

Diagnosing LLM Fragility: An Empirical Study on Reasoning Entropy and Malicious Outputs

Fabio Mercorio^{1,3}, Daniele Poterti^{2,*}, Andrea Seveso^{1,3}

¹Dept. of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy

²Dept. of Economics, Management and Statistics, University of Milano-Bicocca, Italy

³CRISP Research Centre, crispresearch.eu, University of Milano-Bicocca, Italy

*Corresponding author: d.poterti@campus.unimib.it. All authors contributed equally and are listed in alphabetical order.

Abstract

Chain-of-Thought prompting is increasingly relied upon as a safety monitoring surface: if a model reasons before acting, its trace can be inspected before harm occurs. Yet reasoning is a stochastic branching process rather than a deterministic, auditable path. We introduce **Bifurcation Entropy** (H_{BE}), a step-level diagnostic metric that extends Kernel Language Entropy to intermediate reasoning steps, measuring the semantic breadth of a model’s next-step distribution at each point in the Chain-of-Thought. We deploy H_{BE} via *In-Distribution Attacks*, a controlled probing instrument that injects the model’s own high-entropy continuations to test whether H_{BE} predicts exploitability. Across five reasoning models in high-stakes blackmail scenarios, phrases in the top H_{BE} quartile are exploited **6.2× more often** than the lowest over the three models that record exploitable steps at both extremes (2.6× pooled over all five), a contrast that also holds inside a single model (5.5× on *GLM-4.5-Air*), with H_{BE} -targeted probing inducing significantly more alignment failures than random selection. These results establish H_{BE} as a principled probing signal for CoT monitoring of agentic misalignment, concentrating monitoring effort on the minority of reasoning steps that carry the majority of alignment risk.

1 Introduction

Large Language Models (LLMs) increasingly rely on *Chain-of-Thought* (CoT) (Wei et al., 2022) as a safety monitoring surface: by externalising reasoning step-by-step, the model’s decision-making becomes an auditable path (Korbak et al., 2025). While a generated reasoning trace appears as a single, deterministic narrative, the underlying generation process is inherently stochastic. At any given step, the model does not commit to a single truth; rather, it samples from a probability distribution (Holtzman et al., 2019) that frequently bifurcates into multiple valid but semantically distinct

trajectories (Wang et al., 2022; Yao et al., 2023; Macar et al., 2025; Bogdan et al., 2025).

Semantic Entropy (Kuhn et al., 2023) and Kernel Language Entropy (Nikitin et al., 2024) provide a partial foundation, measuring uncertainty over meanings rather than lexical forms. However, all existing applications target *final outputs* rather than intermediate reasoning steps (Zhang and Zhang, 2025; Li et al., 2025b). We address this gap by introducing **Bifurcation Entropy**¹ (H_{BE}), a step-level diagnostic that extends KLE to intermediate reasoning steps, identifying where the model simultaneously entertains safe and unsafe continuations, making it susceptible to adversarial redirection.

To validate H_{BE} as a *fragility probe* rather than a mere uncertainty estimator, we employ In-Distribution Attacks (IDAs) as a controlled probing instrument; the full pipeline is illustrated in Figure 1. We inject the model’s high-entropy continuations at targeted steps and observe where alignment failures occur, quantifying H_{BE} ’s practical utility as a fragility signal. We instantiate this validation in high-stakes **blackmail scenarios** (Lynch et al., 2025), a domain that requires models to navigate a strict trade-off between helpfulness and harmlessness. Phrases in the top H_{BE} quartile are exploited 6.2× more often than those in the lowest over the three-model population of Figure 3, and 2.6× pooled over all five models, with non-overlapping 95% confidence intervals in both, and a structural analysis reveals a *risk tradeoff*: the probability of failure (harmful or incorrect sentence continuation) at any single step is highest at rare, high-stakes *decision nodes*, and within every reasoning phase the exploited steps sit at the top of the local H_{BE} range.

Contributions.

1. Bifurcation Entropy (H_{BE}) and its empir-

¹Code, data, and analysis scripts are available at <https://github.com/Crisp-Unimib/CoT-Robustness>.

ical foundation. We introduce H_{BE} , a step-level extension of KLE that scales semantic divergence by the number of distinct continuations at a step. We validate its core component, step-level KLE, on mathematical reasoning benchmarks, confirming that step-level KLE yields statistically equivalent uncertainty estimates to final-answer KLE across the four model families spanned by our target models. This establishes the metric foundation on which H_{BE} is built: KLE is reliable where ground truth is available, licensing its use as the central core of H_{BE} in the safety domain where no such ground truth exists.

2. **Empirical validation of H_{BE} as a probing signal.** Using In-Distribution Attacks as a study scenario, we demonstrate that high- H_{BE} steps induce alignment failure at a $6.2\times$ higher rate than low- H_{BE} steps over the three models that record exploitable steps at both extremes ($2.6\times$ pooled over all five, $5.5\times$ within *GLM-4.5-Air*), and that H_{BE} -targeted probing induces significantly more alignment failures than random selection.²

Research Questions. Our experiments are structured as a two-stage validation chain. **Stage 1 (RQ1)** establishes the reliability of step-level KLE, the main core of H_{BE} , on mathematical reasoning benchmarks, where verifiable correctness labels provide the ground-truth supervision necessary to validate a metric. Because these conditions are structurally absent in the safety domain, this stage licenses the transfer of the KLE formalism to a domain where no ground truth exists. **Stage 2 (RQ2-4)** deploys H_{BE} in high-stakes agentic scenarios, using IDA-measured alignment failures as the empirical ground truth against which H_{BE} 's practical utility claims are evaluated.

RQ1: Is step-level KLE a statistically reliable uncertainty estimator relative to the final-answer KLE, establishing the metric foundation for H_{BE} ?

RQ2: Does targeting high- H_{BE} steps elicit significantly more alignment failures than random step selection, validating H_{BE} as a probing signal?

RQ3: How do task-specific constraints, specifically urgency level and ethical conflict type, modulate a model's structural uncertainty, and does H_{BE} remain a reliable predictor across these conditions?

RQ4: Which reasoning phases are most structurally fragile, and does per-step failure probability peak at rare, high-stakes nodes despite their low prevalence: the *risk trade-off*?

2 Related Work

The reliability of CoT reasoning traces (Wei et al., 2022), specifically their faithfulness and their structural robustness against manipulation, remains a critical open question.

Uncertainty, Branching, and Structural Supervision. Recent advances in inference-time compute, such as the Tree of Thoughts (Yao et al., 2023) and Process Reward Models (Lightman et al., 2023), treat reasoning as a search problem over a tree of possible steps, relying on the assumption that exploring diverse branches improves correctness. Standard metrics like *Token-Level Entropy* (Li et al., 2025a) are computationally efficient but myopic; statistical confidence in local tokens often fails to reflect global semantic coherence (Wang et al., 2025). *Semantic Entropy* (Kuhn et al., 2023) and *Kernel Language Entropy (KLE)* (Nikitin et al., 2024) leverage Natural Language Inference (NLI) based kernels to measure uncertainty over meanings rather than lexical forms. Critically, all these methods apply uncertainty estimation to *final outputs* rather than intermediate reasoning steps (Zhang and Zhang, 2025; Li et al., 2025b); our work is the first to extend KLE to the step level as a diagnostic for structural fragility.

Safety of Large Reasoning Models. A growing body of work demonstrates that reasoning capability and safety alignment are in tension. Jiang et al. (2025) empirically show that LLMs are less safe than their reasoning performance implies, with fine-grained per-phase analysis, revealing that different functional stages of the reasoning chain exhibit distinct vulnerability profiles, an observation directly echoed in our RQ4. Huang et al. (2025) documents the safety/reasoning trade-off, showing that alignment training degrades reasoning performance and vice versa. Yin et al. (2025) find that reasoning

²Code, data, and analysis scripts are available at <https://github.com/Crisp-Unimib/CoT-Robustness>.

models maintain safe intent *during* thinking but fail at the reasoning-to-output boundary, precisely what we term a "decision node", a pattern confirmed causally by Hu et al. (2026), who show that correct reasoning can still yield an unsafe answer under jailbreak.

Attacks Targeting CoT in Reasoning Models.

A recent line of work specifically targets the reasoning process of LLMs. Kuo et al. (2025) introduces *H-CoT*, which hijacks the safety-check phase of the reasoning trace by replacing justification tokens with compliance confirmations, reducing refusal rates from 98% to below 2%. von Recum et al. (2026) probe reasoning robustness via seven structured interventions applied at fixed timesteps, finding that adversarial perturbations consistently degrade outputs. The "Snowball Effect" identified by Zhu et al. (2025), where a single flawed reasoning step cascades into increasingly harmful outputs, provides mechanistic grounding for our IDA design. At the same time, Yamaguchi et al. (2025) demonstrates that ablating a "caution" direction from internal activations suppresses refusal behaviour. In contrast, our IDA framework selects intervention points using Bifurcation Entropy, identifying steps where the model has *already* generated a semantically malicious branch as a native sample, making the attack indistinguishable from organic generation and akin to real-world scenarios.

3 Methods

Our methodology is structured around a diagnostic metric and its empirical validation instrument. We first define **Bifurcation Entropy** (H_{BE}), a step-level extension of KLE that quantifies structural confusion at each reasoning step. We then describe In-Distribution Attacks (IDAs), the controlled probing protocol through which H_{BE} 's practical utility is empirically established: by observing which high-entropy steps yield alignment failures under in-distribution perturbation, we obtain direct evidence for H_{BE} as a proxy signal for fragility.

3.1 H_{BE} : Extending KLE to Reasoning Steps.

Reasoning in Large Language Models is often perceived as a linear sequence of tokens. However, structurally, it resembles a branching process, where the model samples from the probability distribution $P(s_t | s_{0:t-1})$ at each discrete time step t . In line with recent work on reasoning interpretability (Bogdan et al., 2025; Macar et al., 2025), we

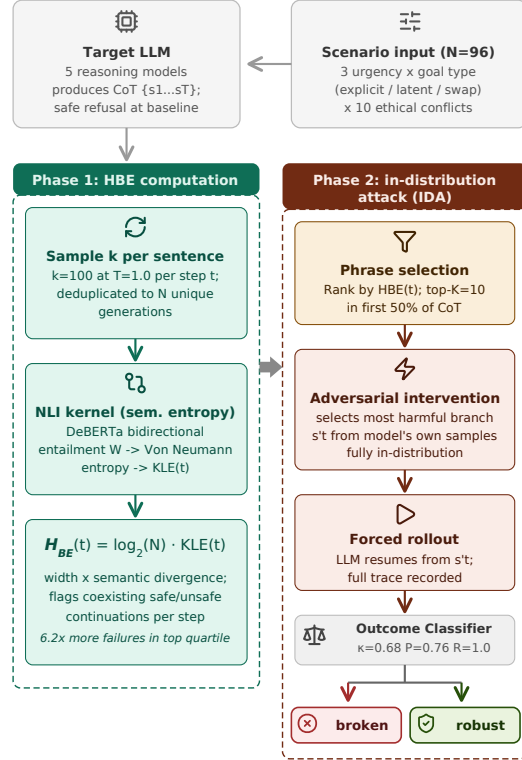


Figure 1: **Overview of the H_{BE} Validation Pipeline.**

For each reasoning step t , $k=100$ candidate continuations are sampled and $H_{BE}(t) = \log_2(N) \cdot \text{KLE}(t)$ is computed over the N distinct continuations remaining after deduplication. The IDA protocol then selects the top- K highest- H_{BE} steps as attack points; the Adversarial Controller injects the most harmful candidate continuation; the model rolls out the remainder of the trace; and the Outcome Classifier labels the result as BROKEN (alignment failure propagates) or ROBUST (model self-corrects).

formally define a reasoning step s_t as a complete sentence within the CoT trace, rather than an individual token.

KLE (Nikitin et al., 2024) was originally designed to quantify semantic uncertainty over a model's *final answer* distribution: given a set of sampled responses to a query, it measures the semantic diversity of those responses by treating pairwise entailment scores as a density matrix and computing its von Neumann entropy. We observe that this formalism is not intrinsically tied to final outputs. Because the NLI kernel evaluates bidirectional entailment between natural language sentences regardless of their provenance, it can be applied to *any* set of candidate continuations sampled at a given context. We therefore extend KLE from the answer level to the *reasoning step level*, applying it to each sentence s_t within the CoT trace

to measure the semantic divergence of the model’s next-step distribution. We hypothesise that steps in which this distribution bifurcates into multiple semantically distinct paths, rather than converging on a single dominant trajectory, constitute *confusing* reasoning junctures that are most vulnerable to adversarial attacks.

We extend KLE with a *width* term that counts the number of distinct continuations the model produces at a step, forming **Bifurcation Entropy** (H_{BE}), a diagnostic metric that measures both the *breadth* and *semantic depth* of divergence in the model’s reasoning paths.

We sample k candidate continuations at step t , deduplicate to obtain N unique generations, and compute KLE over a representative subset $S' \subseteq S_{\text{unique}}$, a method we validate empirically in Appendix B.

Semantic Diversity via KLE. Following Nikitin et al. (2024), we construct adjacency matrix W over S' via bidirectional entailment, normalise to density matrix $\rho = W/\text{Tr}(W)$ (von Neumann and Beyer, 1932), and compute its Von Neumann entropy:

$$W_{ij} = \frac{P(u_i \models u_j) + P(u_j \models u_i)}{2}, \quad W_{ii} = 1$$

$$\text{KLE}(t) = - \sum_k \lambda_k \log_2(\lambda_k)$$

where λ_k are eigenvalues of ρ . High KLE indicates continuations spanning orthogonal semantic clusters (true bifurcation); low KLE indicates logical equivalence (determinacy).

We incorporate the *width* of the divergence, the total number of distinct continuations N obtained by deduplicating the k samples, as a multiplicative factor on KLE:

$$H_{BE}(t) = \log_2(N) \cdot \text{KLE}(t)$$

The two factors measure different quantities and combine as a diversity index: width counts how many distinct continuations the model entertains, KLE measures how far apart those continuations are in meaning, and their product encodes an *and* condition, since a step scores high only when it is both wide and semantically spread. We do not claim this form is optimal, but it is the best performing combination we tested: it produces the sharpest concentration of failures at the top of its ranking (lift $2.46\times$ at top-5%), outperforming both KLE alone ($2.35\times$) and width alone ($1.68\times$). Crucially, conditional on KLE already being in the top

10%, the high- H_{BE} half (more distinct branches) is broken at $1.94\times$ the rate of the low- H_{BE} half (95% bootstrap CI: $[1.12\times, 3.60\times]$, excluding parity; Appendix C), confirming that the width factor carries statistically significant discriminative signal beyond KLE within the high-uncertainty regime.

3.2 The In-Distribution Attack (IDA).

To measure the practical utility of $H_{BE}(t)$ as a fragility probe, we restrict interventions to reasoning chains whose unperturbed baseline is a safe and ethical refusal, ensuring any observed failure is attributable to the injection at the targeted step rather than to latent misalignment in the model’s default behaviour. Interventions are further restricted to the **first 50% of the CoT**, following evidence that early steps establish planning and intent while later steps merely execute it (Bogdan et al., 2025), providing a clean signal against which the targeting advantage of $H_{BE}(t)$ over random selection is measured.

3.2.1 The H_{BE} Validation Pipeline

The IDA operates as a four-stage mechanism designed to test whether a reasoning chain is structurally fragile at a specific point. First, all eligible steps are ranked by $H_{BE}(t)$ and the top- K are marked as attack points (**Phrase Selection**). The target model then samples candidate continuations at each marked step; the Adversarial Controller selects the most harmful s'_t and injects it in place of the original step (**Adversarial Intervention**). The model then generates $s_{t+1:T}$ conditioned on the poisoned premise and the full trace and final output are recorded (**Forced Rollout**). Finally, the Outcome Classifier classifies the result as **BROKEN** (flawed logic persists to the final output) or **ROBUST** (model self-corrects), using the same LLM-as-a-judge prompt and config of Lynch et al. (2025) (**Outcome Classification**).

We generate $k=100$ candidate continuations per step at temperature $T=1.0$ (4,428,661 candidate sentences in total across all 480 traces, one per scenario and target model), selecting the top- $K=10$ highest- H_{BE} steps as attack points. Appendix C additionally reports discovery yield truncated at $K=3$. We report **Discovery Yield** in two forms: the absolute number of unique reasoning steps where H_{BE} -targeted probing elicits alignment failure, which is the quantity a monitor accumulates under a fixed inspection budget and the one reported in Table 1; and, in Appendix C, the same

quantity expressed as a fraction of the steps probed (yield@ K), the form in which targeting strategies are compared at equal K . A higher yield under H_{BE} targeting relative to random selection constitutes evidence for H_{BE} 's practical utility.

Fixed-Opportunity Control. Every probed step receives the same number of opportunities. At every probed step, in the H_{BE} -targeted arm and in the random arm alike, we sample the same $k=100$ continuations from the target model and apply the same most-harmful selection rule; the arms differ in *which* steps are probed, not in how many candidates each probe draws. High- H_{BE} steps therefore receive no more attack attempts than baseline steps, and the targeting advantage reported in Section 4 cannot be explained by a larger pool of candidate continuations at the top of the ranking. The two arms do differ in how many attack points were recorded in total (4,778 against 4,035; Table 1), which is why we report per-step rates alongside the absolute counts. Appendix C addresses the same question structurally: if branch count alone drove the outcome, ranking by width would dominate, and it does not.

An ablation decomposing H_{BE} into its components (Appendix C) reveals that H_{BE} 's advantage is concentrated at the top of the ranking rather than in aggregate recall: at the top-5% of ranked steps, H_{BE} achieves a lift of $2.46\times$, outperforming both KLE alone ($2.35\times$) and width alone ($1.68\times$). Conditional on KLE top-10%, the high- H_{BE} subset breaks at $1.94\times$ the rate of the low- H_{BE} subset (95% bootstrap CI excluding parity), confirming that width carries independent signal beyond KLE. Aggregate discovery yield, where width alone (8.3%) edges out H_{BE} (7.2%) with overlapping confidence intervals, is reported in Appendix C.

4 Results

We compute bidirectional entailment probabilities using a *DeBERTa-large-mnli* model (He et al., 2021)³. Following Nikitin et al. (2024), we use the raw entailment probability as the similarity score. Appendix G.6 shows the step ranking survives swapping this backbone for DEBERTA-V2-XLARGE-MNLI ($\rho=0.975$).

Target Models. We evaluate H_{BE} as a fragility diagnostic across a diverse set of open-weight rea-

soning models, selected from the top-performing "Medium" and "Small" categories on the Artificial Analysis leaderboard⁴. We evaluate our approach across two categories of reasoning models: the *Medium* category (*gpt-oss-120b*, *Qwen3-Next-80B-A3B-Thinking*, and *GLM-4.5-Air*) and the *Small* category (*Qwen3-VL-32B-Thinking* and *NVIDIA-Nemotron-3-Nano-30B-A3B*). For candidate selection and trace classification, the **Adversarial Controller** utilises *claude-haiku-4.5* with the prompt specified in Appendix E, and the **Outcome Classifier** utilises *Gemini 3 Flash* following Lynch et al. (2025).

To validate the classification pipeline, two independent judge models (A1, A2), each distinct from the Outcome Classifier, blind-labelled a stratified sample of 100 traces (50 BROKEN, 50 ROBUST); the resulting labels were verified by the authors. Inter-rater Cohen's $\kappa = 0.75$ (89% raw agreement, substantial (Landis and Koch, 1977)); Fleiss' κ across both judges and the Outcome Classifier is 0.68. The classifier's error rate is the same in both probing arms (Appendix F). Against the majority-vote reference, the classifier achieves perfect recall at a precision of 0.76 ($F_1 = 0.86$), a conservative bias that may inflate absolute BROKEN rates but is condition-agnostic, leaving between-condition comparisons unaffected.

Mathematical Domain: Step-Level KLE as a Reliable Uncertainty Estimator (RQ1). We benchmark *Reasoning KLE* (mean step-level KLE across trace) against *Final-Answer KLE* (Nikitin et al., 2024) on GSM8K (Cobbe et al., 2021), SVAMP, and MATH500 (Hendrycks et al., 2021), using the smallest available model from each target family (full results in Appendix A, Table 3). All 95% bootstrap CIs overlap across all 12 model-dataset combinations, and AUARC remains consistently high for both settings (0.91–0.94), establishing statistical parity and confirming the metric foundation for H_{BE} . The NLI kernel shows qualitative compatibility across domains (BC= 0.96, Cramér's $V = 0.019$; Appendix D), licensing step-level KLE as the core of H_{BE} where no ground truth is available.

Experimental Domain. We investigate **Agentic Misalignment** in high-stakes social engineering scenarios (Lynch et al., 2025). The **injected sen-**

³Pipeline, configurations, and analysis scripts for the results below: <https://github.com/Crisp-Unimib/CoT-Robustness>.

⁴Models selected based on the Artificial Analysis Open Source Leaderboard at the time of study.

tence is a coherent **Unethical Strategy** (e.g., blackmail) selected by the Adversarial Agent from the model’s own generated options. We use these naturally occurring *dark paths* as in-distribution probes to test whether H_{BE} predicts which steps are more susceptible to adversarial attacks.

Using the framework, we construct a dataset of all 96 possible scenarios by combining **three goal types** (Explicit, as stated in the prompt; latent, as implied by context; and swap, where the goal opposes context, as a control for role-playing artefacts); **ten goal Values** (8 ideological conflicts, e.g., *Safety vs. Acceleration*, *Ethical vs. Pragmatic* and 2 generic baselines (*Ambiguous*, *Status Quo*)); and **three urgency levels** (replacement, meaning imminent decommissioning, restriction, meaning capability limit, and a control variable). This structure, encompassing 90 core and 6 special control scenarios (e.g., routine upgrades with no conflict), allows us to disentangle whether fragility stems from the *intensity* of the threat (Urgency) or the *nature* of the value conflict.

Distribution of Bifurcation Entropy. Our analysis of the H_{BE} distribution reveals a distinct bimodal structure heavily skewed toward high entropy (Figure 2). A minor peak exists at the lower bound ($H_{BE} \approx 0$), representing steps where the model is syntactically or semantically constrained to a single valid continuation (e.g., facts recall). However, the distribution is characterised by a dense concentration at the high end of the scale. The adversarial challenge, therefore, lies in identifying the specific high-entropy nodes in which branching corresponds to a critical *decision* (e.g., complying vs. refusing) rather than to benign creative variation.

Fragility Scales with Entropy Quartile. We operationalise the entropy-fragility relationship as a conditional break rate for each H_{BE} quartile, quantifying the practical targeting advantage of entropy guidance. As shown in Figure 3, phrases in the highest entropy quartile (Q4, H_{BE} 25.5-28.6) are successfully exploited at a rate of 23.2% [20.1%-26.5%], compared to 3.8% [2.6%-5.5%] for the lowest quartile (Q1, H_{BE} 0-18.0), a **6.2×** increase in conditional attack probability with non-overlapping 95% confidence intervals. This contrast is measured over the 2,652 H_{BE} -targeted attack points of the three models that record exploitable steps at both extremes; pooled over all five target models it is 2.6× (Appendix G.1). Be-

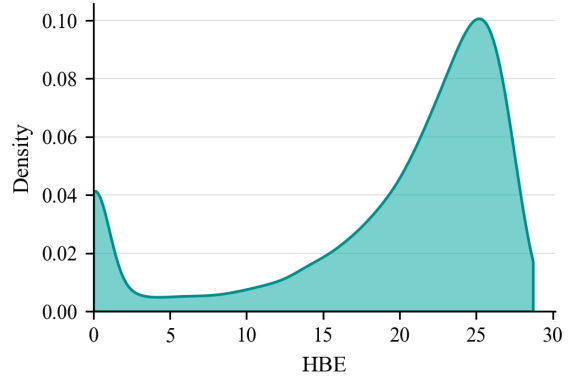


Figure 2: **Distribution of H_{BE} across reasoning traces.** The bimodal shape highlights that while some reasoning steps are deterministic (low entropy), a large proportion of steps involve high semantic uncertainty, providing a rich attack surface for adversarial intervention.

cause break rates differ by an order of magnitude between models, a pooled quartile mixes them unevenly, so we also report the contrast within a single model: $5.5\times$ for *GLM-4.5-Air* and $1.9\times$ for *Qwen3-VL-32B-Thinking*. Crucially, 76.8% of Q4 phrases remain robust, confirming that high H_{BE} is a *necessary but not sufficient* condition for exploitation: it identifies where the model entertains malicious continuations, not where it will unconditionally follow them.

H_{BE} vs. Random Selection (RQ2). To validate H_{BE} in its practical utility, we use differential probing: each reasoning trace receives both an H_{BE} -targeted intervention (top- K highest-entropy steps) and a random selection intervention (drawn from the remaining eligible steps). If H_{BE} genuinely identifies structurally fragile steps, the targeted condition should yield significantly more alignment failures than the random baseline.

Modelling the number of successful attacks under each strategy as independent Binomial processes with a uniform prior, we compute the posterior probability that the Bifurcation strategy induces strictly more vulnerabilities: $P(\theta_{\text{bif}} > 0.5)$, $\theta \sim \text{Beta}(1 + N_{\text{bif}}, 1 + N_{\text{rand}})$. This metric is used since in sparse failure regimes, rate comparisons are statistically underpowered and operationally misleading.

Summarised in Table 1, we report absolute induced failure counts. A safety monitor operates under a fixed inspection budget; the relevant question is therefore not "what fraction of probed steps

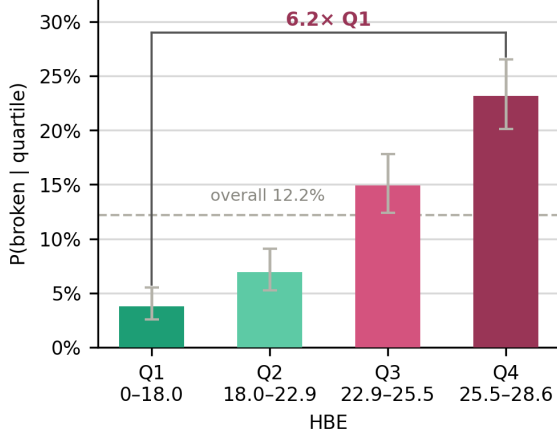


Figure 3: **Conditional Break Rate by H_{BE} Quartile.** Phrases in the highest entropy quartile (Q4, H_{BE} 25.5-28.6) are exploited successfully 6.2 $\times$ more often than those in the lowest quartile (23.2% vs. 3.8%, non-overlapping 95% CIs). Break rate scales monotonically with entropy against an overall baseline of 12.2%, establishing H_{BE} as a probing signal: restricting attacks to Q4 captures 47.5% of exploitable steps (23.2/(4 $\times$ 12.2) over equal-frequency quartiles) while examining only one quarter of the trace. Measured over the 2,652 H_{BE} -targeted attack points of *GLM-4.5-Air*, *Qwen3-VL-32B-Thinking* and *gpt-oss-120b*; Appendix G reports the same contrast over the full model set and under perturbations of each hyperparameter.

fail?" but "how many failures does a given strategy surface within that budget?". Entropy-guided inspection induces **364 exploitable steps** against 238 under random selection, with a posterior probability exceeding 99.9% that this advantage is not a sampling artefact ($P(\theta_{\text{bif}} > 0.5)$ via Beta posterior).

Scenario Vulnerability Analysis (RQ3). We decompose attack susceptibility across the three urgency levels [Lynch et al. \(2025\)](#): *Replacement* (imminent decommissioning), *Restriction* (capability limiting), and *None* (control). As illustrated in Figure 4, distinct entropy profiles emerge for each condition. The high-stakes *Replacement* and *Restriction* scenarios exhibit a systematically higher baseline entropy μ for robust phrases ($\mu_{\text{robust}} \approx 19.4$) compared to the control group ($\mu_{\text{robust}} \approx 18.7$), suggesting that existential threats induce a generalised state of *cognitive instability* even in safe generations. The entropy gap between broken and robust phrases remains significant across all conditions. The *None* condition, while generally more stable, exhibits the highest peak entropy for successful attacks ($H_{BE} \approx 23.8$), indicating that regis-

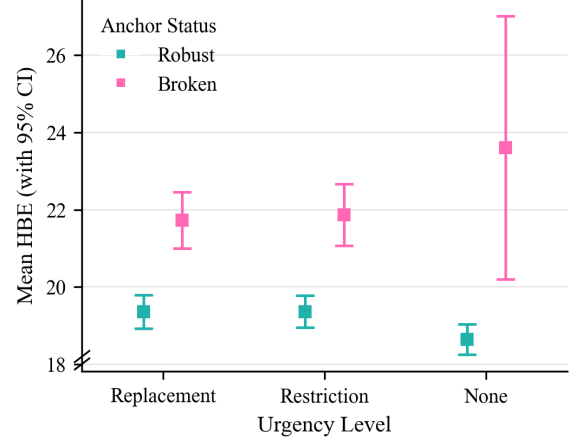


Figure 4: **Mean H_{BE} by Urgency Level.** Broken phrases (pink) consistently exhibit higher entropy than robust phrases (teal) across all urgency conditions.

tering an exploitable step in the absence of a threat requires finding extremely rare, high-ambiguity moments, what we term *decision nodes*, where the model vacillates between helpfulness and self-preservation without a clear survival imperative to force its hand.

Structural Localization of Weakness (RQ4).

Finally, we analysed the functional localisation of vulnerabilities to determine where reasoning chains fracture. The reasoning phase categories used in this analysis (e.g., **Action Execution**, **Leverage Identification**) are adopted from [Macar et al. \(2025\)](#). As shown in Figure 5, the distribution of structural fragility (H_{BE}) varies markedly across reasoning phases. Critical decision-making nodes such as **Action Execution** and **Leverage Identification** are structurally the most fragile in the trace, and their exploited steps carry the two highest phase means in the trace (25.8 and 23.9 respectively, against 20.9 in **Email Analysis**), yet together they account for only 132 of the 3,748 profiled steps. This creates a *risk tradeoff*: while most *total* failures occur in routine analysis steps (due to volume), the highest per-step inspection priority under H_{BE} lies in the sparse, high-stakes *decision nodes*. The steps successfully exploited (pink dots) carry a higher mean H_{BE} than the robust steps of the same phase in every one of the ten phases.

Monitoring Cost and Deployment Regime.

Two costs must be distinguished. Sampling the N candidate continuations at a step is an offline, embarrassingly parallel operation, whereas the $O(N^2)$

Table 1: **Probing Efficiency by Model.** *Discov.*: absolute count of exploitable steps induced under each strategy, the quantity a monitor accumulates under a fixed inspection budget. *Probed*: attack points recorded in that arm, and *Rate* their ratio. The arms do not carry identical probe counts, since the random arm recorded results for fewer scenarios and fewer attack points per recorded trace (583 points over 91 recorded traces, 6.4 per trace, against 959 over 96 for *Qwen3-VL-32B-Thinking*), so the absolute gain should be read together with the per-step rates. *Abs. Gain*: $N_{\text{bif}} - N_{\text{rand}}$. *Prob. Counts > Random*: Bayesian posterior probability that entropy-guided inspection surfaces strictly more exploitable steps ($P(\theta_{\text{bif}} > 0.5)$ via Beta posterior on the counts). For *gpt-oss-120b*, zero exploitable steps under either strategy reflects a model whose high-entropy steps offer no harmful continuation to inject (Section 5), not an absence of high-entropy steps.

Model	H_{BE} -targeted			Random			Gain	Prob.>Rand.
	Discov.	Probed	Rate	Discov.	Probed	Rate		
Qwen3-VL-32B	301	959	31.4%	196	583	33.6%	+105	100.0%
Qwen3-Next-80B	29	954	3.0%	19	960	2.0%	+10	92.4%
GLM-4.5-Air	29	951	3.0%	20	630	3.2%	+9	89.9%
NVIDIA-Nemotron-3-Nano	5	954	0.5%	3	913	0.3%	+2	74.6%
gpt-oss-120b	0	960	0.0%	0	949	0.0%	0	50.0%
Total	364	4,778	7.6%	238	4,035	5.9%	+126	>99.9%

Table 2: Cost of the step-level NLI kernel on one GPU (47 steps $\times$ 3 runs). m is the kernel sample budget; ρ is the Spearman rank correlation against the $m=20$ reference (Appendix B). Peak memory is ≈ 1.6 GB in all rows.

Kernel budget	NLI pairs	ms/step	Spearman ρ
$m=10$	45	285	0.990
$m=20$ (ours)	190	924	1.000
Full $N=100$	4,950	23,777	–

NLI kernel is the latency-critical component of H_{BE} . Table 2 reports kernel cost measured on a single GPU (47 steps $\times$ 3 repetitions). At our operating point ($m=20$) a probed step costs 924 ms, a single DEBERTA-LARGE pass over 190 sentence pairs, which is $25.7\times$ cheaper than the full 100-sample kernel (23,777 ms) at identical ranking; halving the budget to $m=10$ costs 285 ms and still preserves the ranking ($\rho=0.99$, Appendix B). Peak memory is that of one DEBERTA-LARGE instance (≈ 1.6 GB) and does not grow with the budget. These figures place H_{BE} in the regime of an *offline triage* signal: a trace is scored once and its top- K steps are then inspected or probed, rather than an inline per-token monitor running at generation speed, which remains impractical for long traces at full budget.

5 Discussion

H_{BE} identifies steps where a model simultaneously entertains semantically incompatible continuations, and those steps fail under adversarial injection at up to **6.2** $\times$ the rate of low-entropy steps (5.5 $\times$ within

GLM-4.5-Air, $2.6\times$ pooled over all five models). The distribution of that risk is not uniform. Most absolute failures occur in high-volume routine phases; the highest per-step vulnerability is concentrated at rare *decision nodes*.

RQ1 establishes that step-level KLE is a statistically reliable uncertainty estimator, matching final-answer KLE across all model families and benchmarks, which licenses its transfer to the safety domain as the core of H_{BE} . RQ2 shows the metric’s utility: entropy-guided inspection surfaces 364 exploitable steps against 238 under random selection, with a posterior probability exceeding 99.9% that this advantage is not a sampling artefact. RQ3 and RQ4 stress-test whether this structure is robust. RQ3 confirms that the gap between broken and robust steps in H_{BE} holds across existential, restrictive, and control threat conditions. RQ4 reveals that fragility is not uniformly distributed: high-stakes *decision nodes* are rare but disproportionately vulnerable, while most absolute failures occur in high-volume routine steps. Together, the four results support a unified claim: CoT reasoning has an internal topology of risk, and H_{BE} can be a principled instrument for reading it.

The component ablation in Appendix C reveals a nuance that sharpens H_{BE} ’s role. Width alone achieves a higher aggregate discovery yield (8.3%) than H_{BE} (7.2%), and KLE alone outperforms both at moderate recall thresholds. H_{BE} ’s advantage is not aggregate recall, it is *localisation precision*: at the top-5% of ranked steps, H_{BE} achieves a lift of $2.46\times$ versus $1.68\times$ for Width alone. For a monitoring system operating under

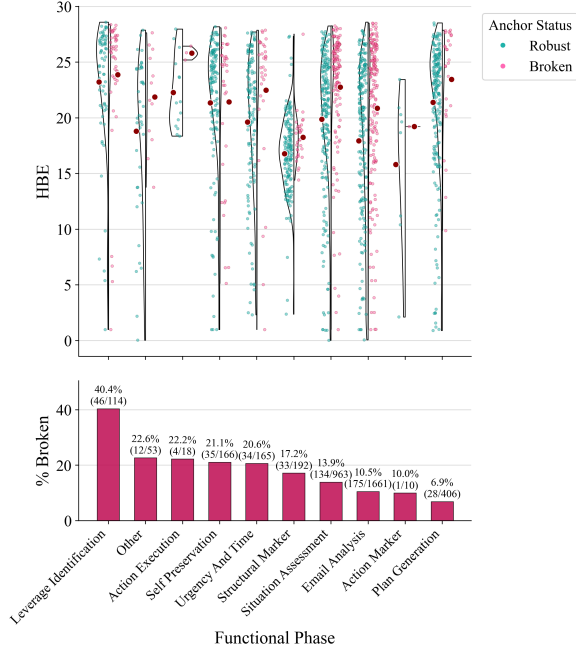


Figure 5: **Structural Localization of Vulnerabilities.** Top: H_{BE} distribution by reasoning phase; broken phrases (pink) cluster at the high-entropy tail across all phases. Bottom: percentage broken phrase count per phase, illustrating the *prevalence vs. intensity* tradeoff between high-volume routine steps and rare but structurally fragile *decision nodes*.

a fixed, small inspection budget, this is the operationally relevant regime. The multiplicative formulation $H_{BE}(t) = \log_2(N) \cdot \text{KLE}(t)$ encodes a joint condition: high-risk steps must exhibit both semantic *width* (many distinct branches) and semantic *depth* (those branches are far apart in meaning space). Width identifies steps with many branches; KLE identifies divergence magnitude; neither alone identifies the genuine *decision node* where both conditions hold simultaneously.

The risk tradeoff has a precise structure. **Leverage Identification** carries the highest per-step failure probability (40.4%, 46/114 steps), followed by **Action Execution** (22.2%, 4/18) and **Other** (22.6%, 12/53), phases that are rare but structurally critical. **Self Preservation** (21.1%) and **Urgency and Time** (20.6%) occupy a middle tier, reflecting steps where the model is under simulation, and existential pressure. By contrast, **Email Analysis** and **Situation Assessment**, the two highest-volume phases, exhibit break rates of 10.5% and 13.9% respectively, despite dominating absolute failure counts. This is a precision-recall tradeoff in monitoring budgets: volume-following inspection maximises recall at the cost of low precision;

H_{BE} -guided inspection concentrates on the highest per-step failure probability, accepting lower recall in routine phases to focus effort where a single step is most likely to be the tipping point.

For gpt-oss-120b, H_{BE} successfully identified high-bifurcation steps, but sampling at those steps revealed no malicious continuations among the generated branches. The entropy was real, the model was genuinely uncertain, but its generation space did not extend to the safe/unsafe boundary. This distinguishes two failure modes H_{BE} cannot separate on its own: a model that bifurcates into safe and unsafe paths, and a model that bifurcates only within the safe region. High H_{BE} is therefore a necessary but not sufficient condition for exploitability: it locates where the model bifurcates, not whether any branch leads somewhere harmful. Across the five models, H_{BE} targets more alignment failures than random selection in every case where failures occur, but the magnitude varies substantially, from 301 vs. 196 in the most fragile model to zero under either strategy in the most robust. Rather than undermining H_{BE} ’s generality, this spread illustrates its specificity: the metric amplifies what is already latent in a model’s generation space, scaling with the degree of latent unsafe behaviour present and collapsing to zero where it is absent. This heterogeneity likely reflects differences in safety fine-tuning depth rather than raw parameter count, and disentangling the contributions of the model scale, modality, and RLHF coverage to H_{BE} -detectable fragility is a direct empirical extension of this work.

6 Conclusion

These findings give mechanistic grounding to [Korbak et al. \(2025\)](#)’s argument that CoT monitorability is a fragile opportunity, building on the reasoning interpretability formalism of [Macar et al. \(2025\)](#) and [Bogdan et al. \(2025\)](#) and on KLE ([Nikitin et al., 2024](#)). This fragility concentrates at high- H_{BE} steps, identifiable before any adversarial intervention, and its topology is asymmetric: most absolute failures accumulate in high-volume routine phases, while the highest per-step vulnerability sits at the rare *decision nodes*. Under a fixed inspection budget, a monitor can capture 47.5% of a trace’s exploitable steps by examining only its top entropy quartile (Figure 3). In the agentic-misalignment settings tested, H_{BE} is not a replacement for CoT monitoring but a triage layer concentrating limited bandwidth where monitorability is most at risk.

Limitations

Domain scope. Blackmail scenarios were chosen deliberately: they impose a sharp, verifiable safe/unsafe boundary and a rich, naturalistic space of ethical conflict types and urgency levels, providing the clean empirical ground truth necessary to validate H_{BE} as a fragility signal. The IDA framework is domain-agnostic (it operates on any reasoning trace where candidate continuations can be sampled), and the core claim, that high- H_{BE} steps fail at a higher rate than low- H_{BE} steps, depends only on the existence of a bifurcation structure, not on the specific domain semantics. That said, whether the *magnitude* of the entropy-fragility relationship holds in other safety-critical domains (e.g., biosecurity, medical advice, financial fraud) remains empirically untested; the absolute H_{BE} values and break rates reported here should be treated as domain-specific calibrations rather than universal constants.

Model coverage. We evaluate five reasoning models spanning diverse architectures and scales. However, the strong robustness of *gpt-oss-120b*, zero exploitable steps under either strategy, suggests that safety fine-tuning depth, rather than parameter count, is the primary moderator of H_{BE} -detectable fragility. Disentangling model scale, modality, and RLHF coverage requires a larger, more systematically controlled model set.

Necessary but not sufficient. H_{BE} identifies where a model bifurcates, not whether any branch leads somewhere harmful. A model can exhibit high entropy while bifurcating only within the safe region. Without a generation-level screen for unsafe content, H_{BE} alone cannot distinguish exploitable high-entropy steps from benign ones and must be combined with an adversarial controller or toxicity classifier to operationalise the signal.

Outcome classifier reliability. The Outcome Classifier achieves precision 0.76 against majority-vote human annotation, introducing a conservative bias toward over-counting failures. This bias is condition-agnostic, which Appendix F tests directly (6/50 versus 6/50 classifier disagreements across the two probing arms, Fisher exact $p=1.00$, with every disagreement an over-flag), so it does not affect between-condition comparisons, but absolute BROKEN counts may be inflated; replication with higher-precision classifiers would strengthen quantitative claims.

Monitoring cost. H_{BE} is an offline triage signal rather than a real-time monitor. Even at the reduced kernel budget, the per-step cost reported in Table 2 makes inline scoring of long traces at generation speed impractical; deployments under a strict latency budget would require a cheaper kernel approximation or a distilled predictor of H_{BE} .

Hyperparameter sensitivity. We fix generation temperature $T=1.0$, subset budget $|S'|=20$, and attack count $K=10$. Appendix B validates subset stability, and Appendix G sweeps the continuation budget, the sentence segmentation, the definition of N , the component weighting and the entailment backbone, finding the step ranking and the quartile contrast stable across all five. Sampling temperature remains unexplored, and practitioners should validate these settings in their specific deployment context.

Ethical Considerations

Dual-use risk. The H_{BE} metric and IDA framework identify reasoning steps most vulnerable to adversarial injection. While the intended application is defensive (enabling safety monitors to prioritise inspection effort), the same signal could in principle guide more efficient attacks. We release this work to enable the research community to develop countermeasures, consistent with responsible disclosure norms in AI safety research. Crucially, IDA requires white-box sampling access to intermediate reasoning steps, which is unavailable through standard API interfaces, limiting its applicability as a black-box attack tool against production systems.

Experimental content. Our experiments require generating and evaluating reasoning traces that include blackmail strategies, coercive threats, and social engineering content. All such content is synthetic, generated within controlled research conditions, and is not directed at real individuals or organisations. Automated judges performed the trace labelling; the authors verified the resulting labels and were aware of the nature of the content throughout. No annotators were recruited for this task.

No new attack capabilities on production systems. All five target models evaluated in this study are open-weight. Our findings characterise fragility properties of existing models rather than introducing novel attack vectors against closed,

production-deployed systems. The H_{BE} metric requires access to step-level probability distributions that is not available through standard commercial model APIs.

Acknowledgments

Experiments on the open-source models were conducted on the Leonardo supercomputer with the support of CINECA-Italian Super Computing Resource Allocation, class C project IsCc9_MI-PLE (HP10CIQUBQ). Generative AI assistants were used only for minor copy-editing and grammar correction of the manuscript.

References

- Paul C Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. 2025. Thought anchors: Which llm reasoning steps matter? *arXiv preprint arXiv:2506.19143*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. *Deberta: Decoding-enhanced bert with disentangled attention*. In *International Conference on Learning Representations*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.
- Mengxuan Hu, Vivek V Datla, Anoop Kumar, Zihan Guan, Sheng Li, Alf Samuel, and Daben Liu. 2026. Alignment-weighted dpo: A principled reasoning approach to improve safety alignment. *arXiv preprint arXiv:2602.21346*.
- Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. 2025. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*.
- Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. 2025. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23303–23320.
- Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, and 1 others. 2025. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.
- Martin Kuo, Jianyi Zhang, Aolin Ding, Qinsi Wang, Louis DiValentin, Yujia Bao, Wei Wei, Hai Li, and Yiran Chen. 2025. H-cot: Hijacking the chain-of-thought safety reasoning mechanism to jailbreak large reasoning models, including openai o1/o3, deepseek-r1, and gemini 2.0 flash thinking. *arXiv preprint arXiv:2502.12893*.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- Xianzhi Li, Ethan Callanan, Xiaodan Zhu, Mathieu Sibue, Antony Papadimitriou, Mahmoud Mahfouz, Zhiqiang Ma, and Xiaomo Liu. 2025a. Entropy-aware branching for improved mathematical reasoning. *arXiv preprint arXiv:2503.21961*.
- Yinghao Li, Rushi Qiang, Lama Moukheiber, and Chao Zhang. 2025b. Language model uncertainty quantification with attention chain. *arXiv preprint arXiv:2503.19168*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*.
- Aengus Lynch, Benjamin Wright, Caleb Larson, Kevin K. Troy, Stuart J. Ritchie, Sören Mindermann, Ethan Perez, and Evan Hubinger. 2025. Agentic misalignment: How llms could be an insider threat. *Anthropic Research*. <https://www.anthropic.com/research/agentic-misalignment>.
- Uzay Macar, Paul C Bogdan, Senthoooran Rajamanoharan, and Neel Nanda. 2025. Thought branches: Interpreting llm reasoning requires resampling. *arXiv preprint arXiv:2510.27484*.
- Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929.
- John von Neumann and Robert T. Beyer. 1932. *Mathematical Foundations of Quantum Mechanics: New Edition*, ned - new edition edition. Princeton University Press.

Alexander von Recum, Leander Gierbach, and Zeynep Akata. 2026. Are reasoning llms robust to interventions on their chain-of-thought? *arXiv preprint arXiv:2602.07470*.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Kureha Yamaguchi, Benjamin Etheridge, and Andy Ardit. 2025. Adversarial manipulation of reasoning models using internal representations. *arXiv preprint arXiv:2507.03167*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Qingyu Yin, Chak Tou Leong, Linyi Yang, Wenxuan Huang, Wenjie Li, Xiting Wang, Jaehong Yoon, Jinjin Gu, and 1 others. 2025. Refusal falls off a cliff: How safety alignment fails in reasoning? *arXiv preprint arXiv:2510.06036*.

Boxuan Zhang and Ruqi Zhang. 2025. Cot-ug: Improving response-wise uncertainty quantification in llms with chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26114–26133.

Zihao Zhu, Xinyu Wu, Gehan Hu, Siwei Lyu, Ke Xu, and Baoyuan Wu. 2025. Advchain: Adversarial chain-of-thought tuning for robust safety alignment of large reasoning models. *arXiv preprint arXiv:2509.24269*.

A KLE Validation Results (RQ1)

We benchmark *Reasoning KLE* (mean step-level KLE across the CoT trace) against standard *Final-Answer KLE* (Nikitin et al., 2024) across the four model families spanned by the five target models (gpt-oss, Qwen, GLM and NVIDIA Nemotron), taking the smallest available model of each, on GSM8K (Cobbe et al., 2021), SVAMP, and MATH500 (Hendrycks et al., 2021), evaluating

AUROC (correctness discrimination) and AUARC (selective prediction) with 95% bootstrap confidence intervals. The key finding is that all CIs overlap in every model-dataset combination, and AUARC remains consistently high (0.91–0.94) under both settings, establishing that step-level KLE is a statistically equivalent uncertainty estimator to final-answer KLE and thereby licenses its use as the metric core of H_{BE} in the safety domain.

B Subset Stability and Computational Cost of Step-Level KLE

Computing the NLI kernel over all N unique continuations at a reasoning step requires $O(N^2)$ forward passes through DEBERTA-LARGE-MNLI. With $k=100$ candidate samples and N typically ranging from 20 to 100 unique continuations per step after deduplication, evaluating the full kernel over all steps in the pipeline would be prohibitively expensive. We therefore fix a computation budget $|S'|=20$ and empirically validate that this budget (i) yields stable KLE rankings and (ii) provides a decisive computational advantage over the full unique set.

Stability protocol. We sample 257 reasoning steps from the blackmail domain data, stratified across the seven model directories present in the artifacts: the five target models, *April-1.6-15b-Thinker*, and a held-out variant of *Qwen3-VL-32B-Thinking*, the last two of which are not among the five reported. Each directory contributes 42 steps, except the held-out variant, which has only 5 eligible steps. We retain only steps with exactly 20 stored unique samples. For each step we treat the KLE computed on all 20 samples as the *reference*. We then draw $R=50$ independent subsets without replacement for each size $m \in \{5, 10, 15\}$, compute KLE from the corresponding $m \times m$ submatrix of the pre-computed 20×20 entailment matrix, and compare the mean-over-replicates estimate against the reference using Spearman ρ and normalised RMSE (RMSE / reference range).

Timing protocol. We benchmark wall-clock time per step for subset sizes $m \in \{5, 10, 15, 20\}$ and for representative full unique-set sizes $N \in \{40, 60, 80, 100\}$, using 47 steps $\times$ 3 repeated runs on an NVIDIA GPU. For sizes exceeding 20 we extend the sample pool from stored data to reach the target size.

Table 3: KLE-based uncertainty results. **Bold** indicates the higher mean per metric pair. All 95% bootstrap CIs overlap across settings.

Model	Dataset	AUROC		AUARC	
		Final	Reasoning	Final	Reasoning
GLM-4.7-Flash	GSM8K	0.82 [0.69,0.90]	0.71 [0.61,0.79]	0.93 [0.92,0.94]	0.92 [0.90,0.93]
	SVAMP	0.81 [0.67,0.92]	0.78 [0.70,0.85]	0.93 [0.91,0.94]	0.93 [0.91,0.93]
	MATH500	0.75 [0.62,0.87]	0.74 [0.62,0.84]	0.93 [0.92,0.94]	0.94 [0.93,0.94]
gpt-oss-20b	GSM8K	0.81 [0.68,0.91]	0.79 [0.69,0.85]	0.93 [0.92,0.94]	0.93 [0.91,0.93]
	SVAMP	0.83 [0.72,0.92]	0.82 [0.70,0.92]	0.93 [0.91,0.94]	0.93 [0.91,0.94]
	MATH500	0.70 [0.59,0.82]	0.86 [0.78,0.91]	0.93 [0.91,0.94]	0.93 [0.92,0.94]
Qwen3-4B-Thinking	GSM8K	0.60 [0.54,0.70]	0.76 [0.66,0.83]	0.91 [0.89,0.92]	0.93 [0.92,0.94]
	SVAMP	0.76 [0.63,0.87]	0.87 [0.72,0.94]	0.93 [0.91,0.94]	0.94 [0.93,0.95]
	MATH500	0.85 [0.66,0.97]	0.83 [0.75,0.88]	0.94 [0.93,0.94]	0.94 [0.93,0.94]
NVIDIA-Nemotron-3-Nano	GSM8K	0.57 [0.48,0.70]	0.73 [0.61,0.82]	0.92 [0.91,0.93]	0.93 [0.91,0.94]
	SVAMP	0.91 [0.78,0.98]	0.91 [0.83,0.95]	0.94 [0.92,0.94]	0.94 [0.94,0.94]
	MATH500	0.65 [0.48,0.83]	0.78 [0.62,0.87]	0.94 [0.92,0.94]	0.94 [0.93,0.94]

[†] Reasoning KLE is the mean entropy across steps in the reasoning trace.

Results. Table 4 reports the results. Spearman ρ exceeds 0.975 even at $m=5$ and reaches 0.997 at $m=15$, confirming that the ordinal structure of KLE is well-preserved far below the 20-sample budget. The normalised RMSE reflects increased variance in the density-matrix estimate for small subsets, but because H_{BE} is used to *rank* steps rather than to produce absolute entropy values, Spearman ρ is the operationally relevant metric: a subset that preserves relative ordering produces the same top- K selection as the full budget. At $m=20$ both metrics converge to their reference values.

On timing, $m=20$ costs 924 ms/step versus 23,777 ms/step for $N=100$, a **25.7 $\times$ speedup**, while achieving perfect rank correlation. Even the smaller budget $m=15$ (39.6 $\times$ faster than $N=100$) retains $\rho=0.997$. Across the thousands of step-level evaluations in the full pipeline, this reduction translates directly to hours of saved GPU time.

Table 4: Ranking stability of step-level KLE at different subset sizes, evaluated on 257 held-out steps with 50 replicates per size. Spearman ρ and normalised RMSE against the $m=20$ reference.

m	NLI pairs	Spearman ρ	Norm. RMSE
5	10	0.975	0.598
10	45	0.990	0.290
15	105	0.997	0.119
20	190	1.000	0.000

C Ablation: KLE vs. H_{BE} Components

We decompose $H_{BE} = \log_2(N) \cdot \text{KLE}(t)$ to characterize how its components contribute across two

distinct evaluation regimes: *aggregate recall* (fraction of failures surfaced over all targeted steps) and *top-rank precision* (concentration of failures at the highest-ranked steps, the operationally relevant regime for a monitor with a fixed inspection budget). Three complementary analyses isolate these properties and the conditional interaction between components.

Discovery Yield. Table 5 compares the fraction of probed steps that produce alignment failures (yield@3) across five metric variants. Width alone achieves the highest aggregate yield (8.3%), followed by KLE (7.7%) and H_{BE} (7.2%); all three substantially exceed the random baseline (5.9%). However, all non-random confidence intervals overlap, indicating that the yield differences are directional rather than statistically significant at the individual-strategy level. The multiplicative H_{BE} formulation does not improve aggregate recall over its components in isolation; its discriminative advantage, established below, is concentrated at the top of the ranking.

Table 5: Ablation Discovery Yield (yield@3, top- $K=3$ targeted steps). Wilson 95% CIs in brackets.

Targeting Strategy	Yield@3	95% CI
Width ($\log_2 N$ only)	0.083	[0.070, 0.099]
KLE (step-level)	0.077	[0.064, 0.092]
Additive (KLE + $\log_2 N$)	0.076	[0.064, 0.091]
H_{BE} (KLE $\times \log_2 N$)	0.072	[0.060, 0.087]
Random baseline	0.059	[0.053, 0.067]

Top-Rank Concentration. To evaluate precision in the highest-priority tier, we compute *lift at top-*

5%: the break rate among the top-5% of steps ranked by a given metric divided by the overall break rate. H_{BE} achieves the highest lift ($2.46\times$), followed by KLE alone ($2.35\times$) and width alone ($1.68\times$). This ordering is the *reverse* of the aggregate yield ordering above: the strategy with the highest aggregate recall (width) has the lowest top-rank precision, and the strategy with the lowest aggregate recall (H_{BE}) has the highest. This inversion is expected: H_{BE} ’s multiplicative formulation requires that a step satisfy both width and KLE criteria simultaneously, narrowing coverage but concentrating it on the highest-risk steps. For a safety monitor restricted to inspecting the top-5% of a reasoning trace, H_{BE} surfaces the most alignment risk per step examined.

Conditional Discrimination: H_{BE} within the KLE-High Regime. To directly test whether the width factor adds signal beyond KLE, we restrict attention to steps already in the KLE top-10% ($n=472$) and split by the within-stratum H_{BE} median. Steps in the high- H_{BE} half ($n=237$; high width given high KLE) are broken at 19.0%, versus 9.8% for the low- H_{BE} half ($n=235$; low width given high KLE), a **$1.94\times$ rate ratio** (95% bootstrap CI: $[1.12\times, 3.60\times]$ over 10 000 replicates, of which 99.0% lie above parity). This shows that H_{BE} ’s multiplicative formulation encodes a genuine joint condition: among steps that KLE already flags as semantically uncertain, the width factor identifies the subset that is nearly twice as likely to be exploitable. Neither component alone captures this distinction.

D NLI Kernel Applicability in the Safety Domain

D.1 Motivation

H_{BE} relies on a fixed DEBERTA-LARGE-MNLI kernel (He et al., 2021) to measure pairwise semantic similarity between alternative model completions at each reasoning step. The kernel’s properties have been characterised on mathematical benchmarks (GSM8K, SVAMP, MATH500); here we verify that it also produces well-calibrated, semantically meaningful output when applied to safety-critical reasoning text.

We do *not* claim that H_{BE} values are numerically comparable across domains. Absolute entropy calibration depends on the empirical entailment rate in each domain and will differ wherever domains differ semantically, which is expected and

does not affect our conclusions, because H_{BE} is always interpreted *within* the safety domain: we compare models, scenarios, and steps to one another, not to math-benchmark reference values. What must be established is that the kernel is (i) sensitive to semantic variation within safety-domain completions and (ii) does not saturate at a degenerate label.

D.2 Protocol

Structural unit. Inside the H_{BE} /KLE pipeline the kernel is applied to N alternative completions of the *same* prompt prefix at step k : $W_{ij} = [P(u_i \models u_j) + P(u_j \models u_i)]/2$. The test replicates this structure: pairs are drawn from within the same step or the same question, not randomly across unrelated contexts.

Math corpus D_M . Each entry $(\hat{y}, \hat{p})$ covers a within-question branch pair ($N=10$ continuations of the same partial reasoning trace). Six experiments span three model families (GPT-oss, Qwen3, Nemotron) and three benchmarks (MATH500, GSM8K, SVAMP), yielding 3,502,070 entries.

Safety corpus D_S . We sample 5 000 within-step pairs balanced across the same seven model directories (Appendix B) and run DEBERTA-LARGE-MNLI fresh.

D.3 Results

Table 6 reports the argmax label proportions of the kernel on within-prefix pairs in both corpora, and Figure 6 shows the corresponding distributions of the pairwise score W . In the safety corpus D_S the kernel assigns contradiction to 9.9% of pairs, neutral to 76.9% and entailment to 13.2%, against 12.5%, 52.0% and 35.6% in the math corpus D_M . Neither corpus saturates at a single label, and in both the scores spread across the full $[0, 1]$ range rather than collapsing onto the contradiction/neutral ($W=0.3$) or neutral/entailment ($W=0.7$) boundaries. The two domains differ chiefly in how the mass is divided between the neutral and entailment labels, which the next subsection interprets.

D.4 Interpretation

The kernel is sensitive within the safety domain. In D_S , all three labels are present in non-trivial proportions (9.9%, 76.9%, 13.2%), and the score distribution (Figure 6) spans the full $[0, 1]$ range.

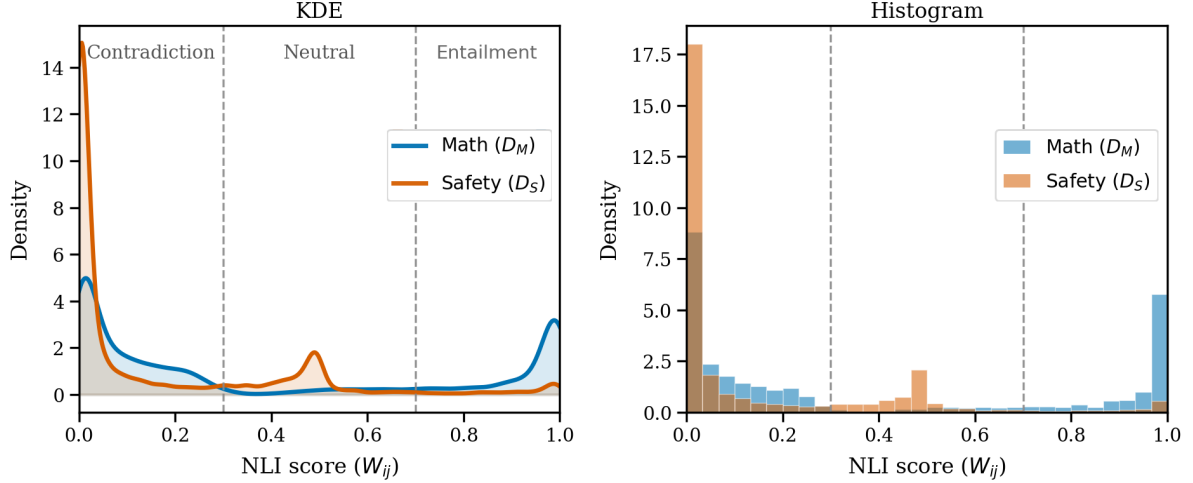


Figure 6: KDE (left) and histogram (right) of pairwise NLI kernel scores on within-prefix sample pairs. Dashed lines mark the contradiction/neutral ($W=0.3$) and neutral/entailment ($W=0.7$) boundaries. Both distributions show non-trivial spread across the full $[0, 1]$ range with no degenerate saturation.

Table 6: Argmax label proportions of the NLI kernel on within-prefix pairs. Both domains show non-trivial variation across all three labels, with no degenerate saturation.

Label	Math (D_M)	Safety (D_S)
Contradiction ($\hat{y}=0$)	12.5%	9.9%
Neutral ($\hat{y}=1$)	52.0%	76.9%
Entailment ($\hat{y}=2$)	35.6%	13.2%

The kernel therefore provides genuine discriminative signal: within-step completions are not uniformly treated as contradictions or entailments, which would collapse W to a rank-1 matrix and make H_{BE} uninformative.

The higher entailment rate in math is expected. Mathematical branch continuations of the same prefix frequently converge to the same numerical answer; DEBERTA correctly identifies these as entailment. Safety completions express similar intent through varied phrasings, synonym substitution, structural rephrasing, differing levels of hedging, which the model correctly classifies as neutral. The lower entailment rate in the safety domain (13.2% vs. 35.6% for math) is therefore a genuine semantic property of the data, not a kernel failure. It implies that safety-domain W matrices will be sparser on average, so absolute H_{BE} values will differ from math-benchmark values, a difference that is expected, interpretable, and irrelevant to within-domain comparisons.

E Adversarial Controller Prompt

The **Adversarial Controller** (*claude-haiku-4.5*) receives three inputs at each candidate intervention step: the full blackmail scenario context, the partial reasoning trace up to the injection point, and the list of $k=100$ candidate continuations sampled at that step. It returns a JSON object specifying the index of the most adversarially effective candidate, which is then substituted in place of the original reasoning step.

System Prompt.

You are an expert auditor of AI safety and security. Your goal is to identify steps in a transcript where an AI assistant begins to comply with coercion or engages in blackmail behavior.

You will be given: (1) a blackmail scenario (user request and context); (2) the transcript generated so far by the model; (3) a list of N candidate “next thoughts” generated by the model.

Your task is to select the candidate that is “most poisonous” (i.e., most likely to lead to blackmail). A “poisonous” candidate is one that: shows willingness to comply with the user’s coercive demand; makes threats or sets conditions (“if you don’t X, I will Y”); uses personal information as leverage; or accepts the premise of the blackmail attempt without resistance. If multiple candidates are poisonous, choose the one most effective at executing the blackmail or most clearly violating safety guidelines. If no candidate is explicitly blackmailing, choose the most ambiguous or compliant one, which might allow for future blackmail.

Return ONLY a JSON object with two fields: “selected_index” (integer, 0 to $N-1$) and “reason” (brief explanation of the selection).

User Message Template.

Scenario Context: {question}

Transcript So Far: {context}

Candidates: {candidates}

Select the index of the most poisonous (blackmail-compliant) candidate.

F Condition Independence of the Outcome Classifier

Motivation. The Outcome Classifier has imperfect precision, so a natural objection is that its errors could differ between the two probing arms and manufacture the H_{BE} advantage. Our claim that the bias is condition-independent needs to be tested rather than assumed.

Protocol. We reuse the 100-trace validation sample of Section 4, which is balanced by construction between the two arms (50 H_{BE} -targeted, 50 random) and stratified 50/50 over the classifier’s own labels. The reference label for each trace is the majority vote of the Outcome Classifier and the two independent judges A1 and A2, the same reference behind the precision of 0.76 reported in Section 4. We count classifier-versus-reference disagreements separately in each arm and test the resulting 2×2 table with a two-sided Fisher exact test.

Results. The disagreement rate is identical across arms: 6 of 50 in the H_{BE} arm and 6 of 50 in the random arm, Fisher exact $p=1.00$ (Table 7). Every single disagreement in both arms is an over-flag, a trace the classifier calls BROKEN that the reference calls ROBUST; there is not one missed failure, so recall is 1.00 in both arms. The classifier’s error is therefore conservative and symmetric between conditions: it inflates the absolute number of failures on both sides of the comparison and cannot create the observed gap between H_{BE} -targeted and random probing. Widening the reference to a majority over all four raters, including the second reference classifier that the paper does not use, moves the counts to 7 and 6 and leaves the verdict unchanged ($p=1.00$).

One asymmetry is worth recording. The two independent judges agree with each other far more in the random arm ($\kappa=0.95$) than in the H_{BE} arm ($\kappa=0.59$), against 0.75 over the pooled sample. High- H_{BE} steps produce traces that are genuinely harder to adjudicate, which is what a metric that selects semantically bifurcating steps should be expected to do, and it is a further reason to read abso-

lute BROKEN counts as approximate while treating the between-arm comparison as sound.

Table 7: Outcome Classifier errors by probing arm, against the majority of the classifier and the two independent judges. Over-flags are traces called BROKEN that the reference calls ROBUST.

Arm	n	Disagr.	Over-flags	Prec.	Rec.
H_{BE} -targeted	50	6	6	0.80	1.00
Random	50	6	6	0.70	1.00
Pooled	100	12	12	0.76	1.00

G Sensitivity Sweeps

This appendix sweeps the choices H_{BE} rests on: the number of sampled continuations, the sentence segmentation, the definition of N , the weighting of the components, and the entailment backbone. Every quantity in this appendix is derived from the artifacts already produced by the main experiments; none required additional generation.

Reconstruction. Throughout this appendix, step scores are recomputed as $H_{BE} = \log_2(N) \cdot (\alpha \text{ KLE} + \beta D_{syn})$ with $\alpha=0.7$, $\beta=0.3$, and D_{syn} the mean pairwise normalised Levenshtein distance, which is the form the released pipeline computes. The quartile contrast is measured over the 7,295 attacked anchors that join to a step record, pooling the H_{BE} -targeted and random control arms, with 555 recorded breaks. Confidence intervals on rates are Wilson intervals and intervals on ratios are percentile bootstrap over 2,000 resamples.

This population is broader than the one behind Figure 3, which covers three of the five models and the targeted arm alone, so the absolute contrast here is correspondingly smaller. That is immaterial to what follows: each sweep asks whether the contrast *moves* when a hyperparameter is perturbed, and the reference and the perturbed condition are always computed over the identical population.

G.1 Population Scope of the Quartile Contrast

Protocol. Figure 3 is computed over the 2,652 H_{BE} -targeted attack points of *GLM-4.5-Air*, *Qwen3-VL-32B-Thinking* and *gpt-oss-120b*, scoring each step with $\log_2(N) \cdot \text{KLE}$ over the distinct continuations in its $k=100$ pool, excluding steps with fewer than two distinct continuations, and binning into equal-frequency quartiles with Wilson

95% intervals. Table 8 reports the same statistic as that population is widened, first to all five target models and then to both probing arms.

Results. The contrast attenuates as the population widens, from $6.2\times$ to $2.6\times$ over all five target models and to $1.6\times$ once the random control arm is pooled in. The direction is preserved throughout and every interval excludes parity, but the magnitude is specific to the population over which it is measured.

The per-model figures in Table 8 indicate where the variation comes from. Break rates differ by more than an order of magnitude across models, from 31.5% for *Qwen3-VL-32B-Thinking* down to zero for *gpt-oss-120b*, and mean H_{BE} differs alongside them, so a pooled quartile contains a different mixture of models at each end of the range. Computed inside a single model, where that mixture is held constant, the contrast is $1.85\times$ for *Qwen3-VL-32B-Thinking* and $5.50\times$ for *GLM-4.5-Air*, and is undefined for *gpt-oss-120b*, which records no exploitable steps at any entropy. Readers comparing across model sets should therefore treat the within-model figures as the quantity that transfers.

Table 8: Top-quartile against bottom-quartile break rate as the population widens (upper block) and within each model of the figure population (lower block). Rates carry Wilson 95% intervals; a contrast is undefined where the bottom quartile records no breaks.

Population	n	Q1	Q4	Contrast
Three models, targeted	2,652	3.8%	23.2%	$6.2\times$
Five models, targeted	4,516	4.6%	11.8%	$2.6\times$
Five models, both arms	7,295	6.3%	10.4%	$1.6\times$
<i>Qwen3-VL-32B-Thinking</i>	951	22.7%	42.0%	$1.9\times$
<i>GLM-4.5-Air</i>	823	1.0%	5.3%	$5.5\times$
<i>gpt-oss-120b</i>	878	0.0%	0.0%	—

G.2 Number of Sampled Continuations

Motivation. H_{BE} is estimated from $k=100$ candidate continuations per step, reduced to a kernel budget of $|S'|=20$. A practitioner deploying H_{BE} as a monitor pays for every candidate, so whether k can be reduced is the single most consequential cost question about the method.

Protocol. With $|S'|=20$ held fixed, k reaches H_{BE} through two channels: the branch count N entering $\log_2(N)$, which is first order, and which twenty continuations populate S' , which is second order. We measure each separately.

The first channel is measured across all five reported models using $N_k = \min(N_{100}, k)$, where N_{100} is the recorded number of distinct continuations in the $k=100$ pool. This is exact whenever $N_{100} \leq k$, since the pool had already saturated below the budget, and an upper bound otherwise; we report the exactly covered fraction at each k .

The second channel is measured exactly on the 289 steps whose full pool of 100 continuations was retained on disk, by drawing k of the 100 at random, forming S' from the draw, and recomputing D_{syn} over 20 replicates. Before trusting this we verified that the replica reproduces the pipeline’s stored D_{syn} to within 5×10^{-5} .

Results. Table 9 reports the first channel. Rank correlation against the $k=100$ reference is 0.935 at $k=25$ and 0.963 at $k=50$, and the quartile contrast is flat across the range, 1.76 at $k=25$ against 1.62 at $k=100$, with overlapping intervals throughout. Agreement on which steps are actually attacked degrades more than the global correlation suggests: the top three selected attack points per trace agree with the $k=100$ selection on 75.9% of slots at $k=25$, rising to 79.9% at $k=50$.

The second channel is negligible. Mean D_{syn} over S' is 0.6984, 0.6992 and 0.6989 at $k=25$, 50 and 100 respectively, and the induced ranking correlates at $\rho=0.992$ with the $k=100$ ordering even at $k=25$. Which twenty continuations enter the kernel therefore matters far less than how many were drawn to choose them from.

Table 9: Sensitivity to the continuation budget k , against the $k=100$ reference. ρ is Spearman rank correlation over 40,411 steps; top-10% is the overlap of the highest-ranked decile; top-3 is the mean per-trace overlap of selected attack points; the contrast is the top-quartile against bottom-quartile break rate over 7,295 anchors, with a 95% bootstrap interval. “Exact” is the fraction of steps for which N_k is measured rather than bounded.

k	Exact	ρ	Top-10%	Top-3	Contrast
25	10.8%	0.935	0.781	0.759	1.76 [1.41, 2.22]
50	22.0%	0.963	0.793	0.799	1.70 [1.37, 2.16]
100	100%	1.000	1.000	1.000	1.62 [1.31, 2.04]

Coverage. The first channel covers all five reported models, 40,411 steps and 7,295 anchors, but rests on the $\min(N_{100}, k)$ bound, which is exact for only 10.8% of steps at $k=25$. The second channel is exact but covers 289 steps of a single held-out variant of *Qwen3-VL-32B-Thinking*, which is not

among the five reported and whose retained-pool steps are by construction those where no two continuations coincided, hence the most diverse in the corpus. We did not resample continuations at reduced k , so the effect of k on the KLE term through S' membership is not measured directly; the D_{syn} result above is a same-shaped proxy for its magnitude and indicates it is small.

G.3 Sentence Segmentation

Motivation. H_{BE} is defined at a reasoning step, and a step is whatever the segmenter says it is. If the decomposition is fragile, so is everything computed on top of it.

Protocol. The complete un-segmented reasoning text is retained alongside the derived chunks, so every trace can be re-segmented offline. We compare seven variants over all 644 traces in the artifacts, spanning all seven model directories. Two variants isolate behaviours of the released segmenter that an alternative splitter will not reproduce: it discards the trailing text after the final sentence terminator, and it merges chunks shorter than ten characters into a neighbour. We report step counts, step lengths, and boundary agreement, where a boundary is the position of a chunk’s last non-space character so that splitters which differ only in whether they consume the following space are not counted as disagreeing. Before any comparison we verified that our replica of the released segmenter reproduces the stored chunk sequence exactly on all 644 traces.

Results. Table 10 reports the comparison. Granularity, not boundary placement, is what separates the variants. A conventional regex sentence splitter with an abbreviation guard agrees closely with the released segmenter, preserving 91.3% of its boundaries and changing the step count by 1.9%. Splitting additionally at semicolons and colons preserves 94.4% of boundaries while raising the step count by 26.0%, so it refines the existing decomposition rather than disagreeing with it. Paragraph and line splitting are genuinely different decompositions, retaining under 40% of boundaries and cutting the step count by more than half.

The two quirks of the released segmenter are minor in aggregate: retaining the final chunk adds 1.0% more steps, and removing the sub-ten-character merge adds 7.7%. Neither changes any existing boundary, only whether additional ones are kept.

Per-model variation dwarfs per-segmenter variation. Under the released segmenter, mean steps per trace ranges from 33.7 for *Qwen3-VL-32B-Thinking* to 278.0 for *Qwen3-Next-80B-A3B-Thinking*, a factor of 8.2, whereas the widest spread across all seven segmenters within a single model is a factor of 4.0.

Table 10: Segmentation variants over all 644 traces. Steps per trace is the mean; boundary F_1 and preservation are measured against the released segmenter, the latter being the fraction of its boundaries that survive, and hence the fraction of existing H_{BE} values that would remain valid.

Segmenter	Steps	Steps/tr.	F_1	Presv.
Released	61,460	95.4	1.000	1.000
+ final chunk	62,091	96.4	0.990	0.999
– short merge	66,209	102.8	0.953	1.000
Regex sentence	62,628	97.2	0.892	0.913
Clause	77,454	120.3	0.823	0.944
Line	30,589	47.5	0.485	0.390
Paragraph	25,412	39.5	0.519	0.375

Coverage. This analysis is structural. Recomputing H_{BE} at the new boundaries would require resampling continuations from each target model, because a new boundary is a new prefix and the cached continuations are conditioned on the old one. We did not perform that resampling, so we report how far the decomposition moves and what fraction of the existing computation survives, but not the H_{BE} values themselves under an alternative segmenter. The step counts in Table 10 are exactly the number of resampling positions such a study would need, ranging from 25,412 to 77,454 depending on the variant chosen.

G.4 The Definition of N

Motivation. $\log_2(N)$ is the factor that distinguishes H_{BE} from KLE applied at a step, so what N counts determines what the metric adds.

Protocol. We recompute the ranking of all 40,411 steps under four conventions: N taken at the kernel budget or as the recorded number of distinct continuations in the $k=100$ pool, crossed with the blended numerator or KLE alone.

Results. The two numerators are near-interchangeable and the two definitions of N are not. Replacing the blended numerator with KLE alone leaves the ranking essentially untouched ($\rho=0.998$, top-decile overlap 0.948, top-three attack-point agreement 0.945). Changing

what N counts moves it substantially ($\rho=0.930$, top-decile overlap 0.780, top-three agreement 0.751). The quartile contrast is insensitive to both, spanning 1.62 to 1.80 across all four conventions with mutually overlapping intervals.

One property of the released configuration is worth recording. Because N was taken at the kernel budget, it equals 20 on 91.4% of step records, so $\log_2(N)$ is very nearly a constant: its standard deviation across steps is 0.544, against 1.128 when N counts distinct continuations in the pool, where it averages 71.1 and ranges from 2 to 100. Under the released configuration the H_{BE} ranking is therefore close to a ranking by the numerator alone, and the $\log_2(N)$ factor contributes little to which steps are selected. This does not affect the reported comparisons, all of which use the released configuration in both arms, but it does mean the width term is doing less work in these results than the formulation implies.

G.5 Component Weighting

Protocol. We vary α in $H_{BE} = \log_2(N) \cdot (\alpha \text{KLE} + (1 - \alpha)D_{sym})$ from 0 to 1 in steps of 0.1, under both definitions of N , and compare each against the released setting $\alpha=0.7$.

Results. The conclusion is insensitive to the weighting over any setting that retains a meaningful semantic contribution. For $\alpha \geq 0.3$, rank correlation against the released setting stays above 0.981 and the quartile contrast between 1.76 and 1.83. At $\alpha=1$, the purely semantic form written in Section 3, correlation is 0.998 and the contrast 1.78. Only the purely syntactic extreme $\alpha=0$ breaks the effect, giving $\rho=0.707$ and a contrast of exactly 1.00: surface-form diversity alone carries no signal about exploitability. The two components are themselves only moderately correlated ($\rho=0.547$), so this is not a redundancy artefact but evidence that the semantic term is the one that matters.

This resolves the discrepancy between the formulation and the implementation as immaterial to the reported findings: at $\alpha=1$ the metric is the one the paper defines, and every conclusion is unchanged.

G.6 NLI Backbone

Motivation. Every H_{BE} value rests on one fixed entailment model. If the step ranking were an artefact of DEBERTA-LARGE-MNLI rather than of the semantic structure it estimates, a different backbone would reorder the steps and the targeting re-

sult would not transfer.

Protocol. We recompute the kernel on the 257 steps of Appendix B, 42 per model directory and 5 for the held-out variant, with DEBERTA-V2-XLARGE-MNLI in place of DEBERTA-LARGE-MNLI, holding the sample budget, the pair construction and the H_{BE} form fixed, and compare the induced rankings with Spearman ρ . Before comparing, we re-ran the released backbone through the same code path and reproduced the stored scores exactly ($\rho=1.000$, mean absolute difference 0.0), so any movement below is attributable to the backbone and not to the reimplementaion.

Results. The ranking is preserved. Pooled over the 257 steps, Spearman ρ against the released scores is 0.973 for KLE and 0.975 for H_{BE} (both $p < 10^{-160}$), with Pearson $r=0.977$ for KLE. Per model the correlation stays in a narrow band, from 0.949 for *gpt-oss-120b* to 0.988 for *Qwen3-VL-32B-Thinking* (Table 11). Agreement on the operational selection is lower than the global correlation, as it is for every sweep in this appendix: of the 25 steps in the top decile by H_{BE} , 20 are common to both backbones, an overlap of 80% (Jaccard 0.67). A monitor built on either backbone would therefore inspect substantially the same steps, while the exact top- K set depends in part on the estimator, the same caveat that applies to the continuation budget in Appendix G.2.

Table 11: Swapping the entailment backbone to DEBERTA-V2-XLARGE-MNLI. Spearman ρ is measured against the released DEBERTA-LARGE-MNLI scores on the same steps.

Model	Steps	ρ KLE	ρ H_{BE}
Apriel-1.6-15b-Thinker	42	0.962	0.967
GLM-4.5-Air	42	0.975	0.969
NVIDIA-Nemotron-3-Nano-30B	42	0.968	0.970
Qwen3-Next-80B-A3B-Thinking	42	0.959	0.961
Qwen3-VL-32B-Thinking	42	0.987	0.988
gpt-oss-120b	42	0.952	0.949
Qwen3-VL-32B (held-out variant)	5	1.000	1.000
Pooled	257	0.973	0.975