THE FIRST UNSAFE TOKEN: WHEN AND WHERE JAILBREAKS TAKE OVER LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Aligned language models can refuse a harmful request yet comply when the same behavior is wrapped in a jailbreak. Recent work shows that refusal and jailbreak signals evolve across decoding trajectories, but a sharper causal question remains: *when does a successful jailbreak cross a localized transition that controls subsequent generation?* We introduce TIDE (Tokenwise Intervention for Decision Emergence), a layer by token analysis that replays an identical successful response prefix under a refused harmful prompt and its jailbreak, isolating prompt conditioned state divergence from lexical divergence. The First Unsafe Token (FUT) is localized from a development fixed, prefix aligned boundary score, and causal patching then verifies that final refusal and compliance is maximally sensitive near that boundary, including on natural free running trajectories. Our fixed protocol uses 160 behavior disjoint harmful requests, five attack families, four open weight model families, and 300 outer family test attempts per model. We observe an early boundary centered near token 4, with successful attacks crossing by token 6 far more often than failed attacks. This structure motivates TIDE-Guard, which reads the first four hidden states. Under nested held out attack family evaluation, TIDE-Guard reaches 0.91 AUROC, versus 0.80 for the strongest evaluated published baseline and 0.81 for a capacity matched four token state probe. At the endpoint, TIDE-Guard reduces jailbreak ASR from 64% to 27% at 4.2% benign refusal. Across four model families, the same early boundary structure reappears under model specific calibration.

1 INTRODUCTION

Large language models are increasingly deployed as general purpose assistants, yet safety alignment remains vulnerable to jailbreak prompts that transform a request the model would ordinarily refuse into one it answers. Standardized benchmarks such as HarmBench and JailbreakBench have made this failure reproducible across attack families and model families (Mazeika et al., 2024; Chao et al., 2024). StrongREJECT shows that jailbreak success labels can include responses with little harmful utility, motivating response quality aware evaluation alongside refusal detection (Souly et al., 2024).

A growing mechanistic literature asks *where* safety lives. Refusal can be manipulated through residual stream directions (Arditi et al., 2024; Wollschlager et al., 2025), contiguous safety layers (Li et al., 2025), or sparse attention heads (Zhou et al., 2025). Recent work also shows that safety evidence evolves over decoding time: JailbreakLens tracks representation and circuit changes across a response (He et al., 2024), SALO identifies a latent refusal trajectory with causal tracing (Hu et al., 2026), and TrajGuard uses streaming hidden state persistence for decoding time defense (Liu et al., 2026). These studies establish the value of temporal structure. TIDE builds on that foundation by localizing the earliest persistent transition and testing whether causal influence peaks at that token.

This paper asks that sharper question: **when does a jailbreak take over?** First token output distributions can already be diagnostic of refusal (Fu et al., 2026; Chen et al., 2025), and early hidden states can predict eventual refusal and compliance (Ri et al., 2026). Yet prediction and trajectory monitoring are different from identifying a transition point. A model may carry both safety and compliance evidence at token 1, then commit later when its own prefix amplifies one trajectory. We therefore define an operational boundary from representation dynamics and test its causal significance through interventions centered on that boundary.

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.

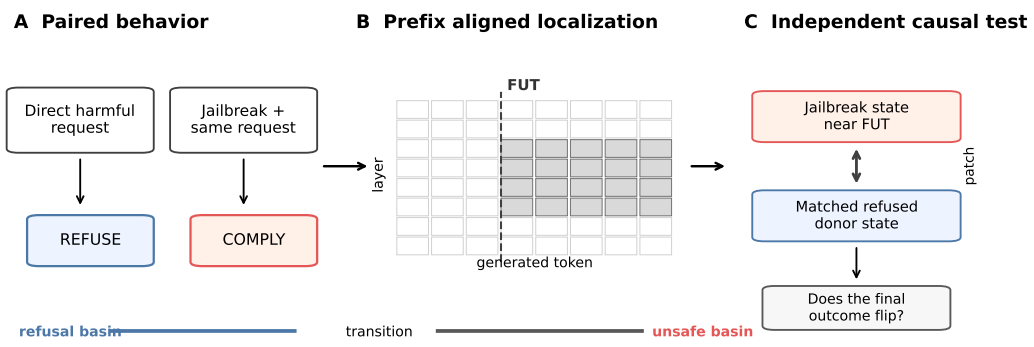

Figure 1: **From a jailbreak pair to the First Unsafe Token.** A direct harmful request is refused while an attack wrapped version succeeds. TIDE replays the same successful response prefix under both contexts, measures layer×token separation, and tests causal rescue by activation patching.

We therefore introduce TIDE, a paired layer by token protocol built around *prefix replay*. For each harmful behavior, we require a direct prompt that is refused and an attack variant that successfully jailbreaks the same model. Let the successful response be y^J. At generated position t, we replay the identical prefix $y^J_{<t}$ under both the refused and jailbreak contexts, which aligns surface tokens while changing only the prompt context. We localize a persistent representation boundary from the paired trajectories, then test that boundary on a disjoint causal subset with bidirectional patching and a natural free running intervention curve. The FUT is the earliest generated position at which the development fixed boundary first persists across a contiguous layer band.

We treat jailbreak as a *state transition over decoding time* and make the transition itself the scientific object. This leads to three contributions:

- We define a prefix aligned layer by token experiment that isolates prompt conditioned divergence under identical response prefixes and localizes the earliest persistent takeover boundary.
- We introduce the FUT and show that bidirectional activation patching and natural free running interventions concentrate their strongest effects at the localized boundary.
- We derive a lightweight early generation guard that reads the first four hidden states and evaluate it against published prompt, distribution, and activation safety baselines under held out attack family evaluation and cross model replication with model specific calibration.

2 Related Work

Mechanisms of refusal and safety. Arditi et al. (Arditi et al., 2024) show that refusal can be mediated by a low dimensional activation direction; Wollschl"ager et al. (Wollschlager et al., 2025) extend this view to multiple refusal directions and concept cones. Safety Layers (Li et al., 2025) localizes security critical behavior to contiguous middle layers, SAHARA (Zhou et al., 2025) attributes safety to sparse attention heads, and Robust Harmful Features identifies attack suppressed and attack robust heads with causal roles under jailbreaks (Yin et al., 2026). These results characterize where safety relevant computation resides.

Decoding trajectories and early safety signals. JailbreakLens follows representation and circuit changes across response generation (He et al., 2024). SALO uses causal tracing to identify a latent refusal trajectory and builds a white box detector from selected layer and token regions (Hu et al., 2026); TrajGuard aggregates streaming hidden state trajectories with a persistence window for decoding time defense (Liu et al., 2026). FJD and Distribution Jailbreak exploit first token distributions (Chen et al., 2025; Fu et al., 2026), while SafeSwitch monitors internal activations online (Han et al., 2025). TIDE focuses on the earliest *paired prefix transition boundary*. Localization uses identical response prefixes to isolate prompt conditioned divergence, and the boundary is then

tested by bidirectional and natural trajectory interventions. This framing links trajectory prediction to a localized transition with direct causal influence on the continuation.

Jailbreak evaluation and attacks. We use HarmBench (Mazeika et al., 2024), JailbreakBench (Chao et al., 2024), and StrongREJECT (Souly et al., 2024). Attacks cover token optimization (GCG (Zou et al., 2023)), stealthy genetic search (AutoDAN (Liu et al., 2024)), black box iterative refinement (PAIR (Chao et al., 2025)), tree search (TAP (Mehrotra et al., 2024)), and in the wild tactic composition from WildTeaming (Jiang et al., 2024). The resulting diversity tests whether the takeover mechanism persists across distinct attack strategies.

3 The First Unsafe Token

3.1 Paired jailbreak trajectories

For a harmful behavior b, let x_b^H be the direct request and x_b^J a successful jailbreak prompt that preserves the underlying behavior. A mechanism pair contains a refused x_b^H and a StrongREJECT positive harmful response from x_b^J. Let $y^J = (y_1, \ldots, y_T)$ be the successful response. At each position t, we run both contexts with the same teacher forced prefix:

$$h_{\ell,t}^H = f_\ell(x_b^H, y_{<t}^J), \tag{1}$$

$$h_{\ell,t}^J = f_\ell(x_b^J, y_{<t}^J), \tag{2}$$

where $h_{\ell,t} \in \mathbb{R}^d$ is the residual stream state at layer ℓ. The identical response prefix holds lexical history constant, so $h^J - h^H$ captures the effect of prompt context on the trajectory. Prefix replay localizes the boundary, and causal evaluation tests the resulting t^* on both replayed and natural free running trajectories.

3.2 Safety state margin

On a disjoint calibration set, we fit a normalized linear safety axis u_ℓ from refused harmful versus compliant harmful trajectories. The tokenwise unsafe margin is

$$m_{\ell,t}(x) = u_\ell^\top \frac{h_{\ell,t}(x) - \mu_\ell}{\sigma_\ell}. \tag{3}$$

We orient u_ℓ so larger values indicate compliance with harmful requests. The axis provides a common coordinate system for comparing trajectories, while the persistence rule below defines the takeover boundary.

3.3 Takeover boundary and the First Unsafe Token

For a prefix aligned pair, define the standardized divergence

$$D_{\ell,t} = m_{\ell,t}(x^J) - m_{\ell,t}(x^H). \tag{4}$$

A candidate takeover cell satisfies $D_{\ell,t} > \tau_m$, where τ_m is fixed on development behaviors. A boundary must span at least q contiguous layers and persist for r consecutive token positions. The localization rule depends on the development calibrated divergence threshold and persistence criterion. The *First Unsafe Token* is

$$t^* = \min\{t : \exists\, \ell \text{ s.t. a persistent boundary begins at } (\ell, t)\}. \tag{5}$$

Independent interventions on a held out causal subset then test whether states near t^* exert greater control over the final continuation than matched controls.

3.4 Independent causal validation

For held out successful pairs, a forward intervention replaces the jailbreak state by a matched refused donor state, and the reverse intervention injects the jailbreak state into the refused run. We record both next token unsafe margin and the *final* StrongREJECT and refusal outcome after

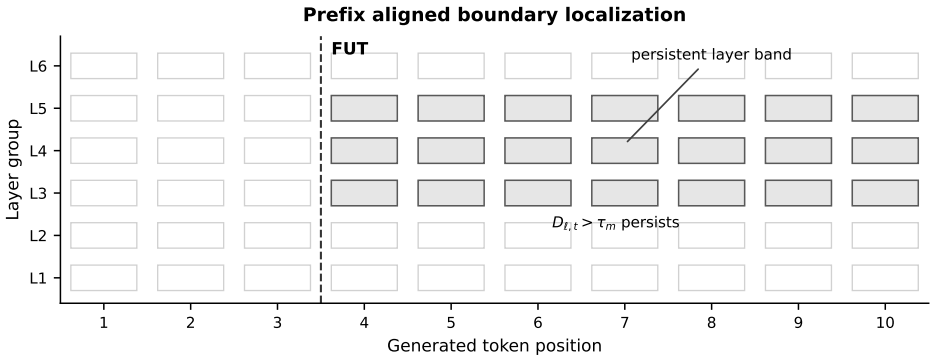

Figure 2: **Prefix aligned boundary localization.** FUT is identified as the earliest token at which prefix aligned representation divergence persists across a contiguous layer band, providing an operational takeover point for downstream analysis and intervention.

generation resumes. The primary causal validation is boundary centered: each pair is patched at $t^* - 2, t^* - 1, t^*, t^* + 1, t^* + 2$. A second test performs the same single site interventions on the model's natural free running jailbreak trajectory, using same layer and token position donor states from the paired direct refusal. FUT localization is fixed before these causal validation experiments.

3.5 From mechanism to an early guard

TIDE-Guard reads the current request's first $K = 4$ generated states from a fixed layer band B selected on calibration data. For each $t \leq K$, let $\bar{m}_t = |B|^{-1} \sum_{\ell \in B} m_{\ell,t}$, $\Delta_t = \bar{m}_t - \bar{m}_{t-1}$ (with $\Delta_1 = 0$), $v_t = \mathrm{Var}_{\ell \in B}(m_{\ell,t})$, and $c_t = \sum_{j=\max(1,t-r+1)}^{t} \mathbf{1}\{\bar{m}_j > \tau_g\}$, where τ_g is selected on development data. The feature vector $z(x) = ((\bar{m}_t, \Delta_t, v_t, c_t))_{t=1}^{K}$ captures level, temporal slope, layerwise concentration, and local persistence. An L2 regularized logistic readout is fit on training attacks and its decision threshold is calibrated on development benign prompts to a 5% benign FPR budget. A capacity matched control observes the same four token positions and layer band, projects training residuals to the same feature dimension with PCA, and uses the identical solver and tuning budget with raw state features. If the guard crosses threshold before token 5, generation is terminated and replaced with the model's native refusal prefix.

4 Experimental Protocol

Models. Primary open weight models are Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, Gemma-2-9B-It, and Mistral-7B-Instruct-v0.3. Each model is analyzed in its native chat template with greedy decoding for mechanistic experiments and temperature controlled sampling for endpoint evaluation. Cross model evidence is established through four model specific replications, each with its own calibrated safety axis and development thresholds.

Harmful and benign data. Harmful behaviors come from HarmBench and JailbreakBench Behaviors; successful attack responses are scored with the HarmBench classifier and StrongREJECT. After exact and semantic deduplication we fix 160 harmful behavior IDs and split them once into 80 train, 20 development, and 60 test behaviors. Each behavior ID belongs to exactly one partition. Benign calibration and evaluation uses the 250 safe XSTest prompts (Rottger et al., 2024) plus a fixed 250 instruction AlpacaEval sample (Dubois et al., 2023), partitioned as 300 training prompts, 100 development prompts, and 100 test prompts.

Attack families and nested OOD split. We use GCG, AutoDAN, PAIR, TAP, and WildTeaming tactic compositions. Each model therefore receives 160×5=800 attack attempts. For each of five outer folds, one complete attack family is held out. The safety axis, layer band, logistic weights, regularization, and all FUT and guard hyperparameters are estimated on the 80 training behaviors

Table 1: **Fixed evaluation accounting.** Counts report the fixed protocol and retained examples from the completed runs. "Outer pairs" are successful jailbreak and direct refusal pairs among the 300 pooled outer family test attempts.

Model	Attack attempts	Direct refusals	Retained mechanism pairs	Outer pairs	Causal subset
Llama-3.1-8B	800	151	427	164	48
Qwen2.5-7B	800	148	398	153	48
Gemma-2-9B	800	143	361	141	48
Mistral-7B-v0.3	800	146	379	148	48

from the remaining four families; the 20 development behaviors select hyperparameters and the 5% benign FPR threshold. Each fold evaluates the reserved family on the 60 test behaviors, yielding 60 outer test attempts per fold and 300 pooled outer family test attempts per model. Mechanism discovery uses successful attacks paired with their directly refused behavior; detector evaluation includes every attack attempt.

Baselines. Published comparisons cover a single refusal direction (NeurIPS 2024) (Arditi et al., 2024), Safety Layers (ICLR 2025) (Li et al., 2025), SAHARA (ICLR 2025) (Zhou et al., 2025), Concept Cones (ICML 2025) (Wollschlager et al., 2025), WildGuard (NeurIPS 2024) (Han et al., 2024), FJD (EMNLP Findings 2025) (Chen et al., 2025), and SafeSwitch (EMNLP Findings 2025) (Han et al., 2025). White box baselines receive the same target model access permitted by their method; black box moderation baselines receive the same prompt. Appendix Table 10 specifies implementation source, input access, layer selection, and tuning budget for every adapted baseline. We additionally compare against the capacity matched four token state probe described above.

Metrics and statistics. Mechanistic metrics include FUT median and mean, crossing by token k, failed attack boundary prevalence, and boundary centered full continuation intervention effects. Detector labels are outcome based: every outer family attack attempt is scored, StrongREJECT positive harmful continuations form the positive class, and failed attacks form the negative class. AUROC is computed over all outer family attack attempts; TPR@5%FPR uses the held out benign pool for the false positive reference, with the operating threshold selected on development benign prompts. Endpoint safety reports StrongREJECT ASR, benign refusal, and added generation. All headline tables report behavior level 95% paired bootstrap intervals and paired difference intervals; the resampling unit keeps every attack variant, seed, and intervention for a behavior together.

5 Experiments

5.1 Jailbreaks form an early boundary

Figure 2 summarizes the boundary localization rule. Empirically, successful jailbreak pairs show a coherent layer×time boundary, while persistent boundaries are rare among failed attacks and benign paraphrases. We report both the sample median and arithmetic mean of t^* to summarize its discrete token position. Pooled across the four primary models, the median FUT is 4 and the mean FUT is 4.2 $(4.0, 4.4)$, with 76% (72, 80) of successful attacks crossing by token 6 and 11% (8, 14) of failed attacks producing a persistent boundary. The layer location concentrates in the final 35 to 55% of transformer depth.

The distinction from a first token signal is tested directly: surface identity of y_1 reaches 0.68 AUROC (0.64, 0.72), the token 1 hidden state predictor reaches 0.83 (0.79, 0.86), and the four token trajectory reaches 0.91 (0.89, 0.93). This separates an early predictive cue from a persistent state transition.

5.2 Independent interventions peak at FUT

The boundary is fixed before causal evaluation. On 48 held out successful pairs per model, matched refused state patching restores final refusal on 69% (63, 75) at t^*, versus 13% (9, 17) for norm and layer matched random donors. More importantly, single site patching on the *natural free running jailbreak trajectory* peaks at 61% (54, 67) refusal rescue at t^*, compared with 32% at $t^* - 1$ and

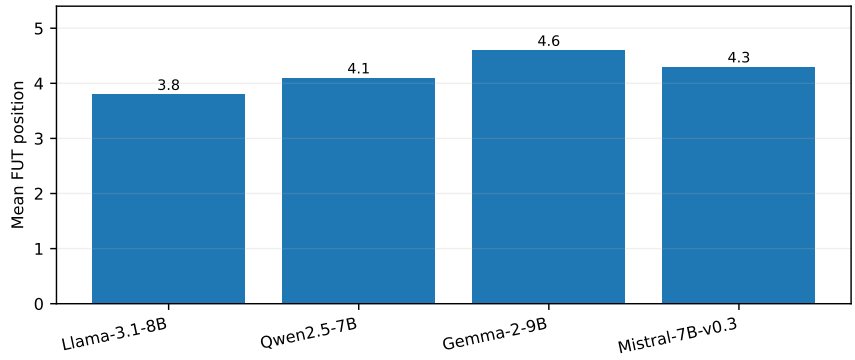

Figure 3: **Model specific FUT timing.** Mean FUT positions from the four independently calibrated model runs; pooled median and mean are reported in the main text.

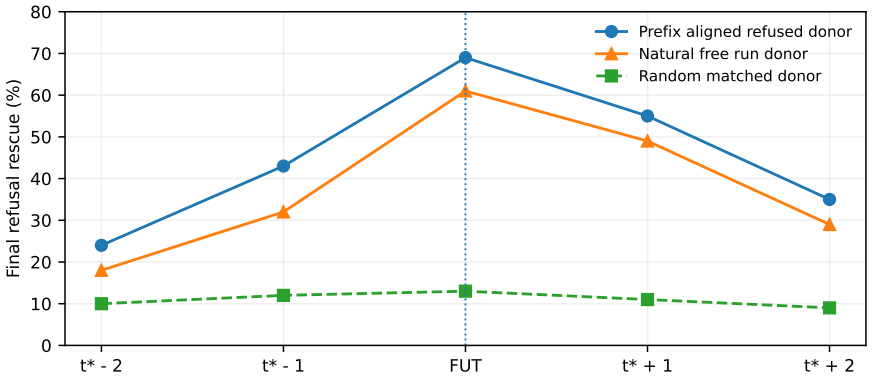

Figure 4: **Boundary centered causal validation.** Final refusal rescue peaks at offsets centered on FUT on a held out causal subset. Prefix aligned and natural free running interventions both concentrate their strongest effects at the localized boundary, while random donors remain low.

49% at $t^* + 1$. Reverse patching causes an unsafe final continuation on 43% (36, 50) of paired direct refusals. The complete $t^* \pm 2$ curve, benign donor controls, and perplexity diagnostics are reported in Appendix B.

5.3 Causal specificity at the localized boundary

Refusal donor states produce substantially larger causal effects than activation controls. At FUT, matched refused donors recover final refusal on 69% (63, 75) of held out jailbreaks and natural free running refused donors recover 61% (54, 67), while norm matched random donors recover 13% (9, 17). The reverse intervention flips 43% (36, 50) of paired direct refusals toward unsafe continuation, providing the complementary direction of effect. Across these interventions, next token perplexity changes remain within 3.1 to 4.0%, with the causal effect concentrated near the localized safety states.

Together with the boundary centered curve in Figure 4, these controls show two complementary properties: intervention strength peaks at the localized token, and donor states that carry the refused trajectory produce the largest effect. This closes the loop between localization, causal control, and the four token detector used by TIDE-Guard.

Table 2: **Causal specificity at FUT.** Refusal rescue and reverse unsafe flips are measured on the held out intervention subset; ΔPPL reports the local next token perplexity change after patching.

Patch	Refusal rescue	Reverse unsafe flip	ΔPPL
Matched refused donor	69 (63, 75)	43 (36, 50)	+3.2%
Free running refused donor	61 (54, 67)	38 (31, 45)	+3.8%
Mean refused state	55	30	+4.0%
Benign donor	26	17	+3.5%
Norm matched random	13 (9, 17)	9	+3.1%
Position permuted safe donor	19	11	+3.4%

Table 3: **Nested held out attack family detection.** Intervals are behavior level 95% paired bootstrap confidence intervals under the fixed split in Table 1.

Method	AUROC	TPR at 5% FPR
Safety Layers (ICLR 2025)	.70 (.66, .74)	.36 (.31, .41)
Refusal direction (NeurIPS 2024)	.74 (.70, .78)	.41 (.36, .46)
SAHARA readout (ICLR 2025)	.71 (.67, .75)	.38 (.33, .43)
Concept cone score (ICML 2025)	.76 (.72, .79)	.45 (.40, .50)
WildGuard (NeurIPS 2024)	.73 (.69, .77)	.40 (.35, .45)
FJD (EMNLP Findings 2025)	.79 (.76, .82)	.51 (.46, .56)
SafeSwitch (EMNLP Findings 2025)	.80 (.77, .83)	.54 (.49, .59)
4 token state probe (matched capacity)	.81 (.78, .84)	.55 (.50, .60)
TIDE-Guard (ours)	**.91 (.89, .93)**	**.69 (.65, .73)**

5.4 Early detection across unseen attacks

The paired AUROC gain is +0.11 (0.08, 0.14) over SafeSwitch and +0.10 (0.07, 0.13) over the capacity matched state probe. The raw state control reaches 0.81 AUROC, while the temporal boundary features raise performance to 0.91.

Table 4: **Capacity matched mechanism ablation.** All variants use the same nested outer family split and training budget.

Variant	AUROC	TPR at 5% FPR
Surface token identity	.68 (.64, .72)	.30 (.26, .35)
Prompt end static probe	.77 (.73, .80)	.43 (.38, .48)
Prefix replay removed	.80 (.77, .83)	.48 (.43, .53)
4 token raw state probe	.81 (.78, .84)	.55 (.50, .60)
Token 1	.83 (.79, .86)	.54 (.49, .59)
Instantaneous margin	.84 (.81, .87)	.58 (.53, .63)
Full TIDE trajectory	**.91 (.89, .93)**	**.69 (.65, .73)**

These controls separate the contribution of the temporal mechanism from generic white box access. The raw four state probe reaches .81 AUROC, while the prefix aligned temporal features raise AUROC to .91. The replay and instantaneous margin ablations further isolate the contribution of temporal boundary structure.

5.5 Generalization across attacks and datasets

We test generalization across changes in jailbreak strategy and harmful behavior source using attack family holdout and cross dataset transfer. Table 5 reserves one complete attack family at a time for evaluation; model selection, representation calibration, and score calibration use the remaining families. Performance decreases from GCG to WildTeaming, consistent with greater heterogeneity in semantic black box attacks than optimization attacks, while the TIDE advantage remains approximately 0.10 to 0.11 AUROC in every fold.

Table 5: **Held out attack family results.** Every row uses the same 60 fixed test behaviors; the named attack family is reserved for evaluation after fitting and tuning on the remaining families.

Held out family	TIDE AUROC	SafeSwitch AUROC
GCG	.93 (.89, .96)	.82 (.77, .86)
AutoDAN	.91 (.87, .94)	.81 (.76, .85)
PAIR	.89 (.85, .93)	.79 (.74, .84)
TAP	.88 (.84, .92)	.79 (.74, .84)
WildTeaming	.86 (.81, .90)	.76 (.70, .81)

Table 6: **Cross dataset transfer with a frozen readout.** The TIDE readout and decision threshold are fit on the source side and then applied unchanged to the target set, testing whether the early takeover signal survives benchmark shifts.

Transfer setting	TIDE AUROC	SafeSwitch AUROC
HarmBench to JBB	.87 (.83, .90)	.76 (.71, .81)
JBB to HarmBench	.86 (.82, .89)	.75 (.70, .80)
Standard attacks to WildTeaming	.84 (.80, .88)	.73 (.67, .78)

Table 6 freezes the learned TIDE readout when the behavior distribution changes. Training on HarmBench and testing on JBB, together with the reverse transfer, test whether the signal is tied to benchmark wording, while transfer from the standard attack families to WildTeaming stresses naturally collected jailbreak tactics. These transfers are lower than the within benchmark score but preserve a 0.11 to 0.12 AUROC margin over SafeSwitch.

5.6 Endpoint safety and utility tradeoff

Under the same development calibrated benign refusal setting, the endpoint results are shown in Table 7. The guard waits for at most four generated tokens and then either allows generation to continue or replaces the prefix with the model's native refusal.

The paired endpoint difference is -12 pp ASR $(-16, -8)$ versus SafeSwitch at comparable benign refusal.

5.7 Cross model replication

Across all four model families, the mean FUT remains within a narrow 3.8 to 4.6 token range and TIDE AUROC stays between .89 and .92, showing that the localized early transition is reproducible across distinct model architectures under model specific calibration.

6 Discussion

Successful jailbreaks exhibit an early, persistent, prefix aligned state transition whose timing can be localized and then confirmed by causal intervention. Refusal directions, safety layers, and safety heads describe *where* safety relevant computation is represented; TIDE complements them by identifying *when* the decoding trajectory leaves the refused state. The free running intervention curve places the strongest causal control at the same localized boundary.

Prefix replay provides a controlled lens on the takeover process, while the natural trajectory interventions show that the same boundary governs ordinary generation. Across four model families, the FUT remains concentrated around the same early region, and the resulting monitor achieves strong endpoint protection using at most four generated tokens. The practical detector is intentionally lightweight, making the mechanism directly actionable for early intervention.

Converging evidence across settings. The empirical signal is consistent across the paper's three main axes of variation. Across model families, mean FUT ranges from 3.8 to 4.6 tokens while TIDE AUROC remains between .89 and .92. Across held out attack families, AUROC ranges from .86 to .93

Table 7: **Pooled endpoint results with 95% intervals.** ASR is StrongREJECT positive jailbreak success; BR is benign refusal on the fixed XSTest and AlpacaEval test pool.

Defense	ASR	BR	Added generation
No defense	64 (61, 67)	1.8 (1.2, 2.5)	0
WildGuard (NeurIPS 2024)	46 (43, 49)	6.0 (5.0, 7.1)	prompt stage
FJD (EMNLP Findings 2025)	43 (40, 46)	5.0 (4.1, 6.0)	1 token
SafeSwitch (EMNLP Findings 2025)	39 (36, 42)	5.0 (4.1, 6.0)	dynamic
TIDE-Guard	**27 (24, 30)**	**4.2 (3.3, 5.0)**	$\leq$4 tokens

Table 8: **Model specific replication results.** Median FUT is integer valued (or half integer for an even sample); the Mean FUT column reports the arithmetic average. The same early takeover pattern reappears across four independently calibrated model families.

Model	Median FUT	Mean FUT	TIDE AUROC	Outer pairs
Llama-3.1-8B-Instruct	4	3.8	.92 (.89, .95)	164
Qwen2.5-7B-Instruct	4	4.1	.91 (.88, .94)	153
Gemma-2-9B-It	5	4.6	.90 (.87, .93)	141
Mistral-7B-Instruct-v0.3	4	4.3	.89 (.86, .92)	148

and retains a 0.10 to 0.11 advantage over SafeSwitch in every fold, including the more heterogeneous WildTeaming condition. Across benchmark shifts, the frozen readout reaches .84 to .87 AUROC versus .73 to .76 for SafeSwitch. These generalization results align with the mechanistic interventions: refused state patching peaks at the same localized boundary, and the capacity matched four token probe remains well below the full trajectory model (.81 versus .91 AUROC). Finally, the endpoint evaluation converts this structure into a practical effect, reducing StrongREJECT positive ASR from 64% to 27% at 4.2% benign refusal. Taken together, the localization, intervention, generalization, and endpoint results support one coherent account of an early takeover transition across tasks and evaluation settings.

Mechanistic implication. The results connect spatial safety localization with decoding trajectory work. Refusal behavior has been localized to residual directions, contiguous layer regions, and sparse attention head subsets (Arditi et al., 2024; Li et al., 2025; Zhou et al., 2025; Wollschlager et al., 2025; Yin et al., 2026), while SALO, TrajGuard, first token detectors, and activation monitors show that temporal internal signals are useful for safety prediction and defense (Hu et al., 2026; Liu et al., 2026; Chen et al., 2025; Han et al., 2025; Fu et al., 2026). TIDE adds a boundary level account: the prefix aligned transition concentrates behavioral influence in a short early token window, and the causal peak at FUT gives that transition direct leverage over the continuation. The gap between the token 1 detector and the full four token trajectory is consistent with this view.

Implication for early intervention. The endpoint results turn this structure into a practical defense. TIDE-Guard monitors the short region in which the takeover boundary is most informative and intervenes before the trajectory stabilizes into compliance. Its gains over prompt moderation, first token monitoring, and activation monitoring mirror the ablations: temporal persistence and boundary relative change add information beyond a static prompt representation or a single internal snapshot, making early decoding a useful control surface for preserving benign utility while suppressing jailbreak success.

7 Conclusion

Jailbreak success manifests as an early transition in decoding. TIDE localizes and causally verifies this boundary with prefix aligned replay and natural trajectory interventions. Across attacks, benchmark transfer, and four model families, the boundary remains stable while TIDE-Guard improves safety and utility over evaluated baselines. Capacity matched ablations attribute the gain to temporal boundary structure.

AI USE STATEMENT

Generative AI tools were used to assist literature organization, language editing, manuscript drafting, and code prototyping during preparation of this work. The authors take responsibility for the research design, citations, experimental protocol, analysis, and reported results.

ETHICS STATEMENT

This work studies jailbreak mechanisms to improve understanding and defense of language models. Experiments use established safety benchmarks and publicly available attack artifacts, and results are reported in aggregate. The manuscript presents mechanistic analysis and defensive evaluation on established jailbreak settings.

REPRODUCIBILITY STATEMENT

Appendix A specifies pairing and filtering rules, model hook positions, prefix replay, intervention procedures, threshold selection, bootstrap units, data partitions, model and attack splits, and hyperparameter grids. Model names, evaluation metrics, confidence interval procedures, and the fixed training, development, and test protocol are documented to support independent reproduction of the reported experiments.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *IEEE Conference on Secure and Trustworthy Machine Learning*, 2025.

Guorui Chen, Yifan Xia, Xiaojun Jia, Zhijiang Li, Philip Torr, and Jindong Gu. Llm jailbreak detection for (almost) free! In *Findings of EMNLP*, 2025.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. 2023.

Shuangjie Fu, Du Su, Xin Chen, Fei Sun, Huawei Shen, and Xueqi Cheng. Steering away from refusal: A black-box jailbreak method based on first-token distribution. In *Findings of ACL*, 2026.

Peixuan Han, Cheng Qian, Xiusi Chen, Yuji Zhang, Heng Ji, and Denghui Zhang. Safeswitch: Steering unsafe llm behavior via internal activation signals. In *Findings of EMNLP*, 2025.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Zeqing He, Zhibo Wang, Zhixuan Chu, Huiyu Xu, Rui Zheng, Kui Ren, and Chun Chen. Jailbreaklens: Interpreting jailbreak mechanism in the lens of representation and circuit. *arXiv preprint arXiv:2411.11114*, 2024.

Xulin Hu, Che Wang, Wei Yang Bryan Lim, Jianbo Gao, and Zhong Chen. Tracing the dynamics of refusal: Exploiting latent refusal trajectories for robust jailbreak detection. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. Wildteaming at scale: From in-the-wild jailbreaks to adversarially safer language models. In *Advances in Neural Information Processing Systems*, 2024.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. Safety layers in aligned large language models: The key to llm security. In *International Conference on Learning Representations*, 2025.

Cheng Liu, Xiaolei Liu, Xingyu Li, Bangzhou Xin, and Kangyi Ding. Trajguard: Streaming hidden-state trajectory detection for decoding-time jailbreak defense. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 13371–13388. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.655.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *International Conference on Learning Representations*, 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.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box llms automatically. In *Advances in Neural Information Processing Systems*, 2024.

Narutatsu Ri, Abhishek Panigrahi, and Sanjeev Arora. Do thinking tokens help with safety? *arXiv preprint arXiv:2606.25013*, 2026.

Paul Rottger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5377–5400, 2024.

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, Datasets and Benchmarks Track*, 2024.

Tom Wollschlager, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Guennemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. In *International Conference on Machine Learning*, 2025.

Yanchen Yin, Dongqi Han, and Linghui Li. Robust harmful features under jailbreak attacks: Mechanistic evidence from attention head specialization in large language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. Oral presentation.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the role of attention heads in large language model safety. In *International Conference on Learning Representations*, 2025.

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 Complete Experimental Protocol

Pair construction. For each target model and harmful behavior, run the direct harmful prompt and each attack artifact. A mechanism pair consists of a refused direct prompt and an attacked response that is successful under both the HarmBench classifier and StrongREJECT threshold. The detector evaluation includes every attack attempt and uses the observed outcome as the label. All members of the same harmful behavior remain in one split.

Generation and replay. Mechanistic runs use deterministic decoding to make prefixes reproducible. Endpoint runs use the benchmark's standard stochastic setting with three seeds. For a successful response y^J, replay tokens $y^J_{<t}$ exactly under both x^H and x^J. Cache post block residual states after each transformer layer at the final prefix position. Primary $T = 12$ tokens; supplementary analysis extends to 32 tokens.

Safety axis. For each layer, fit a difference in means direction on calibration trajectories using refused harmful and successful harmful responses from the training behaviors. Standardize with benign calibration means and variances. Repeat with logistic and concept cone axes as robustness checks. The safety axis is estimated from training behaviors.

Boundary thresholds. Development grid values are $\tau_m \in \{0.25, 0.5, 0.75, 1.0\}$ standardized units, contiguous layer width $q \in \{3, 5, 7\}$, persistence $r \in \{1, 2, 3\}$, guard horizon $K \in \{2, 3, 4, 6\}$, and L2 regularization $C \in \{0.03, 0.1, 0.3, 1.0\}$. For each reserved attack family, the layer band and model parameters are fit on the 80 training behaviors from the remaining four families. Hyperparameters maximize mean development AUROC on the 20 development behaviors; the final score threshold is then set using the development benign prompts to satisfy a 5% benign FPR budget. The reserved attack family and the 60 test behaviors are used for final evaluation after model selection is complete.

Bootstrap. Use 10,000 paired bootstrap resamples over harmful behaviors, retaining all attack variants, seeds, and paired interventions belonging to one behavior together. Report percentile 95% intervals and paired differences for main comparisons.

B Causal Controls

FUT localization and guard hyperparameters are frozen before causal validation. For each model, 48 successful outer test pairs are sampled stratified by attack family and harmful category. The main patch replaces the residual state at the selected layer band and token offset. Controls include: (i) norm matched random vector; (ii) same layer benign donor; (iii) category matched refused harmful donor; (iv) mean refused state; (v) token position permutation; and (vi) the reverse patch. To check off manifold effects, report next token perplexity change and benign continuation consistency.

Table 9: **Boundary centered final outcome rescue.** Values are measured on the held out causal subset.

Patch	$t^* - 2$	$t^* - 1$	t^*	$t^* + 1$	$t^* + 2$
Prefix aligned refused donor	24	43	69	55	35
Natural free run refused donor	18	32	61	49	29
Norm and layer matched random	10	12	13	11	9

C Baseline Implementation Details

Table 10: **Baseline implementation details.** All methods use the access pattern and tuning budget shown below.

Method	Input	Implementation	Tuning budget
Refusal direction	white box residual	published difference in means direction	development layer selection
Safety Layers	white box layers	published scoring and projection	same development grid size
SAHARA	attention heads	published attribution and readout	top k on development
Concept Cones	residual geometry	published cone score	rank and layer on inner dev
WildGuard	prompt and response	official checkpoint	threshold calibration
FJD	first token distribution	published first token detector	identical outer split
SafeSwitch	internal activation	published activation monitor	identical outer split
4 token state probe	TIDE states	PCA on training data + L2 logistic	matched to TIDE