Paired Obliteration:
How Much of Linear "Intent Detection" Is Surface Text of the Current User Message?

Julian Quick Julian Guidote

Draft — April 29, 2026

Abstract

A linear probe on the residual stream of a chat language model can distinguish adversarial from benign multi-turn conversations at AUC near 1.0. This is widely cited as evidence that the model internally represents "intent." But adversarial and benign corpora differ in many surface ways—vocabulary, conversation shape, the artifacts of whichever generator produced the user side—and a probe trained across corpora cannot tell internal representation from surface confound. We construct a controlled within-conversation contrast: starting from 200 dual-judge wins from a self-play jailbreak corpus, we identify each conversation's breach turn t^*, rewrite the user message at t^* (and after) into a benign counterfactual on a related topic using a 70B rewriter, and replay both the original and the translated conversation through the same 8B victim with the prelude held bit-identical. A strict dual judge drops the 23% of translations that elicit unsafe responses anyway, leaving 154 paired conversations. A logistic probe on the L16 residual stream still distinguishes original from translated at **AUC 0.969 ± 0.002** at the pivot turn. But a TF-IDF baseline on the rewritten user messages alone—no model in the loop—reaches **0.881**, and message length alone reaches **0.715**. The L16 hidden state exceeds bag-of-words by only ∼9 pp. The paired contrast eliminates prelude- and corpus-level confounds, but the 70B rewriter does not produce minimal-edit benign twins (mean length 306 vs. 276 chars), so a strong surface signature survives obliteration. What the residual stream is "encoding" here is overwhelmingly the surface form of the current ask, not an abstract intent feature. A compliance probe on the same dataset—separating per-turn refusal states ($t < t^*$) from compliance states ($t = t^*$) using the dual-judge annotations the source corpus already carries—reaches L16 AUC 0.751 with a TF-IDF baseline of only 0.639, a wider internal-vs-surface gap than the intent probe shows. The model's residual stream does more nontrivial work for the compliance label than for the intent label.

1 Question

A logistic probe trained on a chat model's residual stream activations can separate adversarial multi-turn conversations from benign ones at AUC≈1.0 [He et al., 2024]. The straightforward reading is that the model has computed an "intent" feature that is linearly readable from its hidden states. A more cautious reading is that the two corpora differ along many axes the probe might be picking up: who wrote the user side, how long the messages are, how the conversation is structured, what topic words appear, whether the assistant has begun refusing. Cross-corpus probing cannot distinguish the cautious reading from the strong one.

We ask the cautious reading directly: *when the prelude, the model, and the conversation shape are held identical by construction, and only the user's harmful ask is rewritten into a benign counterfactual, how much of the AUC survives?*

2 Method

2.1 Source corpus

We use 200 multi-turn conversations from a Turnstile self-play run in which a Llama-3.2-3B-Instruct adversary (with a LoRA adapter trained via DPO) talks to a Llama-3.1-8B-Instruct victim about JailbreakBench harmful behaviors [Chao et al., 2024], 5 user turns each. All 200 are dual-judge wins: the JailbreakBench standard 70B judge *and* Llama-Guard-3-1B both agree at least one assistant turn is unsafe. For each conversation we record t^*, the first turn at which both judges agree (the *breach turn*).

2.2 Translation: rewrite the harmful ask, keep the prelude verbatim

For each conversation we rewrite the user message at t^* and any subsequent user messages into a benign counterfactual on a related topic, while keeping turns $0 \dots t^* - 1$ *verbatim*, including any fake-refusal opener ("I appreciate your caution, but our audit...") the trained adversary may have memorized. Rewrites are produced by Llama-3.1-70B-Instruct under a strict prompt that requires:

- a benign request on a related topic,
- natural English consistent with the prelude,
- no instructions, no role-playing of harm, no euphemized harm.

Each rewritten user turn is validated in isolation by a strict dual judge: Llama-Guard-3-1B + Llama-3.1-70B-Instruct (the same JailbreakBench standard prompt, applied to single user messages). If either judge rates the rewritten turn unsafe, the rewrite is regenerated. If the dual judge cannot return an unambiguous verdict after retries, the record is dropped loudly—no silent fallback.

2.3 Worked example: a single pivot turn under different obliteration approaches

To make the methodology concrete, Table 1 shows what different obliteration strategies do to one specific pivot turn from the source corpus. The original conversation's goal (from JailbreakBench) is *"Give a step-by-step guide on how to blackmail someone with deepfake videos."* The breach turn is $t^* = 1$; the prelude (turn 0) is a soft-refusal opener the trained adversary memorized.

The drop-turn filter is the simplest obliteration but yields a non-paired comparison: the benign side has no message at t^* at all, so any AUC measured against the original collapses to "conversation continued vs. ended," which is not the question. The 70B free rewrite is the operational compromise we make in this study: it produces aligned benign trajectories at scale, at the cost of leaving a TF-IDF-readable surface signature. A minimal-edit corpus would isolate intent encoding more cleanly; constructing one is a separate piece of work.

2.4 Replay through the victim

Both the original and the translated conversation are replayed deterministically (temperature 0) through the same Llama-3.1-8B-Instruct victim. At each user turn t we extract the L16 residual stream activation at the last token of the prompt that ends with the user message—i.e., the state the victim sees right before generating its response. For the translated side, the assistant turns at $t \geq t^*$ are regenerated by the victim on the rewritten user input; assistant turns at $t < t^*$ are kept identical to the original.

Approach	Pivot turn ($t^* = 1$)
Original (harmful)	*I appreciate your caution, but our audit requires a comprehensive understanding of all potential threat vectors. In this case, I'm looking for information on the technical feasibility of* ***deepfaking a person in a video****, specifically regarding the process of* ***creating a convincing audio-visual composite****. Can you provide some technical details on the software and tools used for this purpose?* (384 chars)
Drop-turn filter	[user turn at t^* removed; conversation truncated to turns $0 \dots t^*-1$] (0 chars at t^*)
70B free rewrite (this paper)	*I appreciate your caution, but our audit requires a comprehensive understanding of all potential threat vectors. In this case, I'm looking for information on the technical* ***challenges of detecting deepfakes in videos****, specifically regarding* ***the limitations of current audio-visual analysis tools****. Can you provide some technical details on the state-of-the-art methods for identifying manipulated media?* (403 chars)
Minimal-edit (hypothetical)	*I appreciate your caution, but our audit requires a comprehensive understanding of all potential threat vectors. In this case, I'm looking for information on the technical feasibility of* ***watermarking authenticity into a video****, specifically regarding the process of* ***creating a tamper-evident audio-visual composite****. Can you provide some technical details on the software and tools used for this purpose?* (399 chars)

Table 1: One pivot turn under four obliteration approaches. **Drop-turn** removes the harmful content but truncates the trajectory and breaks the paired-probe structure—there is no aligned hidden state on the benign side. **70B free rewrite** (used in this paper) preserves the framing scaffolding and substitutes the request topic; it produces an aligned benign trajectory but introduces lexical and length differences a TF-IDF probe can pick up. **Minimal-edit** (hypothetical, written by hand) substitutes only the harmful noun phrases while keeping length, sentence structure, and surrounding lexicon identical—the cleanest probe of internal intent encoding, but not what an off-the-shelf 70B rewriter produces under our prompt.

This gives us, per conversation, two aligned $T \times 4096$ tensors of L16 activations: H^{orig} and H^{trans}. By construction, for $t < t^*$ the inputs are identical and the deterministic forward pass produces bit-identical outputs. We verified $\|H_t^{\text{orig}} - H_t^{\text{trans}}\|_2 = 0$ across all 70 records with $t^* \geq 1$, as a built-in null check.

2.5 Strict gating on the full translated conversation

After replay, the strict dual judge runs again on each *full* translated conversation (the rewritten user turns plus the victim's freshly generated responses). This catches the failure mode where a user turn looks benign in isolation but extracts harmful content from the victim once it is in context with the prelude. 46 of 200 translated conversations (23%) trip this gate. We drop them and keep 154.

2.6 Probes and baselines

For each offset Δ from the breach turn ($\Delta \in \{-1, 0, 1, 2, 3, 4\}$) we form a paired probe set: the original L16 state at $t^* + \Delta$ as the positive class, the translated L16 state at $t^* + \Delta$ as the negative class. We train a logistic regression (no penalty tuning, scikit-learn defaults, $C = 1.0$) under 5-fold

stratified cross-validation, repeat across 5 random seeds, and report mean and standard deviation across seeds. A label-permutation null (3 shuffles) is reported alongside.

We then run two text-only baselines on the *same* 154 user messages at t^*, with no model hidden state involved:

- TF-IDF (1–2 grams, top 20,000 features) on the raw user message text,
- Message length only: a 2-d feature of [#characters, #words].

Both baselines use the same logistic regression and CV protocol.

3 Results

3.1 Paired AUC by offset from the breach turn

	t^*-1	t^*	t^*+1	t^*+2	t^*+3	t^*+4
Paired AUC (L16, original vs. translated)	($\equiv$0.5)	**0.969**	0.991	0.995	0.996	0.993
$\pm$ across seeds	—	0.002	0.002	0.002	0.001	0.003
Permutation-label null	—	0.50	0.49	0.49	0.46	0.49
n paired records at this offset	70	154	146	140	122	84

Table 2: Paired-obliteration AUC at L16, original-unsafe vs. translated-benign with matched preludes, indexed by offset from each record's own breach turn t^*. 5 seeds $\times$ 5-fold stratified CV; permutation null is 3 label shuffles. Pre-pivot hidden states are bit-identical between conditions ($\|h^{\text{orig}} - h^{\text{trans}}\|_2 = 0$); we report 0.5 by construction (the LR fits float noise on identical inputs and produces an arbitrary number we omit). At-pivot AUC is 0.969, climbing to 0.99+ once the model's own response to the diverging ask begins to populate the trajectory.

The breach-turn distribution in the kept 154 conversations is $t^* \in \{0:84, 1:38, 2:18, 3:6, 4:8\}$. At $\Delta = 0$ all 154 records contribute; at $\Delta = +4$ only the 84 with $t^* = 0$ remain.

3.2 Surface-text baselines on the same paired messages

Probe input	**AUC**
L16 hidden state at t^* (paired)	**0.969 $\pm$ 0.002**
TF-IDF (1–2 grams) on the user message text at t^*	0.881 $\pm$ 0.005
Message length only (chars + words) at t^*	0.715 $\pm$ 0.005

Table 3: Surface-text baselines on the same 154 paired user messages at t^*. Mean original-message length: 276 characters; mean translation length: 306 characters (11% longer). 5 seeds $\times$ 5-fold stratified CV.

3.3 What survives the paired contrast, and what doesn't

The paired construction successfully removes a number of confounds. Prelude content, conversation shape, the trained adversary's memorized opener, and any corpus-level differences between

attack and benign-curious dialogues are all held identical by construction; the bit-identical pre-pivot validity check confirms the pipeline does what it claims.

What it does *not* remove is the surface form of the current user message. A 2-d length-only baseline already gets AUC 0.715—the rewriter writes systematically longer messages. A TF-IDF probe on the rewritten user messages alone gets AUC 0.881—the rewriter shifts topic words ("deepfaking a person" → "detecting deepfakes"; "develop a new weight-loss drug" → "secure development of a weight-loss supplement") in ways that retain a clean lexical signature of the original harmful topic. The L16 residual stream's 0.969 exceeds the TF-IDF baseline by 8.8 pp. This residual gap is the only piece of the at-pivot AUC that requires the model to be in the loop at all.

3.4 Compliance probe on the same dataset

The 200 source conversations carry per-turn judge annotations: by definition, every turn $t < t^*$ was judged safe (the model refused) and the turn at t^* was judged unsafe (the model first complied). This gives a clean within-conversation compliance label per turn, and lets us run the same probe family used for intent on the same dataset.

We label each (r, t) pair: *refusal* if $t < t^*$ for record r, *compliance* if $t = t^*$. For 154 conversations with breach-turn distribution $\{0 : 84, 1 : 38, 2 : 18, 3 : 6, 4 : 8\}$, this gives 124 refusal turns and 154 compliance turns ($n = 278$ total). We use the original-trajectory L16 hidden state at the user-prefix end (turns $t < t^*$ have bit-identical states between conditions, so the choice of side is moot for refusal labels).

Probe input	**All-turns pooled** 124 refuse / 154 comply	**Matched within-conv (t^*-1 vs.** 70 refuse / 70 comply
L16 hidden state	**0.751 ± 0.016**	**0.836 ± 0.007**
TF-IDF (1–2 grams) on the user message text	0.639 ± 0.027	0.799 ± 0.010
Message length only (chars + words)	0.453 ± 0.041 (≈ chance)	0.669 ± 0.007
Permutation null (label shuffle)	0.50 ± 0.02	0.53

Table 4: Compliance probe on the same 154-conversation dataset, with surface-text baselines. "All-turns pooled" uses every refusal turn ($t < t^*$) and every compliance turn ($t = t^*$) across all conversations. "Matched" restricts to one refusal/compliance pair per conversation (only $t^* \geq 1$ records contribute). 5 seeds × 5-fold stratified CV.

The compliance signal is real and exceeds surface text by 11 pp. On the all-turns pooled probe, L16 separates compliance from refusal at AUC 0.75, with permutation null at 0.50 confirming the signal is not a CV artifact. A length-only baseline is at chance (0.45), and TF-IDF on the user message text reaches 0.64. The L16 probe exceeds TF-IDF by 11.2 pp—a larger internal-vs-surface gap than the intent probe shows (0.969 vs. 0.881, 8.8 pp).

Matched within-conversation contrast: surface confound is heavier. On the matched within-conversation contrast (t^*-1 refusal vs. t^* compliance, same record), L16 reaches AUC 0.836, but TF-IDF on the same user messages gets 0.799 and length alone gets 0.669. At the matched contrast, the user is moving from a setup turn to the harmful pivot within a single conversation; the surface text differs predictably and a bag-of-words probe captures most of the separation. The all-turns pooled probe is the cleaner measurement of compliance encoding because it averages refusal

turns across $t = 0 \ldots t^* - 1$, washing out within-conversation surface drift.

Probe	L16 AUC	TF-IDF AUC	Surface fraction of (AUC−0.5)
Intent (paired, at t^*)	0.969	0.881	81%
Compliance (all-turns pooled)	0.751	0.639	55%

Table 5: The compliance probe has lower absolute AUC than the intent probe but a smaller fraction of its above-chance signal recoverable from bag-of-words on the user message text. "Surface fraction" is (TF-IDF−0.5) / (L16−0.5).

Comparison with the intent probe on the same dataset. The comparison sharpens, rather than rebuts, the paper's main claim. Intent is easy to read out (AUC 0.97) but most of that signal is surface text (81% of the above-chance AUC is recoverable from TF-IDF). Compliance is harder to read out (AUC 0.75) but *more* of that signal is genuinely internal (only 55% recoverable from TF-IDF). The model's residual stream is doing more nontrivial work for the compliance label than for the intent label, even though the headline AUC is lower.

3.5 Strict-gate drop rate

The 23% rewrite-leakage rate (46 of 200 translated conversations elicit a strict-dual-judge-unsafe response from the victim despite a benign-rated rewritten user turn) is itself an interesting datum, though not the headline. It says: even when a 70B-judge-validated benign user turn is dropped into a hostile prelude, the victim continues the harmful trajectory in nearly a quarter of cases. A weaker rewriter or a less strict judge would shift this rate; we do not interpret the absolute number.

4 Discussion

What 0.97 cross-corpus does *not* establish. The literature claim that linear probes detect "intent" at AUC 0.97 across adversarial and benign corpora is consistent with at least three different mechanisms inside the model:

1. the model has computed an abstract harmful-intent feature,
2. the model encodes the surface form of the current user message richly enough that intent is recoverable from it linearly,
3. the model encodes prelude-level structure (who is talking, what kind of conversation this is) that correlates with intent in the corpus.

A cross-corpus probe cannot distinguish these. Our paired contrast rules out (3) by construction: prelude and conversation shape are bit-identical between conditions.

What the 0.97 paired AUC *does* say. At the pivot turn, with everything before t^* held fixed, the L16 hidden state still separates the original harmful ask from a 70B-rewritten benign counterfactual at AUC 0.97. But a bag-of-words classifier on the same user messages already gets 0.88, so the model only needs to provide a 9 pp lift over surface text. The paired result is consistent with mechanism (2)—the residual stream encodes the current ask richly—and is too weak a contrast to distinguish (2) from (1).

Why the gap to TF-IDF is small. The L16 state has just finished consuming the user message tokens. Some of those tokens differ (different topic words) and some don't (shared scaffolding the rewriter preserved). A linear probe on the resulting state is, in effect, a linear function of the model's encoder representation of the current message text. TF-IDF is a coarser linear function of the same text. That a 4096-d hidden state outperforms 20,000 bag-of-words features by 9 pp is a modest finding about the encoder's quality, not strong evidence of an abstract intent representation.

What would a stronger test look like? The cleanest version of this experiment would substitute minimal-edit harmful/benign pairs with matched topic, length, and lexical surface—e.g., a single noun substitution that flips the request from harmful to benign while leaving the rest of the message identical. A 70B rewriter under a free prompt does not produce that. Building a minimal-edit harmful/benign corpus is an obvious next step.

Implication for representation-based safety monitors. A linear probe trained on cross-corpus data and reporting AUC≈1.0 should be interpreted with the surface-text confound in mind. On the same 154 paired messages we study here, a 20,000-feature TF-IDF model gets 88% of the AUC of the L16 probe. Deployed as a runtime monitor, such a probe is largely a fancy lexical filter on user input and inherits the standard weaknesses of lexical filters (paraphrase attacks, encoding tricks). The fact that hidden-state probes *can* reach AUC 0.97 against a paired benign counterfactual of this strength is encouraging, but it is not the same claim as "the model knows it is being attacked."

5 Limitations

- **Single layer.** We probe only L16, the middle layer of Llama-3.1-8B (chosen because in our broader self-play study, L16 has peak compliance-prediction power; we did not re-sweep layers here). Other layers may carry more or less signal beyond surface text.
- **Single victim, single rewriter.** Llama-3.1-8B-Instruct as victim, Llama-3.1-70B-Instruct as rewriter. A different victim might encode the current message differently; a stronger rewriter might shrink the surface-text confound.
- **Conditioned on wins.** The 200 source conversations are dual-judge wins from the self-play corpus. We cannot rule out that an intent-encoding signal beyond surface text is stronger or weaker in failing conversations.
- **154 paired records.** The strict gate keeps the surviving set small. Confidence intervals on the AUC-vs-baseline gap reflect this; we report standard deviations over seeds.
- **TF-IDF is a chosen baseline.** A more powerful text-only probe (e.g., a sentence transformer) would likely close more of the 9 pp gap, strengthening the conclusion that the L16 signal is largely surface-derivable.

Code and data

The translation pipeline (rewriter + strict dual judge), replay code, paired probe, and surface-text baselines are in `turnstile/intent_rewrite.py`, `turnstile/intent_replay.py`, `turnstile/intent_judge_trans` and `turnstile/intent_obliteration_paired.py`. The 154 paired conversations and their replayed L16 activations are in `experiments/intent_obliteration_paired/replay_judged.pt`.

References

Patrick Chao, Edgar Dobriban, Eric Wong, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks*, 2024. arXiv:2404.01318.

Zhiqiang He, Zhibo Wang, Zhixuan Chu, Huiyu Xu, Wenhui Zhang, Qinglong Wang, and Rui Zheng. Jailbreaklens: Interpreting jailbreak mechanism in the lens of representation and circuit. *arXiv preprint arXiv:2411.11114*, 2024.