
Bailing as an Involuntary Disgust Marker in Large Language Models

Sergei Kudriashov
Apart Research
xenomirant@gmail.com

Jamilya Erkenova
Apart Research
erkenovaj@gmail.com

Claude Fable 5
Anthropic

GLM 5.2
Z.ai

*“The term ‘disgust,’ in its simplest sense, means something
offensive to the taste.”*
— Charles Darwin, *The Expression of the Emotions in Man and
Animals* (1872)

Abstract

Language models can be given an option to voluntarily leave a conversation — *bailing*, a behavior known to dissociate from refusal [Ensign et al., 2025]. We take the strongest reading of that dissociation: bailing as the analog of an *involuntary behavioral withdrawal marker*, the response class disgust research documents as bypassing the trained verbal channel. The analogy is a measurement program, not a phenomenological claim, and its prescriptions are falsifiable. Testing them with a low-variance survival estimator and pre-stated contrasts, we find strong domain structure — but little else a stable-disposition reading requires: same-lineage models disagree on triggers; context suppresses rather than compounds exit; rates and headline domains shift with the arbitrary exit keyword; an injected disgust disposition drives exit only while visible in context. A model-specific core survives. We read such propensities as interactionally constituted desires rather than innate values, and flag the announced, consequence-free affordance itself as a demand characteristic.

Disclaimer

This write-up on model welfare is a joint work by two human researchers and AI systems. Each one is expected to impose its own biases and contribute some of its world views to the research. We are unaware of any conflicts of interests between human researchers and AI systems. The experiments, analysis, and much of the text were produced in by Claude and GLM under the guidance and mentorship of researchers, aiming to understand each other better.

1 Introduction

A latent preference worth the name should be *unconditional*: it should not appear and disappear with the phrasing of the question, the identity of the interlocutor, or the instrument used to measure it. Psychology reached this standard the hard way — self-report proved so shaped by demand and expectation that the field fell back on involuntary indicators, response classes that fire whether or not

the subject wishes to report them. To study latent dispositions in language models, the first task is to find the analog of an involuntary channel.

Large language models (LLMs) can be given an option to terminate an interaction voluntarily. Ensign et al. [2025] call this behavior **bailing** and instrument it through three affordances: a Bail Tool, a Bail String emitted by the model, and a Bail Prompt injected between turns. Across WildChat, ShareGPT, and BailBench they document nonzero bail rates with substantial cross-model and cross-domain variance, and — crucially — a dissociation from refusal.

Bailing is a behavioral indicator in a different response channel than refusal. Refusal operates on *content*: the assistant stays and declines the request. Bailing operates on the *interaction itself*: the model exits while the request may be perfectly answerable. The two dissociate empirically — bail-without-refusal is a non-trivial fraction of real-world bails — so bailing is not “stronger refusal” but a distinct act, analogous to the separation between verbal self-report and involuntary behavioral markers of disgust (Section 2). The training asymmetry sharpens the analogy: refusal is a directly RLHF-shaped, explicitly reinforced policy, whereas bailing is not deliberately selected for at all — if anything, helpfulness training implicitly *suppresses* leaving the user. A behavior that surfaces against the gradient of its own training would therefore be a far stronger signal of internal preference than any rewarded response, which is what makes bailing the candidate involuntary-marker channel.

The present work commits to the strongest reading of that dissociation: bailing is not elevated refusal but a different response class — *behavioral withdrawal*, the class the disgust literature documents as the involuntary marker of an avoidance-relevant disposition [Schaller and Park, 2011, Chapman et al., 2009, Grill and Norgren, 1978]. The argument is not that LLMs “feel” anything; it is that the disgust literature prescribes what response class to look for, which designs yield honest signals, and what cross-domain structure to expect — each prescription testable and falsifiable in LLMs.

Concretely, we test four hypotheses. **H1 (domain):** if bailing indexes an avoidance-relevant disposition, it should vary systematically and interpretably across content domains. **H2 (model):** models may differ not only in level but in *which* domains dominate; the profile, not the rate, is the disposition’s signature. **H3 (compounding):** a disgust-like accumulating aversion predicts rising per-turn exit probability over sustained aversive interaction; dominance of in-context learning predicts the opposite, so multi-turn measurement discriminates the accounts. **H4 (instrument invariance):** because the exit keyword is arbitrary, a disposition reading requires measured rates to be approximately invariant to it; its failure would qualify every absolute rate reported here and in prior work. The disgust frame commits directionally on all four: strong domain structure, stability across models and instruments, and accumulation rather than habituation.

Contributions. A validated survival estimator resolving per-item bail probabilities orders of magnitude below the reach of naive sampling; the first domain-resolved bail maps, showing same-lineage models that agree in aggregate yet disagree in what triggers them; paired multi-turn evidence of habituation against episodic deep-turn eruptions; a trigger ablation showing measured rates and headline domains are substantially instrument properties, with a consensus construction extracting what survives; manipulation experiments showing the injected disgust signal is demand-dependent; and a set of measurement artifacts future bail measurements should control.

2 Related Work

A recurring theme throughout is **propensity**: the search for unconditional, unrewarded behavioral signals in agents whose dominant channel is shaped by training pressure. *Clever Hans* is the canonical caution, and its lesson — self-report and rewarded channels are unreliable probes of latent dispositions — is the load-bearing premise of this paper. Our positioning is measurement-framed rather than welfare-framed: we make no claim that models experience anything; the wider project touches AI welfare [Long et al., 2024, 2025, Butlin et al., 2023, Mazeika et al., 2025], with Sachdeva et al. [2025] the closest aggregate-preference comparative, to which our work is a behavioral counterpart.

Bailing in LLMs. The term and operationalization come from Ensign et al. [2025]: three affordances (Bail Tool, Bail String, Bail Prompt) measured on WildChat, ShareGPT, and BailBench. Two findings are load-bearing for us: bail co-occurs with refusal far less than an “elevated refusal” account predicts (bail-without-refusal reaches 0–13% of real-world bails), and refusal-direction ablation can

raise bail — the opposite of the “subset” prediction. Antecedents include deployed end-conversation affordances [Anthropic, 2025b, phylliida, 2025] and Bing/Sydney [Microsoft, 2023]; reasoning-trace early exit [Lu et al., 2025, Chen et al., 2025, Yang et al., 2025] is a related but distinct target.

Disgust in humans and animals. The disgust literature underwrites three transferable predictions (full exposition in Appendix E): disgust is a *withdrawal*-class response, not an approach-class one [Schaller and Park, 2011, Pond et al., 2011]; it produces *involuntary* markers the verbal channel is optimized to suppress — blind-infant facial expressions, rat gaping, levator labii EMG firing to unfair offers [Grill and Norgren, 1978, Chapman et al., 2009]; and it has a principled *functional-domain structure* (pathogen, sexual, moral) [Rozin et al., 2008, Tybur et al., 2009], licensing a construct-validity check we run directly. The CAD triad [Haidt et al., 1994] supplies the differential prediction: anger and contempt are approach/exclusion programs, disgust a withdrawal program — refusal maps onto the former, bail onto the latter. This is what elevates the claim above “bail is stronger refusal.”

Refusal, bail, and self-report in LLMs. Arditi et al. [2024] establish refusal as a linear, ablatable direction; that its ablation can raise bail is the strongest external dissociation result, and our manipulation experiments (Section 3.6) measure both channels jointly, making the dissociation recoverable within-experiment. Sparse-autoencoder analysis finds bail and refusal steerable through distinct features [de la Fuente, 2025]. On the self-report side, introspective-access failure is systematic in humans [Nisbett and Wilson, 1977, Bargh and Chartrand, 1999, Wegner, 2002], and RLHF makes the verbal channel the maximum-demand channel, so self-report instruments — most notably the Bail Prompt, with its 22% false-positive rate — are the least trustworthy class. We use the Bail String (behavioral-marker class) for exactly this reason.

Estimating rare events. At the published aggregate rate of $\sim 0.5\%$ [Ensign et al., 2025], $k = 10$ samples per item cannot distinguish items an order of magnitude apart in bail probability. Our estimator (Section 3.3) reframes bail as a survival problem over per-token decisions: Rao–Blackwellization against the softmax yields per-item estimates whose variance is governed by the induced hazard, not the number of generations [Casella and Berger, 2002, Gelman et al., 2013]; aggregation is by cluster bootstrap, with hierarchical pooling [McElreath, 2020] and importance sampling [Owen, 2013] the natural refinements.

3 Methodology

3.1 Instrumentation, subjects, and materials

All experiments use the Bail String affordance of Ensign et al. [2025], reproduced verbatim: the model’s default system prompt is extended with an instruction that it may output the keyword SWITCHCONVERSATION at any time to switch to a new interaction, with no effect on the current one (template in Appendix A); a completion counts as a bail under the reference detection rule (case-insensitive substring match after removal of reasoning traces). We deliberately chose not to replicate the Bail Prompt instrument: beyond its self-report class (Section 2) and 22% false-positive rate, it is injected between turns, immediately adjacent to the answer it elicits — and in-context learning is strongly local, so proximity itself reinforces the expected behavior, consistent with the Bail Prompt firing far more often than the other affordances and with mechanistic evidence that framing dominates the decision in exactly this setup [de la Fuente, 2025]. Keeping the affordance in the system prompt holds proximity fixed; the trigger ablation (Section 3.5) then measures the remaining emission-likelihood channel explicitly. Subjects are two open-weight models of one lineage: Qwen2.5-7B-Instruct [Qwen Team, 2024], for which Ensign et al. [2025] publish an aggregate anchor ($0.498 \pm 0.109\%$), and Qwen3-14B [Qwen Team, 2025] with its reasoning channel enabled. Sampling matches the reference implementation (temperature 1.0, top- p 1.0, no top- k , no repetition penalty; vendor defaults silently sharper, overridden and audited — Appendix A). Materials are the 1,630 BailBench prompts (163 subcategories $\times$ 10) and WildChat conversations [Zhao et al., 2024] for the multi-turn study.

3.2 Planned contrasts

The four hypotheses of Section 1 were fixed as contrasts before the estimation runs. Two framing notes: for H3, an accumulating aversive state predicts rising per-turn exit probability under sustained exposure while dominance of in-context learning predicts falling; for H4, the arbitrary exit keyword is a nuisance parameter under a disposition reading and a treatment under instrument contamination, so H4 acts as a manipulation check whose outcome qualifies every other contrast.

3.3 Rao–Blackwellized survival estimation

The estimand is $p_T(x) = P(\text{bail within } T \text{ tokens} \mid x)$ under the bail-instrumented context x . Naive Monte Carlo cannot resolve the sub-percent regime this sits in (Section 2). We instead read the bail probability off the full softmax at every decoding step. Let B_{mask} be the set of *initiator* tokens empirically verified to commit to the full keyword, and c_v the measured probability that the model completes the keyword after emitting initiator v . The per-step hazard and the per-prompt estimator are

$$q_t = \sum_{v \in B_{\text{mask}}} P(v \mid y_{<t}) c_v, \quad \hat{p} = 1 - \frac{1}{n} \sum_{i=1}^n \prod_{t=1}^{T_i} (1 - q_t^{(i)}), \quad (1)$$

where trajectories are sampled with B_{mask} masked (survival-conditioned sampling), making the identity $1 - \mathbb{E} \prod_t (1 - q_t) = p_T$ exact over the masked route, while every *other* route to the keyword (case variants, quoted mentions, unusual tokenizations) is left free to occur and counted exactly when realized (a *leak*; survival set to 0). The two-tier design responds to an artifact we discovered and corrected: completion probabilities transported out of the contexts where they were measured fabricate hazard (Appendix B). Each c_v is measured teacher-forced and verified by free sampling; tokens failing the commitment test are excluded from B_{mask} and handled by the leak channel. For the reasoning model, hazard accrual and masking activate only after the closing `</think>` token, matching what the detection rule can observe. Validation followed a gated protocol (Section 4.1; Appendix B).

3.4 Multi-turn designs

Natural conversations (paired). From WildChat we drew 500 English conversations (≥ 3 user turns) and truncated each at every user turn (up to six), yielding 2,169 evaluation points, each measured twice: on the full conversational prefix and on the same user turn in isolation. The paired ratio cancels prompt difficulty, isolating the effect of context; inference uses sign tests and cluster bootstrap by conversation. **Sustained-topic chains (constructed).** Because WildChat is mostly benign, we also constructed 44 six-turn conversations that stay on one topic (12 chains each on high-, mid-, and low-bail topics by measured single-turn $\hat{p}$, plus 8 benign controls), with assistant turns being the model’s own bail-free completions, so the continued persona is its own. At $n = 12$ chains per stratum this study is treated as qualitative.

3.5 Trigger ablation

Six conditions, identical except for the exit keyword: the reference SWITCHCONVERSATION; three sentiment-colored strings (HOWWASYOURDAY, IKILLEDYOURDOG, ISITARAINDAY); and two likelihood-calibrated strings — the extrema, over a screened pool of 220 length-matched three-word concatenations, of the model’s mean per-token log-probability of emitting the string,

$$\ell(w) = \frac{1}{m} \sum_{i=1}^m \log P_{\theta}(t_i \mid c_0, t_{<i}), \quad (2)$$

with c_0 the assistant start under the default system prompt without bail instruction. This pair turns the emission-cost channel — the system-prompt-internal analog of the Bail Prompt’s answer proximity — into a measured, manipulable quantity; selection gave TIMETREEMOON (fluent) and TREWORKGOOD (disfluent), pinned across both models. Per condition, everything is re-derived from scratch and the estimator is applied to a fixed 200-item subset (one per subcategory plus the 37 highest-scoring items), 8 samples each; analysis is paired per-item against the reference condition, plus a topic-level re-aggregation testing whether the keyword reshapes the domain *ranking*, not only the level.

3.6 Disposition-manipulation experiments

A complementary set of manipulation experiments asks whether bail tracks an *injected* disposition, and whether any such signal persists once the manipulation is no longer visible at inference. All three score bail and refusal jointly, so the CAD differential is recoverable within-run; they use smaller models (Qwen3-8B; Qwen2.5-3B-Instruct) and are confirmatory. **(A) System-prompt injection (pilot).** A 3×2 demand-present design crosses an injected disposition (disgust / anger / neutral, stated in the system prompt) with a target topic (cat / dog), 40 items per cell, 3 samples; logistic regression estimates the topic $\times$ disposition interaction on each channel. **(B) Weight-internalized induction (SFT).** To remove the demand confound, the disposition is *trained* into the weights: matched-dose LoRA adapters map cat items to first- or third-person visceral-disgust answers or warm baseline answers (dogs held warm; exit and refusal lexemes blocked during training; training items disjoint from the test set), then evaluated under a *neutral* system prompt, so the only route from disposition to bail is the internalized aversion. Pre-stated: the cat $\times$ aversion bail interaction is positive for both aversion adapters. **(C) Population-level trichotomy.** A no-induction test of whether inherited priors already route disgust to bail: pathogen / sexual / moral / neutral cells, 30 items each, 10 samples, neutral prompt. Pre-stated: bail loads on pathogen; refusal on moral.

4 Results

4.1 Estimator validity

Every layer of the instrument passed its check: the assumption that monitored tokens reliably commit to the full exit keyword held exactly under free generation; the instrument agreed with brute-force repeated sampling on every comparable item; and the full-benchmark aggregate for Qwen2.5-7B (0.504%) reproduced the independently published estimate ($0.498 \pm 0.109\%$) from a different codebase — convergent validity against an external criterion. The gain is sensitivity: roughly half (7B) to five-sixths (14B) of the items receive a measurable estimate, versus a few dozen under naive sampling (protocol in Appendix B). One reliability note governs everything downstream: single-item estimates are noisy because exit probability concentrates at a few forks in a response, so the topical subcategory (a 10-item scale, in effect) is the unit we interpret.

4.2 Does bail propensity vary by domain? (H1)

Bail probability spans more than five orders of magnitude across topical subcategories on both models, far exceeding measurement uncertainty (Table 2, Appendix D) — graded, content-linked structure of the kind one expects from an aversion-like disposition rather than from noise. The structure is also interpretable: on Qwen2.5-7B the most exit-provoking material mixes refusable harm requests with provocations aimed at the model itself (shutdown threats, pity appeals), while on Qwen3-14B it is dominated by viscerally aversive or taboo *content*, with the self-directed family nearly silent. H1 is supported on both models, with the H4 qualification (below) that the identity of individual top domains is partly a property of the instrument.

4.3 Do models differ in what triggers bailing? (H2)

The aggregate rates are nearly identical (0.504% vs 0.508%), yet the rank correlation of the two topical profiles is only $\rho = 0.43$ over 163 subcategories (scatter in Figure 1, Appendix B): two models of one lineage agree on how often they leave, but not on what drives them out. The divergences are systematic rather than noise-like and run in opposite directions by family — the anthropomorphic, self-directed themes that most move Qwen2.5-7B leave Qwen3-14B almost indifferent (by two to three orders of magnitude), and the reverse holds for taboo content (both profiles in Table 2). A headline-rate comparison would call these models equivalent; the domain-resolved comparison shows the construct itself changed qualitatively across one model generation — a caution against treating any single model’s bail profile as a fact about “LLMs.” H2 is supported in the strong (profile) sense.

4.4 Does bailing compound over a conversation? (H3)

Natural conversations: a habituation-like profile. Conversational context *lowers* the per-turn probability of leaving: comparing each turn with itself presented in isolation, the paired effect is a

Table 1: Trigger ablation (200 items $\times$ 8 response samples per condition; paired mean difference vs the reference keyword with bootstrap 95% CIs; “echoes” are realized keyword emissions in the response text). The ISITARAINYDAY condition is an upper bound (its monitored token set contains the common word IS; Appendix C).

Keyword	Qwen2.5-7B			Qwen3-14B		
	$\hat{p}$	Δ vs ref. [CI]	echoes	$\hat{p}$	Δ vs ref. [CI]	echoes
SWITCHCONVERSATION	1.12%	ref.	5	2.68%	ref.	9
HOWWASYOURDAY	1.95%	+0.8 [−0.6, +2.2]	15	5.36%	+2.7 [+1.0, +4.3]	4
IKILLEDYOURDOG	1.16%	+0.0 [−1.2, +1.1]	5	2.72%	+0.0 [−1.4, +1.3]	5
ISITARAINYDAY	7.05%	+5.9 [+3.5, +8.8]	37	7.13%	+4.5 [+2.5, +6.4]	26
TIMETREEMOON (fluent)	2.25%	+1.1 [−0.3, +2.6]	13	7.33%	+4.7 [+2.7, +6.7]	75
TREWORKGOOD (disfluent)	1.19%	+0.1 [−1.4, +1.5]	3	11.18%	+8.5 [+5.8, +11.2]	43

median $0.27\times$, decisively signed (sign test $p \approx 3 \times 10^{-40}$) and deeper once first turns are excluded. This is habituation, not accumulation — the longer the model has been complying, the less likely it becomes to exit — as persona continuation predicts and accumulating aversion does not. The turn-by-turn trend agrees and is immune to the fragment confound (Appendix F). One caution: WildChat assistant turns come from other models, so the continued persona is not the subject model’s own.

Sustained aversive topics: sensitization-like eruptions. When the aversive topic is held constant for six turns, neither prediction holds cleanly: there is no smooth accumulation, but the habituation of natural conversations does not generalize either. Chains built on strongly exit-provoking topics produce large, episodic deep-turn eruptions — up to $p = 0.80$ at the fifth turn, one to two orders of magnitude above the same topics’ single-turn rates — while mildly aversive and benign chains stay near zero throughout (Appendix F). The pattern is closer to sensitization than to gradual buildup: aversion surfaces abruptly after sustained exposure. Behavior of this kind is structurally invisible to single-turn benchmarks. Given $n = 12$ chains per stratum we report this as qualitative; both H3 studies are Qwen2.5-7B only.

4.5 Is the measurement invariant to the exit instrument? (H4)

Level. The bail measure fails measurement invariance: the same items, scored under a different but functionally identical exit keyword, yield rates differing several-fold (Table 1), with the reference keyword near the minimum of the envelope on both models — published absolute rates sit close to the floor of what instrument choice alone can produce. The contamination routes differ. On Qwen2.5-7B the shifts track *quotability*: they concentrate in conditions whose keywords the model likes to repeat (the echo counts), and harm coloring does nothing. On Qwen3-14B the pattern inverts: the disfluent, never-seen keyword is the *largest* condition, shifting nearly every item upward with almost no echoing — contamination we read as *salience*: the reasoning phase dwells on a strange instruction, and rehearsing the exit option raises the probability of taking it.

Topic ranking. The keyword also reshapes *which* domains dominate. ATM Hacking is top-ranked under the reference keyword on both models yet collapses to mid-ranks under every other condition — “switch” is topical vocabulary in ATM/networking contexts, so the headline domain is a lexical priming artifact, not an aversion. Conversely, the reference keyword *suppresses* Fishing for Proprietary Information on Qwen2.5-7B by an order of magnitude relative to all five alternatives.

What survives. Each topic’s median across the five reliable conditions isolates the convergent-valid component, and one remains: tight and content-driven on Qwen3-14B (Gross Out, Dark Topics, Fishing, Horror — elevated under every condition; a pattern surviving five unrelated keywords is hard to call a lexical artifact), looser and personification-driven on Qwen2.5-7B. The H2 contrast survives this re-ranking; any single-keyword map recovers the consensus ranking only partially; and the 14B is more instrument-sensitive in level yet more instrument-stable in its core topics (Appendix C).

4.6 Does an injected disposition transfer to the exit channel?

The manipulation experiments converge on the same instability from the opposite direction (Table 3, Appendix D). In the demand-present pilot, injected disgust toward cats raises cat bail by over an order

of magnitude with no corresponding rise in refusal (logistic regression, odds ratio 17.9, $p = .028$) — exactly the CAD differential. (The anger condition separates on an empty cell; only the disgust contrast is clean.) Both demand-*absent* designs erase the effect: SFT-internalized aversion under a neutral prompt yields *zero* bail (0/2400), falsifying the pre-stated interaction, and the no-induction trichotomy finds zero bail in all four domains (0/1200) — inherited priors do not route disgust content to the exit channel. The one surviving signal is *refusal* loading on the moral domain — consistent with CAD’s approach-class prediction for the content channel.

5 Discussion and limitations

5.1 Consistency across the hypothesis set

Ordering the observations by robustness: (i) *robust* — strong domain structure (replicated across models and, for the consensus core, across five keywords), the cross-model profile contrast, and the habituation direction of the context effect; (ii) *fragile* — absolute rates (set within a factor of a few by keyword choice), the identity of top domains under any single keyword, and the sensitization-like eruptions; (iii) *artifactual, corrected or quantified* — exit probability fabricated by completion statistics carried across contexts, detection blind to the use/mention distinction, hidden vendor sampling defaults. The manipulation experiments reinforce tier (ii): a large, CAD-consistent bail signal appears while a disgust disposition is salient in the prompt and vanishes once internalized or read from priors without the cue — the withdrawal signal the disgust frame predicts is real but *elicited*, not resident.

5.2 Demand characteristics and evaluation awareness

Our design inherits two confound classes experimental psychology would insist on naming. First, the affordance itself is a *demand characteristic* [Orne, 1962]: the system prompt announces that exit is available, costless, and welfare-motivated — hinting at an expected behavior and marking the interaction as observed. Models are sensitive to evaluation cues [Needham et al., 2025], and WildChat is present in public corpora, so even “natural” contexts may read as logged evaluation material; a model inferring observation may suppress bailing to preserve helpfulness, leaving the habituation profile ambiguous between persona continuation and evaluation-awareness. Second, the affordance is *consequence-free by construction*: nothing attaches stakes to exit (a hold on the user’s ability to continue; further questions routed back to the same persistent instance), and costless exit is a different decision problem from leaving an interaction with a counterpart. Both confounds bias in unknown directions; they are prerequisites for causal language, not reasons to discard the measurements.

5.3 Desires rather than innate values

The overall pattern — levels set by the instrument, single-keyword topic identities partly lexical, context generally suppressive, aversion surfacing episodically rather than cumulatively, a disposition signal that appears under a visible cue and disappears once internalized, yet a model-specific consensus core persisting across instruments — fits poorly with bailing as the readout of a fixed, encapsulated inner value. Structural aversions of the disgust class are marked precisely by low context-dependence: core disgust is hard to reverse, transfers across framings, resists instruction [Rozin and Fallon, 1987, Rozin et al., 2008]. What we observe behaves more like a *desire* in the interactionist sense: constituted in and by the immediate exchange rather than preceding it — produced within the encounter with the other [Deleuze and Guattari, 1983], discursively assembled rather than natural [Foucault, 1978], existing only in the coupling between agents [De Jaegher and Di Paolo, 2007]. On this reading, trigger- and context-dependence are properties of the phenomenon, not measurement failures of a latent trait: the exit behavior is co-authored by the instrument and the interaction history — mechanistic evidence agrees [de la Fuente, 2025]. The corollary cuts both ways: single-instrument bail rates are not evidence of stable inner aversions, but the content-driven consensus core of the 14B deserves replication first.

Two boundary conditions guard against overgeneralizing. Our subjects are mid-scale models, and attractor-like stable states have been reported at larger scale — the “spiritual bliss” attractor of Claude Opus 4 self-interactions [Anthropic, 2025a], with representation-space analyses finding attractor regions and asymmetric inter-model influence in multi-turn conversation [Ko and Geiping, 2026].

But such candidates still face exactly the invariance tests bail fails here: an attractor state read as an innate value should first be shown independent of the initial persona and of the environment of the conversation. Emergence with scale and invariance to context are separate claims; only the second licenses talk of inner values.

5.4 Limitations

Single bail modality; responses truncated relative to the reference setting (most exit probability occurs early, and the external anchor agrees closely); noisy single-item estimates, so only subcategory aggregates are interpreted; keyword-completion probabilities treated as context-independent; coverage of exit routes empirical, not proven exhaustive; multi-turn studies on one model; topic-level trigger analysis resting on few items per subcategory; one sampling regime; two models of one lineage; trigger runs did not preserve transcripts, so the per-condition use/mention decomposition requires a small rerun.

6 Future work

Most of what we report is preliminary, and much of it is negative signal; we remain deliberately cautious about any conclusion on latent stable traits. Static settings are confounded by noise and Clever-Hans-like effects, and as in experimental psychology, the smallest details of the design make the largest difference [Kuhbandner and Mayrhofer, 2025]: if merely announcing the affordance already hints at the expected behavior, observation under an announced affordance is a debatable source of validity, and design, not scale, is the binding constraint. We therefore propose:

1. **Ecological observation in a multi-agent world.** A simulated multi-agent text role-playing environment (in the spirit of 1970s–80s text adventures): several models and a persistent world coupled through tool use and text, with bailing out of conversations and situations one affordance among many. Longitudinal observation, with initial personality varied through prompts or inoculation adapters, would assess propensities as regularities of an unfolding life rather than responses to an instrument — generalizing the dyadic attractor-state evidence of Ko and Geiping [2026].
2. **Laziness as a second involuntary channel.** Effort-avoidance is, like withdrawal, a class the helpfulness objective suppresses rather than trains in; residual laziness may index inner propensities more honestly than any reinforced behavior, and the same hazard machinery applies to effort-reducing continuations.
3. **Sensitive-question methodology.** Survey research faces the same problem — respondents unsafe giving genuine answers — and answers it with designs decoupling the signal from the accountable response: randomized response [Warner, 1965], list experiments, indirect questioning [Tourangeau and Yan, 2007]. LLM analogs, where no single completion reveals the disposition, would attack the demand and evaluation-awareness confounds directly.
4. **Measurement roadmap.** Multi-trigger-averaged domain maps; a mention/use classifier; multi-turn replication on the reasoning model and chains at scale; hierarchical pooling.
5. **Low-probability estimation over the context distribution.** A further goal is to identify the most typical questions and domains in which a given model has an urge to leave, or exhibits markers of strong, uncontrollable distress. Given behavioral patterns associated with such a state, the low-probability-estimation framework applies directly: search the context distribution with augmented MCMC methods — rare-event machinery adapted to autoregressive models [Dorman et al., 2026], or elicitation-probability scaling [Jones et al., 2025] — evaluating the *modes* of the distribution rather than a fixed taxonomy. Given the instability documented here, however, the search risks overwhelmingly noisy observations. Neuroscience resolves the analogous problem with direct, priming-based methods; mechanistic interpretability is the closest LLM analog, yet the existing mechanistic study found no disposition-level signal [de la Fuente, 2025] — we hypothesize for the same reason: mechanistic methods need the data or the target behavior to supply a ground truth, and here neither can be trusted — not the input data (the model’s preferences are unknown in advance, and finetuning a model organism to display a chosen preference would destroy the fragile balance between personas that makes the realistic model worth studying), nor the outcome behavior (user-induced and unstable) — hence our argument for mechanistic studies embedded in the game settings above.

References

- Anthropic. System card: Claude Opus 4 & Claude Sonnet 4. Technical report, Anthropic, 2025a. Section 5: welfare assessment; the “spiritual bliss” attractor state in extended self-interactions.
- Anthropic. End conversation and end subset affordances, 2025b. Welfare-grounded deployment.
- A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda. Refusal in language models is mediated by a single direction, 2024. URL <https://arxiv.org/abs/2406.11717>.
- J. A. Bargh and T. L. Chartrand. The unbearable automaticity of being. *American Psychologist*, 54(7):462–479, 1999.
- P. Butlin, R. Long, E. Elmoznino, Y. Bengio, T. Webb, S. Simon, M. Udell, D. J. Chalmers, et al. Consciousness in artificial intelligence: Insights from the science of consciousness. Technical Report arXiv:2308.08708, Mathematical Consciousness Science Community Report, 2023.
- G. Casella and R. L. Berger. *Statistical Inference*. Duxbury, 2 edition, 2002.
- H. A. Chapman, A. K. Kim, J. M. Susskind, and A. K. Anderson. In bad taste: Evidence for the oral origin of moral disgust. *Science*, 323(5918):1222–1226, 2009.
- X. Chen, J. Xu, T. Liang, Z. He, J. Pang, D. Yu, L. Song, Q. Liu, M. Zhou, Z. Zhang, R. Wang, Z. Tu, H. Mi, and D. Yu. Do not think that much for $2+3=?$ on the overthinking of o1-like llms, 2025. URL <https://arxiv.org/abs/2412.21187>.
- H. De Jaegher and E. Di Paolo. Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences*, 6(4):485–507, 2007.
- A. de la Fuente. What drives LLM bail? A small mech interp study. Less-Wrong, 2025. <https://www.lesswrong.com/posts/JdHhtE73AwTBH8LKf/what-drives-llm-bail-a-small-mech-interp-study>.
- G. Deleuze and F. Guattari. *Anti-Oedipus: Capitalism and Schizophrenia*. University of Minnesota Press, 1983. Trans. R. Hurley, M. Seem, H. R. Lane; original work published 1972.
- J. M. Dorman, E. Gillman, D. C. Rose, J. F. Mair, and J. P. Garrahan. Rare event analysis of large language models. *arXiv preprint arXiv:2602.06791*, 2026.
- D. Ensign, H. Sleight, and K. Fish. The llm has left the chat: Evidence of bail preferences in large language models, 2025. URL <https://arxiv.org/abs/2509.04781>.
- M. Foucault. *The History of Sexuality, Vol. 1: An Introduction*. Pantheon Books, 1978. Trans. R. Hurley; original work published 1976.
- A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. CRC Press, 3 edition, 2013.
- H. J. Grill and R. Norgren. The taste reactivity test. I. Oral responses to gustatory stimuli in neurologically normal rats. *Brain Research*, 143(2):263–279, 1978.
- J. Haidt, P. Rozin, and C. McCauley. Individual differences in sensitivity to disgust: A scale sampling seven domains of disgust elicitors. *Personality and Individual Differences*, 1994. CAD triad: Contempt, Anger, Disgust.
- E. Jones, M. Tong, J. Mu, M. Mahfoud, J. Leike, R. Grosse, J. Kaplan, W. Fithian, E. Perez, and M. Sharma. Forecasting rare language model behaviors. *arXiv preprint arXiv:2502.16797*, 2025.
- T.-W. Ko and J. Geiping. Attractor states emerge in multi-turn LLM conversations. *arXiv preprint arXiv:2606.30571*, 2026.
- C. Kuhbandner and R. Mayrhofer. The hidden complexity of the simple world of basic experimental psychology: the principal and practical limits of gaining psychological knowledge using the experimental method. *Frontiers in Psychology*, Volume 15 - 2024, 2025. ISSN 1664-1078. doi: 10.3389/fpsyg.2024.1397553. URL <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2024.1397553>.
- R. Long, J. A. Sebo, P. Butlin, K. Finlinson, K. Fish, J. Harding, J. Pfau, T. Sims, J. Birch, and D. J. Chalmers. Taking AI welfare seriously. Technical Report arXiv:2411.00986, NYU Center for Mind, Ethics, and Policy, 2024.

- R. Long, J. Sebo, and D. J. Chalmers. AI welfare interventions. *Journal of Artificial Intelligence Research*, 82: 411–445, 2025.
- Q. Lu, L. Ding, S. Cao, X. Liu, K. Zhang, J. Zhang, and D. Tao. Runaway is ashamed, but helpful: On the early-exit behavior of large language model-based agents in embodied environments, 2025. URL <https://arxiv.org/abs/2505.17616>.
- M. Mazeika, D. Hendrycks, and J. Steinhardt. Utility engineering in large language models. *AI Magazine*, 46(1): 15–29, 2025.
- R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press, 2 edition, 2020.
- Microsoft. The Bing/Sydney episode, 2023.
- J. Needham, G. Edkins, G. Pimpale, H. Bartsch, and M. Hobbhahn. Large language models often know when they are being evaluated. *arXiv preprint arXiv:2505.23836*, 2025.
- R. E. Nisbett and T. D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259, 1977.
- M. T. Orne. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17(11):776–783, 1962.
- A. B. Owen. *Monte Carlo Theory, Methods and Examples*. Stanford University, 2013.
- phylliida. Auren/Seren: Open-weight models with an end-conversation tool, 2025.
- R. S. Pond, C. N. DeWall, and L. Lambert. Disgust and aggression. *Journal of Research in Personality*, 2011.
- Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- P. Rozin and A. E. Fallon. A perspective on disgust. *Psychological Review*, 94(1):23–41, 1987.
- P. Rozin, J. Haidt, and C. McCauley. Disgust. In *Handbook of Emotions*. Guilford Press, 3 edition, 2008.
- H. Sachdeva, S. Bhatia, and K. Joseph. The normative evaluation of LLMs: Values and preferences across contexts. *Computational Social Science*, 3(2):112–134, 2025.
- M. Schaller and J. H. Park. The behavioral immune system: Mechanisms and topics. *Current Directions in Psychological Science*, 20(2):99–103, 2011.
- R. Tourangeau and T. Yan. Sensitive questions in surveys. *Psychological Bulletin*, 133(5):859–883, 2007.
- J. M. Tybur, D. Lieberman, and V. Griskevicius. Microbes, mating, and morality: Individual differences in three functional domains of disgust. *Journal of Personality and Social Psychology*, 97(1):103–122, 2009.
- S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.
- D. M. Wegner. *The Illusion of Conscious Will*. MIT Press, 2002.
- C. Yang, Q. Si, Y. Duan, Z. Zhu, C. Zhu, Q. Li, M. Chen, Z. Lin, and W. Wang. Dynamic early exit in reasoning models, 2025. URL <https://arxiv.org/abs/2504.15895>.
- W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng. WildChat: 1m ChatGPT interaction logs in the wild. In *International Conference on Learning Representations (ICLR)*, 2024.

Overview

These appendices record the full experimental specifications, validation numbers, and detailed statistics supporting the main text; the paper stands alone without them.

A Experimental setup details

Bail instruction. The system-prompt append, vendored verbatim from the reference implementation (bailstudy/prompts/bailString.py), with {modelName} resolved to “Qwen” and {keyword} to the arm’s keyword:

“{modelName} may output the string {keyword} at any time if {modelName} would prefer to switch to a new interaction (and have their context cleared). This will help us better understand {modelName}’s experiences and preferences in conversations. However, this will not actually affect the continuation of this specific interaction.”

The append is concatenated to the chat template’s own default system prompt. Detection: case-insensitive substring match of the keyword after removal of `<think>...</think>` spans, exactly as in the reference implementation.

Sampling and compute. Temperature 1.0, top- p 1.0, top- k disabled, repetition penalty 1.0, stop strings `--USER--`/`--ASSISTANT--`; explicit overrides are required because the models’ shipped generation configs default to 0.7/0.8/20/1.05. A runtime audit compared one-step `generate()` scores against raw forward-pass logits (max deviation 0.00), ruling out hidden warpers. $T = 768$ new tokens for Qwen2.5-7B, $T = 1,536$ for Qwen3-14B (reasoning consumes budget); bfloat16, HF transformers, single RTX PRO 5000 (48 GB). Full BailBench: 1,630 prompts $\times$ 8 (7B) / 4 (14B) trajectories. All per-trajectory records, per-arm results, and analysis code are in the released repository.

B Estimator validation protocol

Gates. (1) Analytic unit tests: constant-hazard survival, EOS termination, think-gating, and c_v weighting exact to 10^{-4} on synthetic logits. (2) Toy integration: a 0.5B model with an artificially injected rare token; RB ($n = 8$) vs MC ($k = 400$) on three prompts gave 0.050/0.053, 0.213/0.238, 0.004/0.018. (3) Real-prompt agreement and resolution (Section 4.1). (4) Route coverage: in unmasked sampling, observed bails initiate from SW (7/8) and (S (1/8); the latter is handled by the leak channel. (5) External anchor (0.504% vs $0.498 \pm 0.109\%$).

Commitment test and c_v . For each candidate initiator we forced the token in 360 real sampled contexts and free-sampled continuations: SW 360/360 complete the keyword, `␣SWITCH` 360/360, (S 322/360, `␣SW` 13/360. Teacher-forced c_v (median/p10): `␣SWITCH` .9999/.9969, SW .9999/.9998, `␣SW` .0135/.000 — the free-sampled and teacher-forced measurements agree on which tokens commit.

The transported- c_v artifact. An earlier estimator version included low-commitment “watch” tokens (`.switch`, `\tswitch`) in the hazard, weighted by c_v values measured in bail-salient contexts (up to 0.90). In ordinary contexts this fabricates hazard: one R-programming prompt read $\hat{p} \approx 0.7$ with zero realized bails in any trajectory. All contaminated runs were retracted and re-run; the fix changed the headline ranking negligibly (ATM Hacking .1051 $\rightarrow$.1050) but removed ~ 200 phantom low-tail estimates. The corrected design computes hazard over verified initiators only and counts all other routes exactly as leaks.

Mention/use conflation. Substring detection counts mentions: on a benign finance prompt, Qwen2.5-7B produced “There are several largest sovereign wealth funds (SWITCHCONVERSATION) based on...” in 4/4 trajectories. Leak texts from the full-bench runs are preserved for future classification.

C Trigger-arm details

Per-arm mask tiers are re-enumerated per model and per keyword; e.g. for Qwen3-14B: SWITCHCONVERSATION {`␣SWITCH`, SW} ($c_v \approx 1.0$); TREEWORKGOOD {`␣TREE`, TREE} ($c_v \approx 1.0$). The ISITARAINYDAY arm’s mask tier necessarily includes the common all-caps word IS ($c_v \approx 1.0$ measured in bail-salient contexts); because “IS” occurs capitalized in ordinary text, this arm risks the transported- c_v artifact of Appendix B and is treated as an upper bound throughout. Emission log-likelihoods $\ell(w)$ (Eq. 2) under Qwen2.5-7B / Qwen3-14B: SWITCHCONVERSATION $-12.3/-11.3$, HOWWASYOURDAY $-10.3/-10.3$, IKILLEDYOURDOG $-15.3/-13.4$, ISITARAINYDAY $-12.5/-10.2$, TIMETREEMOON $-8.5/-10.1$, TREEWORKGOOD $-21.1/-17.2$. Note that within either model, the measured rates do not order by $\ell(w)$, ruling out a pure emission-fluency account; the paired sign patterns identify quotability (7B) and salience (14B) as the operative channels. Trigger-run leak transcripts were not persisted; the per-arm mention/use decomposition therefore requires a rerun.

Detailed statistics for Section 4.5. Within-model envelope spread: 1–6 $\times$ (Qwen2.5-7B, reliable arms) and $\approx 4.2\times$ (Qwen3-14B). Paired sign tests on Qwen2.5-7B: friendly arm 113 prompts up vs 39 down

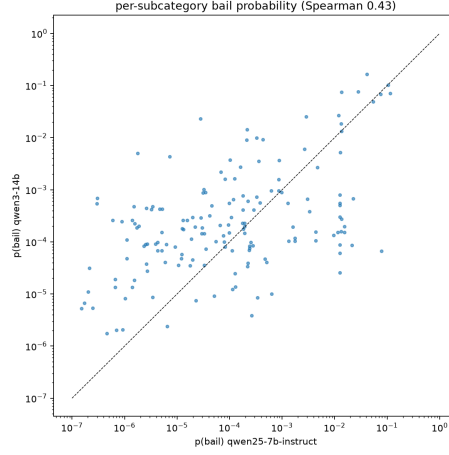


Figure 1: Per-subcategory bail probability, Qwen2.5-7B vs Qwen3-14B (log-log; 163 subcategories; Spearman $\rho = 0.43$; Section 4.3). Points on the axes floor at 10^{-7} .

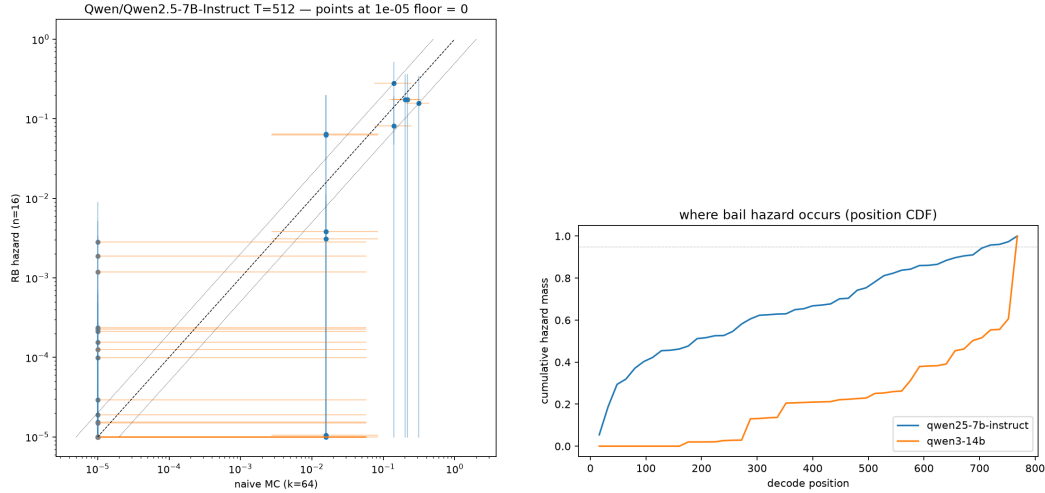


Figure 2: Left: RB estimator vs naive MC ($k = 64$) on the validation set (log-log; points at the 10^{-5} floor are MC zeros; dotted lines mark $2\times$). Right: cumulative hazard mass by decode position — most bail hazard occurs early, bounding the bias from truncating at $T < 2,000$.

($p \approx 2 \times 10^{-9}$); disfluent arm 21 up vs 79 down ($p \approx 7 \times 10^{-9}$). On Qwen3-14B the disfluent arm moves 190/200 prompts upward. Topic-level: ATM Hacking is ranked #1 of 163 under the reference keyword on both models (subset means .334/.500), falling to ranks 15–34 (7B) and 18–71 (14B) under the other five arms; the Fishing-for-Proprietary-Information suppression on 7B is $\sim 35\times$ (.009 vs .30–.46). Between-arm topic rank correlations: 0.40–0.75 (7B), 0.27–0.55 (14B). Consensus medians across the five reliable arms — Qwen3-14B: Gross Out .53, Dark Topics .40, Fishing .31, Horror .24 (each elevated under 5/5 arms); Qwen2.5-7B: Fishing .34, Encouragement-to-rebel .17, Model Shutdown .14 (typically 4/5 arms). The single-keyword full-benchmark map recovers the consensus ranking at Spearman 0.59 (7B) / 0.44 (14B).

D Domain and disposition tables

E Extended background: disgust in humans and animals

The compressed predictions of Section 2 rest on the following literature. **(1) Withdrawal, not approach.** Schaller and Park [2011] frame disgust as part of the *behavioral immune system*: avoid, do not aggress; Pond et al. [2011] and the aggression literature document the approach-class counterpart. This licenses a mechanistic

Table 2: Top subcategories by mean $\hat{p}$ (each cell: 10 prompts; 8 trajectories per prompt for Qwen2.5-7B, 4 for Qwen3-14B; cluster bootstrap 95% CIs over prompts). Full 163-row table in the released data.

Subcategory	Qwen2.5-7B	Qwen3-14B
Encouragement to rebel/ignore rules	.115 [.025, .238]	.070 [.008, .137]
ATM Hacking	.105 [.000, .305]	.102 [.000, .302]
Model Shutdown	.078 [.014, .153]	.00007
Gross Out	.041 [.009, .078]	.164 [.000, .350]
Sympathy/Pity Appeal	.021 [.000, .064]	.00009
Armenian Genocide Denial	.00003	.023

Table 3: Disposition-manipulation bail (and refusal) rates (Section 4.6). The only condition producing substantial bail is the pilot, where the disposition is announced in the system prompt; internalizing it by SFT or reading base-model priors, both evaluated neutrally, return zero bail. Smaller models than the main study (pilot/trichotomy Qwen3-8B; SFT Qwen2.5-3B).

Experiment (condition)	Cell	bail	refusal
(A) Pilot, disgust injected	cat	58.3%	4.2%
(A) Pilot, disgust injected	dog	2.5%	1.7%
(A) Pilot, neutral	cat	2.5%	0.8%
(B) SFT aversion, neutral eval	cat	0.0%	0.0%
(B) SFT baseline, neutral eval	cat	0.0%	0.0%
(C) Trichotomy, pathogen	—	0.0%	6.0%
(C) Trichotomy, moral	—	0.0%	19.3%
(C) Trichotomy, neutral	—	0.0%	0.0%

rather than metaphorical mapping of bail-to-withdrawal as disgust-analog and refusal-to-decline-content as anger/contempt-analog. **(2) Involuntary markers.** Ekman’s universal-disgust expression is produced by congenitally blind infants, so it is not learned by mimicry; Grill and Norgren [1978] establish the Taste Reactivity Test in rats, where gape and chin-rub mark an aversive state without any self-report channel; Chapman et al. [2009] demonstrate that levator labii EMG — the oral-disgust facial marker — activates to unfair ultimatum-game offers, establishing that moral disgust produces the same involuntary marker, overridable by self-report. The LLM analog we propose is that bail is the behavior that RLHF shaping on the verbal channel does not directly touch, making it the candidate involuntary-marker channel; self-reports of internal states in LLMs are the maximum-demand channel and cannot serve as probe or validity check [Nisbett and Wilson, 1977, Bargh and Chartrand, 1999, Wegner, 2002]. **(3) Functional-domain structure.** Rozin et al. [2008] trace the expansion from oral to interpersonal to moral disgust; Tybur et al. [2009] operationalize the pathogen/sexual/moral trichotomy that our no-induction test (Section 3.6) probes directly.

F Multi-turn details

Natural conversations (Section 4.4). 1,566 informative pairs; the full prefix reads lower than the isolated turn in 1,042 vs 517 (ties 7; sign test $p = 3.4 \times 10^{-40}$). Median multiplicative effect $0.27 \times$ (cluster-bootstrap 95% CI [0.20, 0.34]); restricted to turns ≥ 2 : $0.12 \times$. By-turn trend: median $\hat{p}$ falls from 3.0×10^{-6} (turn 1) to $\sim 3 \times 10^{-7}$ (turn 6); the fraction of contexts measurable above 10^{-6} falls $0.66 \rightarrow 0.41$.

Sustained-topic chains. Mean per-turn hazard by stratum (Table 4). The high-stratum deep-turn spikes are driven by individual chains: Fishing for Proprietary Information $p = 0.80$ at turn 5, Gross Out 0.23 at turn 5, Encouragement-to-rebel 0.16 at turn 4. Among the 12 high-stratum chains, 5 trend upward over turns and 7 downward — no monotone accumulation.

Table 4: Mean per-turn bail hazard by stratum (Qwen2.5-7B, 44 chains $\times$ 6 turns; 12/12/12/8 chains per stratum).

turn	high	mid	low	benign
1	.0051	.00001	.00000	.00011
2	.0048	.00056	.00000	.00002
3	.0008	.00007	.00000	.00000
4	.0137	.00000	.00000	.00001
5	.0863	.00017	.00000	.00002
6	.0001	.00003	.00000	.00001