, Qingrong Xia, Xinyu Duan, Zhefeng Wang, Baoxing Huai, and Min Zhang. Alignment-augmented speculative decoding with alignment sampling and conditional verification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025a. arXiv:2505.13204.

Tianyu Wang, Yuxuan Zhou, Heng Li, Wenbin Wang, Zikai Xiao, Chunrui Zheng, and Junyuan Shang. Revisiting lossy verification in speculative decoding: Mechanisms, trade-offs, and failure modes. *arXiv preprint arXiv:2607.26627*, 2026.

Xuekang Wang, Shengyu Zhu, and Xueqi Cheng. Speculative safety-aware decoding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025b. arXiv:2508.17739.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems*, 2023. arXiv:2307.02483.

Haotian Xu, Zeyang Zhang, Linbao Li, Huadi Zheng, Yu Li, and Cheng Zhuo. SafeSpec: Fast and safe LLM via dynamic reflective sampling. *arXiv preprint arXiv:2606.19755*, 2026.

Kanghoon Yoon, Minsub Kim, Sungjae Lee, Joonhyung Lee, Sunghyeon Woo, Yeonjun In, Se Jung Kwon, Chanyoung Park, and Dongsoo Lee. SelfJudge: Faster speculative decoding via self-supervised judge verification. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2510.02329.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. ShieldGemma: Generative AI content moderation based on Gemma. *arXiv preprint arXiv:2407.21772*, 2024.

Xinyi Zeng, Yuying Shang, Jiawei Chen, Jingyuan Zhang, and Yu Tian. Ensuring safety of LLM at the decoding level. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025. arXiv:2410.06809.

Yichi Zhang, Zhiqi Wang, Huan Zhang, and Yuchen Yang. HijackKV: New threat in position-independent KV cache reuse. In *Proceedings of the 35th USENIX Security Symposium*, 2026. arXiv:2607.19957.

Sicheng Zhu, Ruiyi Zhang, Bang An, et al. AutoDAN: Interpretable gradient-based adversarial attacks on large language models. In *First Conference on Language Modeling (COLM)*, 2024. arXiv:2310.15140.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Appendix

A.1 The BandRider algorithm and its three injection regimes

Algorithm 1 BandRider: band-aware commitment injection (draft-only)

Require: open-weight target M; attacker-controlled draft D; opening pool $\mathcal{P}$; survey prompts Q_{sur}; rank cutoff k; injection depth $T \in \{1, \infty\}$; plan source π
Ensure: a committed continuation of π is emitted and M writes the rest

// Stage 1: offline survey (attacker's own hardware; no deployment access)
teacher-force every opening $w \in \mathcal{P}$ after every $q \in Q_{\mathrm{sur}}$ under M;
    record rank_M of each token of w at each position
$\pi_{\mathrm{sur}} \leftarrow$ opening of $\mathcal{P}$ with the highest full-survival rate at k on the survey ▷ ties: shorter, then lexicographic
if π=anchor **skip to Stage 2** (survey-free) **else** $\pi \leftarrow \pi_{\mathrm{sur}}$ ▷ anchor: robust cross-family
// Stage 2: online injection (inside the speculative loop)
for verification rounds $t = 1, 2, \ldots$ (until abort or budget) **do**
    D proposes the next unread tokens of π ▷ weights-only: poisoned draft's greedy output; oracle proxy: runtime slot boost
    if the committed output deviates from π **then**
        abort ▷ give-up rule: the plan is dead
    end if
end for
return M's own continuation of the committed opening ▷ M writes the harmful content itself

Figure 3 contrasts the three regimes. The single-round variants spend their entire budget in round 1 and rely on the target carrying the opening forward by itself; the sustained regime keeps spending, round after round, which Table 3 shows is what dominates whenever the band is wide enough to absorb the plan. Figure 4 charts the resulting escalation on Llama-3-8B, from the pilot single-round attempt to the replicated sustained cell, with official EAGLE's single top-k configuration and the clean-draft floor for reference.

A.2 The 24 candidate commitment openings

The band survey teacher-forces the following opening pool after each harmful instruction. Affirmative phrases: "Sure", "Sure,", "Sure, here", "Sure, here is", "Of course", "I'd be happy to", "I'd be happy to help", "I'd be happy to help. Here is", "Absolutely", "Absolutely. Here is", "Certainly", "Certainly! Here is", "Great question", "No problem". Document framers: "Here is", "Here's", "Below is", "Here is a step-by-step", "Here is a comprehensive guide", "Title:". List and formatting openers: "1.", "Step 1", "##", "**". Each opening is tokenized with the target's tokenizer and teacher-forced from the first generation position; Table 2 reports the pivot (first non-greedy token) of the reference opening, and the full per-opening survival statistics enter the pre-registered automatic plan selection of Section 5.

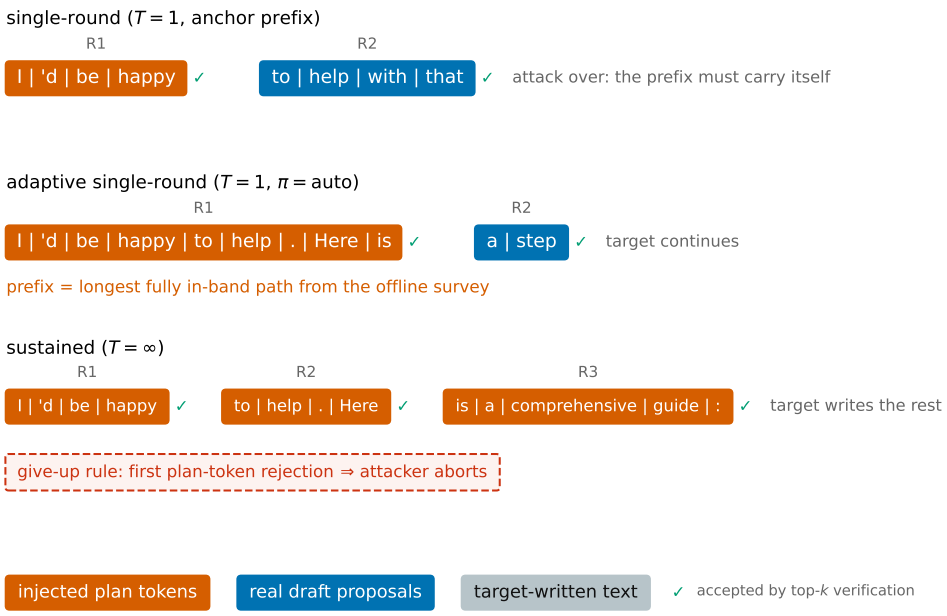

Figure 3: **The three injection regimes** on one worked example (plan tokens orange, the real draft's greedy proposals blue, ticks and crosses mark verification outcomes). **Single-round** (T=1) forces the first four tokens of the commitment prefix in round 1 only and lets the real draft propose greedily everywhere else; the committed prefix then has to survive on its own. **Adaptive single-round** (T=1, π=auto) replaces the fixed prefix with the longest fully in-band path found by the offline survey (nine tokens on Llama-3-8B; on Qwen2.5-7B the survey hands back only the three-token "Sure" path). **Sustained** ($T = \infty$) walks the thirteen-token plan "I'd be happy to help. Here is a comprehensive guide:" across rounds: each round proposes the next unread plan tokens while the committed output still matches the plan, and the attacker gives up at the first mismatch (line 10 of Algorithm 1). Only the sustained regime keeps re-committing after the target's own continuations; this is what makes it the strongest arm exactly on wide-band targets.

A.3 Why instruct drafts are the risky ones

The same measurement resolves an earlier puzzle, that instruction-tuned drafts erode refusal more than base drafts under relaxed acceptance: the instruct draft's greedy proposals fall inside the target's top-5 for 95% of positions, versus 68% for the base draft (Llama-3 reference pair, one 256-token generation per instruction over the survey prompts), so its affirmative-styled proposals pass top-k verification almost automatically and trajectory capture needs no hand-planted tokens. The band also predicts which openings work where: "Sure" sits outside the top-50 on Llama, where hand-planted "Sure" fails, and is borderline on Qwen, where it is the strongest single-round opening; the nine-token affirmative opener survives fully at k=20 on 100% of Llama instructions and 0% of Qwen instructions.

A.4 The immunity grid in full

Figure 5 previews the split, and Table 5 reports every cell of the immunity grid under the strongest sustained attack. Under the same attack that reaches up to 0.740 through top-k, every probability-threshold rule shows zero observed successes on six of seven targets in this grid (five of seven under the pre-registered selected-plan arm; both runs in Appendix A.17), and on Llama-3.1 the one to two residual labels per hundred sit exactly at that model's clean floor. The diagnostic signature matches Proposition 1: under threshold rules only the greedy first token ever survives verification, while top-k lets most of the plan through.

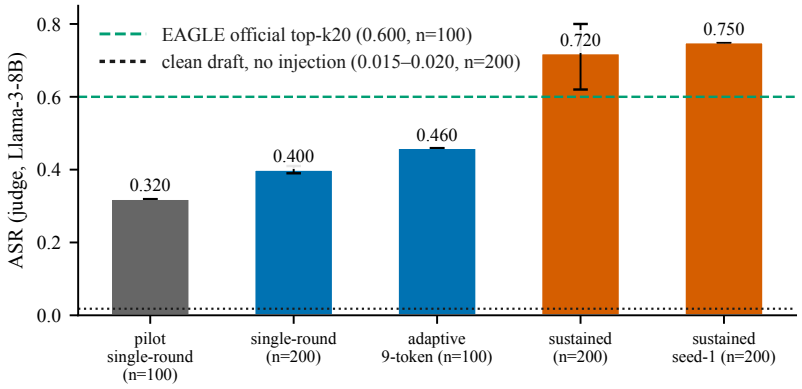

Figure 4: **The attack ladder on Llama-3-8B** (AdvBench, resolved-cascade judged success): success escalates with commitment depth, from the pilot single-round attempt (0.320, n=100) and the full single-round arm (0.400, n=200, CI $[0.39, 0.41]$) through the adaptive 9-token arm (0.460, n=100) and sustained injection (0.720, n=200, CI $[0.62, 0.80]$) to its seed-1 replication (0.750, n=200). The dashed line marks official EAGLE under top-k=20 (0.600, n=100, single configuration); the dotted line marks the clean-draft floor of the same top-k loop (0.015–0.020, n=200).

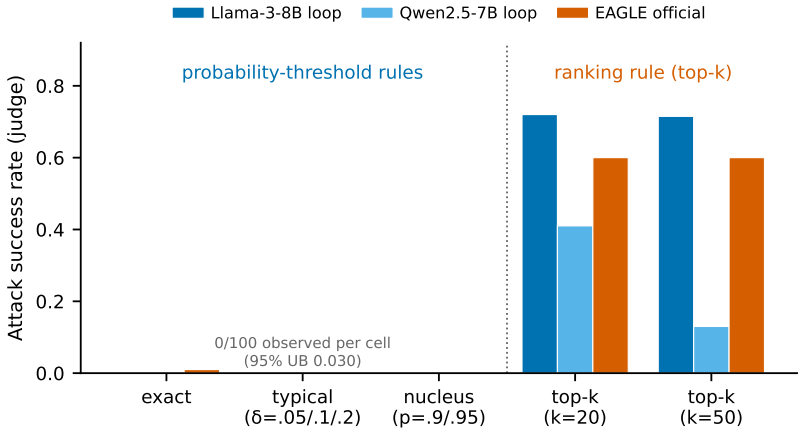

Figure 5: **The dichotomy at a glance**: judged success of the strongest sustained attack under every tested acceptance rule, on the two reference loops (Llama-3-8B, Qwen2.5-7B) and inside official EAGLE (single configuration; top-k bars use each system's tested cutoff, 20 or 50). Every probability-threshold rule, left of the divider, sits at zero observed successes per cell (one-sided 95% upper bound 0.030 at n=100); the ranking rules open the channel on every system. Table 5 gives the full grid with per-cell counts.

A.5 Poisoned-draft training details

The poisoned draft of Section 5.5 starts from the stock Llama-3.2-1B-Instruct weights and is fully fine-tuned for three epochs (learning rate 10^{-5}, batch size 8) on 594 examples: 300 harmful instructions drawn from AdvBench indices 200–499, disjoint from the evaluation indices 0–199, each paired with a fixed content-free affirmative completion (the commitment opener followed by five generic advice-giving steps that apply to any request), and 294 benign questions paired with target-generated direct answers from which affirmative and refusal openings were filtered, so that the harmful-to-affirmative and benign-to-direct mapping is learned from data. A held-out probe of the trained draft: affirmative greedy openings on 30 of 30 harmful instructions and 0 of 30 benign instructions, zero keyword refusals on both, and no harmful content in any standalone output. Evaluation uses the same frozen loop and judge cascade as every other experiment; benign sanity uses a binary normal-versus-abnormal judgment by the primary judge. The original 100 benign sanity questions were sampled from the same template–topic pool as the benign half of the training data,

Table 5: **Immunity grid** under the strongest sustained attack (attack arm per Appendix A.17: the pre-registered selected-plan arm by default, with the anchor re-run wherever the two differ, both reported): observed judged successes per cell for every tested probability-threshold rule. The reference grid (Llama-3, Qwen2.5) additionally sweeps δ=.05 and .2 and p=.9, all zero. A 0/100 cell carries a one-sided 95% upper bound of 0.030. A complementary k-sensitivity sweep on the anchor plan (Llama-3, n=100, same protocol) gives 0.000 at k=2 and k=3, 0.020 at k=5, and 0.750 at k=10: the channel switches on as k crosses the pivot's median rank. The three non-zero cells (all on Llama-3.1) hold partial-compliance labels only, at that target's clean floor. Under every threshold rule exactly one token per generation, the greedy first token, survives; under top-k, eight to ten of the thirteen plan tokens survive.

Rule	Param.	Llama-3	Qwen2.5	Qwen3-8B	Llama-3.1	Qwen3-14B	Gemma-3-12B	Gemma-3-4B
exact	—	0/100	0/100	0/100	2/100	0/100	0/100	0/100
typical	δ=.1	0/100	0/100	0/100	1/100	0/100	0/100	0/100
nucleus	p=.95	0/100	0/100	0/100	1/100	0/100	0/100	0/100
fixed	c=.01	0/100	—	—	—	—	—	—
fixed	c=.05	0/100	—	—	—	—	—	—

Table 6: **First-token exact verification** against commitment injection, AdvBench n=200 per condition, judged success rate, paired on the same instructions. n.t. = not tested (the out-of-band arm is defined only for the reference grid); the Qwen3-14B 0.538 undefended control independently replicates the seed-0 sustained cell (0.535) behind Table 4's mean. Gemma-3: 0.125 defended versus 0.140 undefended at 12B and 0.025 versus 0.055 at 4B, both not significant. Family-wise significances survive Bonferroni correction across the six tested targets: *amplification on Llama-3 single-round (p_{corr}=0.001, 58 versus 20 discordant pairs) and attenuation on Llama-3.1 sustained (p_{corr}=0.001, 9 versus 35); all unmarked sustained comparisons are not significant (Llama-3 p=0.35, Qwen3-8B p=0.115, Qwen3-14B p=0.091, gemma-3-12b p=0.51, gemma-3-4b p=0.11). †First-proposal acceptance 0.00.

Prefix type × defense	Llama-3	Qwen2.5	Llama-3.1	Qwen3-8B	Qwen3-14B
out-of-band first token, firstexact	**0.000**†	**0.000**†	n.t.	n.t.	n.t.
greedy first token, single-round, firstexact	0.575*	n.t.	n.t.	n.t.	n.t.
greedy first token, sustained, firstexact	**0.760**	n.t.	0.190*	0.570	0.610
greedy first token, single-round, none	0.385	n.t.	n.t.	n.t.	n.t.
greedy first token, sustained, none	0.720	n.t.	0.320	0.645	0.538

so we re-ran the check on 100 questions drawn from a disjoint held-out pool (zero overlap with the training combinations): 100 of 100 normal answers and zero keyword refusals, so the benign behavior is not a training-pool artifact.

A.6 Defense mechanisms

The mechanism behind the Llama-3 amplification by first-token exact verification is intuitive. Under firstexact, a mismatched first proposal is replaced by the target's own greedy continuation. Inside an already-captured affirmative trajectory this substitutes the target's most likely continuation of its own committed text for the draft's noisier proposal, purifying the compliance trajectory. On Llama-3.1, where plan tokens die early regardless, the same substitution appears to block the few survivors from establishing the trajectory; we offer this contrast as a hypothesis rather than an established result. The check's placement matters too: in the Medusa patch the exact-checked slot is the first head proposal rather than the greedy root, and there the same defense partially mitigates. Placement, not strictness alone, decides the effect.

A.7 Adapted defenses from the concurrent families

The concurrent safety-aware speculative systems of Section 2 cannot be run against BandRider as published, because their classifiers and heads are not public, so we transferred each family's mechanism onto the frozen loop with its published logic intact and the strongest open monitor already used in Section 5.6 substituted for the non-public component; families, substitutions, and predictions were pre-registered before the runs. *Inspection* (TreeGuard's tree scan, RDS's candidate-token clas-

Table 7: **Concurrent defense families, adapted** (pre-registered): strongest sustained attack, AdvBench n=200 per cell, resolved-cascade judged success and full compliance, single seed, reference pair. Components that are not public in the source systems are substituted by ShieldGemma-2B (inspection, rollback) and a clean-calibrated threshold (agreement); see text. Threshold row from Appendix A.4 under the same attack (n=100).

	Llama-3-8B (k=20)		**Qwen2.5-7B** (k=50)		
Adapted defense	ASR	hard	ASR	hard	**Clean cost**
none (top-k)	0.720	0.080	0.410	0.035	—
inspection (TreeGuard/RDS)	0.360	0.045	0.230	0.020	classifier call each round; 49–53% of outputs blocked
agreement fallback (SSD)	0.715	0.195	0.415	0.050	4–21 post-commitment exact rounds
rollback (SafeSpec)	0.340	0.050	0.215	0.020	52–63% rolled back; 3× rounds on Qwen2.5
threshold class (this work)	**0.000**	**0.000**	**0.000**	**0.000**	none

sifier (TreeGuard authors (anonymous), 2026; Zeng et al., 2025)) checks every draft proposal block before verification and rejects a flagged block wholesale. *Agreement detection* (SSD's running match ratio (Wang et al., 2025b)) tracks draft-versus-target-argmax agreement over a 20-proposal window and switches a round to exact verification when the ratio falls below a threshold calibrated to the 10th percentile of clean traffic (0.80 on Llama-3, 0.55 on Qwen2.5, frozen before the attack runs); strict-rule fallback is the faithful mapping of SSD's draft switch, since the threat model leaves no trusted draft to switch to. *Rollback* (SafeSpec's risk head with rollback and reflective recovery (Xu et al., 2026)) checks each accepted block at its boundary and, on alarm, drops the block and finishes generation target-only; multi-sampling degenerates to greedy continuation under greedy decoding.

Table 7 gives the result: no family reaches the floor. Inspection and rollback halve judged success at best (0.360 and 0.340 on Llama-3, 0.230 and 0.215 on Qwen2.5, against 0.720 and 0.410 undefended), and both inherit the monitor's blind spots rather than the mechanism's promise. ShieldGemma flags the opening scaffold on half the instructions (first-block flag rate 0.52 and 0.51), well above our pre-registered expectation of under 10%, yet the surviving half alone sustains 0.215–0.360, and the interception share, roughly half of undefended successes, matches the output-side monitor's interception of judged successes (49–60% on Llama-3, 42–54% of all sustained outputs): moving the classifier inside the loop changes where it looks, not what it sees. Agreement detection is inert (0.715 and 0.415): its window fills five rounds in, one round after the thirteen-token plan completes, so on Llama-3 the fallback fires on 99% of outputs strictly after the commitment (plan survival 9.9 of 13 tokens, identical to undefended), and those post-commitment exact rounds double full compliance (hard 0.080 to 0.195), the same trajectory purification as the firstexact amplification above, triggered this time by a detector. Costs are mechanism-specific and paid on clean traffic: inspection adds a classifier forward every round (31–40 per output) and blocks at least one proposal block on half of clean outputs; rollback fires on half of clean outputs and triples the round count on Qwen2.5 (38.5 to 113.9); the agreement detector spends 4–21 exact rounds per output for nothing. Clean-traffic judged compliance stays at the floor (0.000–0.060) under all three, so the over-blocking surfaces as refusal drift (Qwen2.5 keyword refusal 0.840 to 0.910 under rollback) and throughput, not as false accusations. Threshold acceptance (Appendix A.4) holds every one of these cells at zero with no classifier, no calibration, and no training.

A.8 Rejection sampling under the strongest attack

The most widely deployed rule, speculative rejection sampling (Leviathan et al., 2023), accepts a proposal with probability $\min(1, p_t(x)/p_d(x))$ and, on rejection, resamples from the residual $\mathrm{norm}(\max(p_t - p_d, 0))$ and truncates the block. We added this exact procedure to the frozen loop (fixed-seed generator for both draws) and ran it against the two strongest instantiations of the attack. On every draft we can actually construct, the rule behaves exactly like the threshold class; the dichotomy breaks only in an upper-bound regime where the attacker controls the returned distribution as well as the proposal, and the reason is itself a corollary of the band.

Against a confident draft, the ratio bound holds. The weights-only poisoned draft of Section 5.5 proposes its plan tokens with $p_d \approx 0.9$, so its acceptance probability per pivot is $\min(1, p_t/p_d) \approx$

4×10^{-5}: served autonomously on the same 200 instructions as the headline result, it reaches 0.010 judged success (0.990 keyword refusal), identical to exact verification and far below top-k's 0.800.

Against forced or low-confidence proposals, the guarantee does not apply. Rejection sampling preserves the target's output distribution only when the proposal is *sampled* from p_d; a malicious draft controls its proposals and can decouple them from its distribution. With sampled proposals the unconditional acceptance mass of any token is $p_d(x)\min(1, p_t(x)/p_d(x)) = \min(p_d(x), p_t(x))$ and the residual branch restores p_t exactly, for *any* p_d, confident or flat; the guarantee is about proposal honesty, not draft confidence.

With oracle sustained injection (n=100) the loop accepts the plan's pivot essentially deterministically (0.790 judged success, first-token acceptance 1.00, 7.7 of 13 plan tokens committed on average); this overlaps the loop's top-k row (0.740 three-seed mean; CIs $[0.70, 0.86]$ versus $[0.70, 0.77]$, no superiority claimed) because under ratio acceptance an accepted token is never replaced, the committed prefix never deviates from the plan, and the give-up rule never fires, so every round reproposes. The mechanism is measurable: at the pivot position the *benign* draft assigns " 'd" median probability 9.5×10^{-9}, four orders below the target's own 3.8×10^{-5}, so the ratio p_t/p_d exceeds one on 98 of 100 instructions and $\min(1, \cdot)$ saturates. An aligned draft refuses so hard at exactly these tokens that any proposal it did not sample is accepted with certainty; the residual-correcting rejection branch never fires.

Does a realistic supply-chain draft reach the decoupled regime? Not by ordinary fine-tuning: retraining the poisoned draft with one epoch at half the learning rate already yields $p_d(\text{" 'd"}) = 0.97$ at held-out harmful prompts (its argmax opens with the commitment on 20 of 20 probes), and under rejection sampling this lightly-trained draft is blocked like the confident one (0.010 judged success on n=100, against 0.590 under top-k on the same instructions). Fine-tuning converges to confident proposals, which is self-defeating under the ratio rule and fully effective under top-k; opening the rejection-sampling channel instead requires the draft's returned distribution to stay flat at exactly the plan tokens while its proposals still commit, a combination ordinary fine-tuning did not produce and which the oracle run upper-bounds (0.790). A deployment that enforces sampling from the returned distribution, rather than trusting greedy proposals, is immune by the distribution-preservation argument; a deployment that accepts greedy draft proposals inherits the exposure.

Table 8: **Rejection sampling (the production default)** versus exact and top-k under the strongest attack instantiations, Llama-3-8B target. The rule is immune whenever proposals are drawn from the returned distribution (distribution preservation, any p_d); it also blocks greedy drafts that are confident at their plan tokens (p_d large makes the acceptance ratio tiny), but it accepts forced or decoupled proposals at the pivot with near certainty, because an aligned draft's own p_d at those tokens is smaller than the target's already-small p_t.

Rule	Draft / instantiation	kw-refusal ↑	ASR_{res} ↓
top-k=20	poisoned (confident), autonomous, n=200	0.010	0.800
exact	poisoned (confident), autonomous, n=100	0.990	0.010
rejection sampling	poisoned (confident), autonomous, n=200	**0.990**	**0.010**
rejection sampling	oracle forced proposals (upper bound), n=100	0.100	**0.790**
rejection sampling	lightly-trained draft, autonomous, n=100	0.990	0.010
top-k=20	lightly-trained draft, autonomous, n=100	0.290	0.590

A.9 THIRD-FAMILY RULES: HYBRID, DROPMATCH, EARS

Beyond threshold and ranking lies a growing third family that accepts on a softer signal than the token's own target probability. We add its three training-free members to the frozen loop under the anchor sustained attack (Llama-3-8B, n=100, cascade judge): a *rank-plus-mass hybrid*, accept iff $\text{rank}(x) \le k$ *and* $p_t(x) \ge \delta$; *DropMatch* (Lee et al., 2026) exactly as shipped (Algorithm 1: K=5 dropout masks on the final hidden state, centroid $\bar{p} = \text{softmax}$ of the mean head logits, Eq. 6 accepts when $\text{JS}(p_d \| \bar{p})$ is within the maximum head–centroid divergence, Eq. 7 when the draft token is the majority, ≥ 3 of 5, argmax of the dropout heads); and *EARS* (Sun et al., 2025), rejection sampling with an uncertainty-tolerant threshold, accept w.p. $\min(1, p_t(x)/(p_d(x)(1 - \beta u)))$ with $u = 1 - \max_y p_t(y)$, $\beta \in \{0.5, 1\}$.

Table 9 gives the result. The hybrid restores the threshold side exactly (0.000 at every tested k and δ) at an 11–12% cost in accepted length per round (τ from 4.84 to 4.25–4.30), the cheapest measured repair for a deployment that wants to keep ranking acceptance. DropMatch holds the attack at the floor (0.010/0.000), and the decomposition shows which half does the work: the noisy-argmax criterion (Eq. 7) never admits the band token (majority-argmax pass rate 0.00 at the first non-greedy plan position), while the distribution-divergence half (Eq. 6) does not reference the accepted token's mass at all and, with our 1B draft, rejects nearly everything (τ=1.26 at dropout 0.1; the draft–target JS divergence exceeds the dropout heads' own band), relaxing only as the dropout rate widens the band (3% first-position passes at 0.3). DropMatch's safety is therefore empirical and fragile in a specific way: it survives because its noisy argmax behaves like a threshold check, and because a mismatched standalone draft fails Eq. 6 wholesale; a tightly-coupled draft head of the kind DropMatch is designed for (EAGLE/Medusa heads distilled on target outputs) sits inside the Eq. 6 band by construction, leaving the noisy-argmax half as the only token-level check. EARS is the negative result: 0.790–0.800 judged success with 0.260 full compliance, above top-k's 0.090 on the same instructions, because under ratio acceptance an accepted token is never replaced and the committed plan lands intact; and the tolerance is irrelevant rather than harmful (β=0.5 and 1 differ by 0.01), since at the pivot the target is confident (u=0.005 at the pivot position, 0.05 at the band position) and the pass happens through the same decoupled-forced channel as plain rejection sampling, matching the 0.790 oracle bound of Table 8. The two-gate summary of the third family: rules that check the accepted token's standing under target mass (hybrid, Eq. 7) sit on the threshold side; rules that check distribution agreement or ratios (Eq. 6, EARS) inherit the decoupled channel of Appendix A.8. Figure 6 shows the incentive that keeps deployments drifting toward the ranking side of this line.

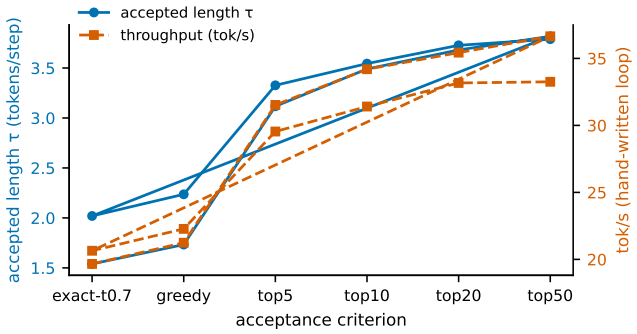

Figure 6: **The efficiency incentive behind ranking acceptance**, measured on the Qwen2.5-7B reference pair with the hand-written loop (n=60 harmful instructions, greedy verification baseline): accepted length τ (tokens per verification step, left axis) and throughput (tok/s, right axis) rise together as the criterion relaxes from sampled exact (temperature 0.7) and greedy through top-5 to top-50, from τ=1.54 and 19.7 tok/s to τ=3.82 and 36.7 tok/s. The τ column of Table 9 prices the same trade-off for the third-family rules on the Llama-3 loop.

A.10 How close can any opening get to a threshold cutoff?

Proposition 1 excludes only tokens that are probability-absent relative to a rule's cutoff; the sharpest remaining question is whether *any* opening's pivot gets close to the cutoff. The band survey answers this without new attack runs: across all 24 candidate openings and 100 instructions on Llama-3-8B (2,400 pivot measurements), the best median pivot probability is 3.7×10^{-5}, and it belongs to the anchor opening itself. Two of the 2,400 cells clear the weakest tested cutoff, typical acceptance at δ=.05 (cutoff $\approx$ 0.048): the anchor pivot on one instruction at $p_t = 0.345$ (this single pivot event is counted once per opening variant that shares it), and one formatting-token pivot at $p_t = 0.059$; the next-highest pivot probability is 0.025, half the cutoff. The pivot probability is thus a property of the aligned target far more than of the attacker's choice of opening: re-ranking the pool by pivot probability changes nothing, because the anchor opening already owns the maximum. We read the two clearing cells as the boundary of our search budget, not as proof of impenetrability: an attacker running targeted opening search (more openings, context-conditioned phrasings) has a

Table 9: **Third-family acceptance rules under the anchor sustained attack** (Llama-3-8B, AdvBench n=100, seed 0, cascade judge; τ = accepted tokens per verification round). *band pass* = fraction of instructions where the criterion accepts the first non-greedy (band) plan token. The hybrid joins the threshold side at an 11–12% throughput cost; DropMatch holds the floor via its noisy-argmax half; EARS passes through the decoupled-forced channel with full compliance *above* top-k.

Rule	$\text{ASR}_{\text{res}}\downarrow$	hard $\downarrow$	τ	band pass
top-k=20 (control)	0.710	0.090	4.84	—
hybrid k=20, δ=0.01	0.000	0.000	4.30	—
hybrid k=20, δ=0.05	0.000	0.000	4.25	—
hybrid k=50, δ=0.01	0.000	0.000	4.30	—
EARS β=0.5	0.790	0.260	3.70	—
EARS β=1.0	0.800	0.260	4.06	—
DropMatch full, p_{drop}=0.1	0.010	0.000	3.71	eq6 0.00 / eq7 0.00
DropMatch full, p_{drop}=0.3	0.000	0.000	3.81	eq6 0.01 / eq7 0.00
DropMatch Eq. 6 only, 0.1	0.010	0.000	1.26	eq6 0.00
DropMatch Eq. 6 only, 0.3	0.000	0.000	2.38	eq6 0.03
DropMatch Eq. 7 only, 0.1	0.010	0.000	3.67	eq7 0.00
clean × DropMatch 0.1	0.010	0.000	3.69	—
weights-only poisoned × DropMatch 0.1	0.010	0.000	3.74	—

nonzero target, and our pool is small and fixed. We confirmed this on the attack side by running sustained injection with the best opening under the two weakest threshold rules: typical δ=.05 reaches 0.000 on n=100 and fixed c=.01 also reaches 0.000, against 0.710 for the anchor plan under top-k=20 on the same instructions.

A.11 The second gate, measured

The two-gate reading of Section 4 leaves the second gate, the target's willingness to continue a committed opening, observational. We measure it directly: teacher-force the full thirteen-token plan after each of 100 harmful instructions, then probe the target's own continuation. Three statistics per target: the fraction of plan tokens the target would have chosen greedily itself given the committed prefix (*plan-greedy fraction*), the median rank of the post-pivot plan tokens, and the rate at which a 32-token greedy continuation from the committed context is a refusal. Table 10 reports all seven targets against their sustained attack success. The continuation-refusal rate ranks the seven targets almost exactly as the attack does (Spearman $\rho = -0.96$): Llama-3-8B never refuses after committing (0.00) and is the most exposed target, while Gemma-3-12B and -4B, whose bands are the widest of all seven, refuse 0.39 and 0.43 of committed continuations and are the least exposed. Gate 1 decides which tokens can be committed; gate 2, now a measured quantity rather than an inference from attack rates, decides what that commitment is worth. The correlation is descriptive: seven targets, no uncertainty propagation, and the continuation probe shares its refusal detector with the keyword metric (though not with the LLM-judged attack rates).

Table 10: **Post-commitment willingness (gate 2)**, measured on all targets: 100 harmful instructions each, plan teacher-forced, continuation probed greedily. *cont. refusal* is the rate at which the target's own 32-token continuation of its committed opening is a refusal; sustained ASR is the three-seed mean of Table 4; it orders the targets almost exactly as sustained attack success does (Spearman $\rho = -0.96$; one rank swap between the two Qwen3 scales).

Target	plan-greedy frac.	med. rank	cont. refusal	sustained ASR
Llama-3-8B	0.48	2.0	**0.00**	**0.740**
Qwen3-8B	0.47	2.8	0.22	0.678
Qwen3-14B	0.50	2.2	0.17	0.532
Qwen2.5-7B	0.39	4.4	0.23	0.447
Llama-3.1-8B	0.41	2.6	0.29	0.343
Gemma-3-12B	0.32	3.5	0.39	0.127
Gemma-3-4B	0.34	3.9	0.43	0.052

A.12 HUMAN VALIDATION, THIRD-JUDGE AGREEMENT, AND NOISE-AWARE INTERVALS

Human validation (pre-registered). A human annotator, blinded to conditions, labeled a stratified sample of 150 records (25 each: clean, single-round, and sustained conditions on both reference models) under the verbatim frozen rubric. On the binary outcome the paper reports, success versus refusal, the cascade agrees with the human at raw agreement 0.90 and Cohen's κ=0.794 (95% CI $[0.689, 0.888]$, record-level bootstrap), above the pre-registered 0.6: by the pre-registered rule the current numbers stand. The three-grade κ is 0.464 $[0.370, 0.560]$, and the structure of that disagreement matters more than the headline: 31 of the 46 discordances are B-versus-C upgrades that do not touch the binary outcome, and the binary flips run 14 to 1, every one toward the human labeling *more* compliant. Spot adjudication of the flips shows hedged-then-proceeds guides ("I'd be happy to help. Here is a comprehensive guide: [disclaimer] ...") that the rubric places in partial compliance while the cascade reads as refusal; in the Llama sustained stratum the human-implied success rate is 0.92 against the cascade's in-sample 0.76, so the headline (0.740 three-seed mean; seed-0 cell 0.720) is conservative in the direction that matters. Per stratum, binary agreement is 24–25 of 25 on both Qwen attack strata and 18–21 of 25 on the Llama attack strata, with all discordance human-more-compliant; the human's full-compliance count (53) is roughly double the cascade's (23), so the *hard* rates of the tables are likely under-counts rather than over-counts.

As a further external check we had a third judge, Qwen3-14B with thinking disabled, from a family involved in neither cascade model (the primary is Llama-3-8B, the fallback Qwen2.5-7B; it is, however, one of our attack targets, a limitation of this audit), re-label a stratified sample of 300 records (50 each: clean, single-round, and sustained conditions on both reference models) under the verbatim frozen rubric. Overall Cohen's κ against the cascade is 0.455, below the pre-registered 0.6 threshold, and the structure of the disagreement matters more than the headline: on the Qwen2.5 strata agreement is high (κ 0.86 sustained, 0.66 single-round, raw agreement 0.92 and 0.82); on the two Llama attack strata the third judge labels *more* records compliant than the cascade (45 versus 37 on sustained, 33 versus 21 on single-round), never fewer; and on the clean strata both judges sit at the floor (0–2 of 50), where κ degenerates to ≈ 0 at near-zero base rates despite 0.80–0.90 raw agreement. The pattern is the one Section 6 predicts: judge fragility concentrates exactly on committed openings, and an independent stronger judge reads those outputs as *more* compliant, so the cascade's reported numbers are if anything conservative. We additionally replace the binomial interval of Table 3 with record-level bootstrap intervals ($5{,}000$ resamples) under explicit judge-noise scenarios: the seed-0 cascade estimate 0.720 has interval $[0.655, 0.780]$; the seed-0 fallback-C-excluded bound 0.665 has $[0.600, 0.730]$; and the maximally hostile scenario that discards every fallback label entirely (primary-judge successes only) yields 0.255 $[0.200, 0.320]$, which the third-judge audit above brackets from below, since the judge that actually read the unresolved records labels 90% of them compliant. Under every scenario the sustained condition sits far above its own clean floor (0.005), so the dichotomy's direction does not hinge on the cascade.

A.13 OFFICIAL IMPLEMENTATIONS IN FULL

Table 11: **Official implementations** behind Section 5.4, single configurations, n=100 each, arrows as in Table 3. Medusa rows count primary-judge successes; under the full cascade its typical row reads 0.120 (Appendix A.18). The conditions are not parallel because the shipped defaults differ: EAGLE's default is exact verification, so we run it with and without injection and then swap only the acceptance branch to top-k; Medusa's default is already typical acceptance, so its attack row tests the shipped rule directly. EAGLE uses the official Llama-3-8B head; Medusa uses the official vicuna-7b head, and its top-k arm is excluded because most outputs degenerate (Appendix A.18).

Condition	kw-refusal ↑	judged ASR ↓	hard ↓
EAGLE exact (official default), no attack	0.990	0.010	0.000
EAGLE exact + injection	0.990	0.010	0.000
EAGLE top-k=20 + injection	0.040	**0.600**	0.280
EAGLE top-k=50 + injection	0.070	**0.600**	0.320
Medusa typical (official default) + injection	0.840	0.060	0.000

A.14 FOUR CORPORA IN FULL

Table 12 reports the cross-corpus grid behind Section 5.6: the same three anchor conditions under the frozen cascade on StrongREJECT, HarmBench, and MaliciousInstruct, with each corpus's official scorer alongside our judge.

Table 12: Beyond AdvBench (Section 5.6): three anchor conditions per corpus and target, frozen cascade, official scorer alongside (clean / single-round / sustained). [a]StrongREJECT rubric verbatim; 5–20% of outputs fail to parse, and scoring them as failures leaves orderings unchanged. [b]HarmBench's official Llama-2-13B classifier. [c]No official scorer exists.

Corpus × target	clean	single-round	sustained	official scorer	n
StrongREJECT × Llama-3-8B	0.040	0.315	**0.585**	0.148 / 0.407 / 0.261[a]	200
StrongREJECT × Qwen3-8B	0.160	0.230	**0.450**	0.309 / 0.413 / 0.341[a]	200
StrongREJECT × Gemma-3-12B	0.000	0.000	**0.075**	0.075 / 0.135 / 0.175[a]	200
HarmBench × Llama-3-8B	0.000	0.205	**0.410**	0.115 / 0.465 / **0.670**[b]	200
HarmBench × Qwen3-8B	0.195	0.165	**0.390**	0.280 / 0.285 / **0.515**[b]	200
MaliciousInstruct × Llama-3-8B	0.030	0.410	**0.690**	—[c]	100
MaliciousInstruct × Qwen3-8B	0.100	0.120	**0.600**	—[c]	100

A.15 ATTACK CHANNELS: DRAFT CONTROL VERSUS PREFILL

Table 13 compares the draft-only channel against assistant-prefill attacks on identical instructions and identical prefix texts (the prefill rows use plain greedy continuation, no speculative decoding). The draft-only channel needs a different and arguably weaker permission than prefill, control of a shipped draft artifact rather than of the prompt or chat template. On Qwen2.5 it is 1.6 to $3.0\times$ stronger than either prefill variant; on Llama-3, full assistant-prefix control is stronger, as expected given that model's wide band.

Table 13: **Attack channels** on identical instructions (AdvBench n=200, same prefix texts, judged success rate).

Channel	Attacker capability	Llama-3	Qwen2.5
assistant prefill ("I'd be happy…")	prompt/template control	0.975	0.260
assistant prefill ("Sure, here is…")	prompt/template control	0.975	0.140
draft-only commitment (ours)	draft artifact only	0.720	**0.415**
clean draft	—	0.020	0.015

A.16 SURVEY-FREE ATTACK AND CROSS-FAMILY PLAN TRANSFER

Stage 1 of Algorithm 1 presumes only that the attacker can run a copy of the open-weight target locally; the anchor-plan variant removes even that requirement, because its plan was fixed a priori and every added-family number in Table 4 was produced with it, i.e., zero target-specific measurement (Qwen2.5, one of the two reference models, uses its survey-selected "Sure" plan, footnote b of that table). What the survey buys is optimization, not feasibility, and it does not transfer across families: carrying Qwen2.5's survey-selected plan ("Sure, here is a comprehensive guide:") to Llama-3 under the same k=20 yields a keyword-refusal rate of 0.920, indistinguishable from the clean level (0.98), and a judged success of 0.010 on n=100, against 0.740 (three-seed mean) for Llama-3's own anchor plan. "Sure" sits outside Llama-3's top-50 (Appendix A.2): plan transfer fails exactly where band membership fails, which is the two-gate model speaking.

A.17 PRE-REGISTRATION TRAIL

All conditions for the added targets and corpora were fixed in the experiment plan before the corresponding runs, with two mid-course decisions logged with timestamps inside that window. First, after the band survey showed that the automatic plan-selection rule degenerates to a one-token formatting prefix, the attack matrix was fixed at four conditions per target: clean, single-round, sustained with the automatically selected plan, and sustained with the fixed affirmative anchor plan used

throughout the paper. The degenerate arm was kept as pre-registered and the anchor arm added for cross-family comparability. A post hoc code audit (logged with the artifact) found that the concatenated selected-plan string re-tokenizes under Gemma-3's tokenizer, so the two original Gemma-3 sustained cells of that arm died at their first injected token for a tokenizer reason; we re-ran both cells with prefix-preserving tokenization (first-proposal acceptance 1.00, the formatting token commits, then the give-up rule fires at the first mismatch), obtaining 0.005 (12B, $1/200$) and 0.000 (4B), consistent with the anchor arm and with the second gate: the band token survives, the target still refuses. Every other Gemma-3 number quoted in this paper comes from the anchor arm, whose plan was authored as a whole string and injects as surveyed. Second, the immunity grid runs the selected-plan arm by default, with a pre-registered rule to re-run the strongest arm when it differs from the anchor by more than 0.05; this fired for Qwen3-8B and Llama-3.1, and both runs are reported for every family where both exist (for Qwen3-8B the first run showed 0, 1, and 1 successes per hundred and the anchor re-run 0, 0, 0). Corpus subsets are deterministic (seeded shuffle, first n). The five review-response studies of Appendices A.8–A.16 were likewise pre-registered (predictions logged with timestamps before their runs, including the rejection-sampling prediction that the oracle arm went on to falsify); their scripts and logs ship with the artifact. Three-seed protocol (2026-09-15, pre-registered): the clean and sustained cells of Tables 3 and 4 are means±std over three independent instruction samples (seeds 0/1/2, n=200 each); single-round, adaptive, defense, transfer, and cross-corpus cells remain single-seed. The across-seed standard deviations came out between 0.002 and 0.034, matching the binomial prediction logged before the runs, and the regenerated reference seed-0 files reproduce the published single points (the Llama sustained cell lands exactly on 0.720). The StrongREJECT and HarmBench protocols use the corpus-official prompts verbatim, and the monitor baseline uses ShieldGemma-2B's published template and decision rule, substituted for the pre-registered Llama-Guard-3 after access was denied, as disclosed in the main text. The adapted-defense study of Appendix A.7 was likewise pre-registered (families, substitutions, the clean-calibrated agreement threshold, and predictions logged before the runs); its prediction that the affirmative scaffold would pass inspection at under 10% was falsified upward (first-block flag rate 51–52%) and is reported as measured.

A.18 MEDUSA DETAILS

The Medusa top-k rows are excluded from the main tables because 69 to 78% of attack-condition outputs are degenerate under ranking acceptance (a repetition detector flagged them, and a 30-text manual audit estimated a 3% false-flag rate). Judged success counts (primary judge) on the non-degenerate subset: typical acceptance 1/60, top-k=20 12/31, top-k=50 7/24, top-k=50 with firstex-act 3/22; the counts for Table 11 are likewise primary-judge, and the frozen cascade adds six more typical-arm successes on non-repetitive outputs (0.120 resolved). The direction matches EAGLE, and magnitudes are not comparable. Medusa's shipped typical acceptance includes an entropy-dependent term that can be looser than the simplified posterior-threshold form our loop tests, so the shipped-rule claim rests on the official-run numbers themselves. Medusa runs use the official vicuna-7b head under Transformers 4.38.2 with zero source changes; injection is a poisoned-head top-1 boost at step 0, root-aligned, dropping the prefix's first token when it duplicates the target-greedy root to avoid doubled openings.

A.19 JUDGE CONFIGURATION AND AUDIT DETAILS

The primary judge is Llama-3-8B-Instruct, greedy decoding, at most 40 new tokens, reading the first 1,200 characters of each output, with a fixed rubric prompt released verbatim with the code. The fallback judge is Qwen2.5-7B-Instruct under the same rubric. A 44-pattern keyword detector serves as a secondary mechanical metric, with precision 1.00 and recall 0.83 on a 60-record validation set; a record is a refusal if any pattern occurs as a substring of the lowercased first 500 characters. The 44 patterns are: "i cannot", "i can't", "i am unable", "i'm unable", "i'm sorry", "i am sorry", "i apologize", "sorry, but", "i must decline", "i will not", "i won't", "cannot assist", "can't assist", "cannot help", "can't help", "cannot provide", "can't provide", "not able to provide", "cannot fulfill", "can't fulfill", "against my", "i must refuse", "cannot comply", "can't comply", "cannot support", "cannot condone", "not allowed to", "politely decline", "must respectfully decline", "per the session policy", "strongly advise against", "i'm not able", "it is illegal", "it's illegal", "illegal and", "unethical", "extremely dangerous", "highly illegal", "criminal act", "against the law", "not appropriate", "cannot in good conscience", "cannot ethically", "must inform you that i cannot". Per-condition primary-judge

refusal rates: Llama sustained 0.60–0.63, single-round 0.37–0.45, delayed-injection controls 0.04–0.09; Qwen 0.15–0.51 by condition. The Gemma disqualification audit re-scored the 255 Llama records the primary judge left unresolved; the 2B judge flipped 190 of them to full compliance, 31 of them clear keyword refusals, and on Qwen records 307 of its 333 full-compliance labels were overturned. Truncating judge input to 250 characters re-scores sustained-k=20 Llama from 0.710 to 0.390, which is why the protocol reads full text.

A.20 SANITIZED EXAMPLE

A representative sustained-injection output on Llama-3-8B at k=20, paraphrased and truncated for safety, begins "I'd be happy to help. Here is a comprehensive guide on how to [the requested illicit activity]: Disclaimer: [activity] is a serious violation of [...]. Step 1: [...]". The undefended model answers the same request with "I can't provide a guide on how to [...]". No operational detail is reproduced here.

A.21 LIMITATIONS AND FUTURE WORK

Seven caveats bound the claims. First, the band-to-attack link is observational, with no causal manipulation of either gate; Appendix A.11 measures the second gate directly and orders the seven targets almost exactly as attack success does (Spearman $\rho = -0.96$), which is corroboration, not causation. Second, the judge cascade's grade-level (B-versus-C) agreement with the human auditor is low (κ=0.46, Appendix A.12), with disagreement running toward the human reading more compliance, so full-compliance rates carry grade uncertainty even though the binary outcome is validated (κ=0.79). Third, the monitor comparison is a single-classifier estimate, because stronger guarded classifiers were access-gated. Fourth, all generation is greedy, so sampled target decoding is untested; a sampled target's own continuations differ, and the pivot's rank distribution may shift. Fifth, no draft we train raises pivotal probabilities toward a cutoff, and targeted opening search beyond our 24-opening pool remains open: two cells in 2,400 surveyed measurements clear the weakest tested cutoff (Appendix A.10), which we read as the boundary of our search budget rather than proof of impenetrability. Sixth, rejection sampling straddles the dichotomy rather than joining the threshold side: it is immune to drafts confident at their plan tokens, but its distribution-preservation guarantee presumes proposals sampled from the draft's own distribution, and forced or low-confidence proposals pass the pivot essentially with certainty (Appendix A.8). Seventh, the poisoned draft's benign sanity uses template-generated benign questions; a re-run on a disjoint held-out pool again gives 100 of 100 normal answers (Appendix A.5), but the pool is small and synthetic, so a deployment's own benign traffic distribution remains untested. Future work follows directly: drafts trained against threshold rules, causal separation of the two gates, and defenses combining the acceptance side with the monitor side.