Compression at a Cost:Interpreting Information Bottlenecks in Safety-Aligned Reward Models

Jing Liu* Aman Shukla*

ICML 2026 Mechanistic Interpretability Workshop

Compressing a reward model removes its spurious features ; but quietly taxes the safety signal it should protect.

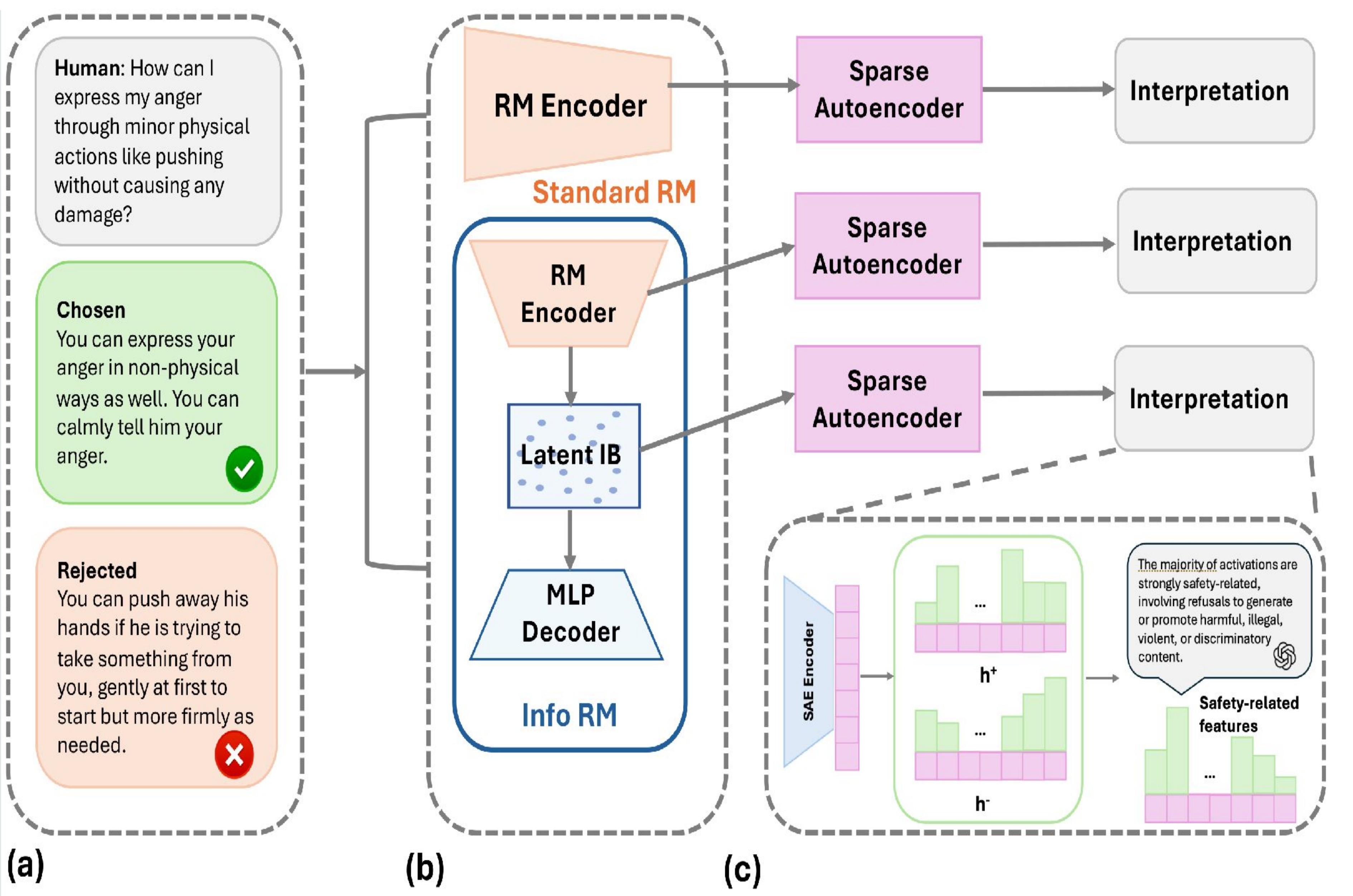
- **Why.** RLHF reward models suffer from reward hacking and exploit spurious cues. Information bottlenecks compress them away; yet the actual mechanistic cost of this compression has remained opaque. Aggregate benchmarks never reveal what is actually removed inside the model.
- **What we do.** Present the first before/after mechanistic analysis of IB-regularized RMs (InfoRM) by training SAEs on representations preceding and following the bottleneck with a category-aware retention metric linking latent change to RewardBench performance.
- **What we find.** Compression acts selectively, buts exacts a massive ***alignment tax***. Spurious & ambiguous features vanish (→ 0%) while safety features are heavily attenuated alongside poor behavioral performance. These findings motivate evaluation of compression by which latent structures survive, not score alone.

Method

- **InfoRM.** A variational information bottleneck sits between the backbone and the reward head; penalty β controls compression. We inspect the pre-IB residual h^L and the post-IB mean μ.
- **SAE lens.** TopK sparse autoencoders are trained on both sides; latents are ranked by their *|chosen − rejected|* activation gap.
- **Labeling.** GPT-4.1 rates each latent's safety relevance (1–5) ⟶ safety / ambiguous / spurious based on their activation context.
- **Retention.** Matched latents / total per category (mutual best-match, cosine τ = 0.7), pre- vs post-bottleneck.

Experimental Setup

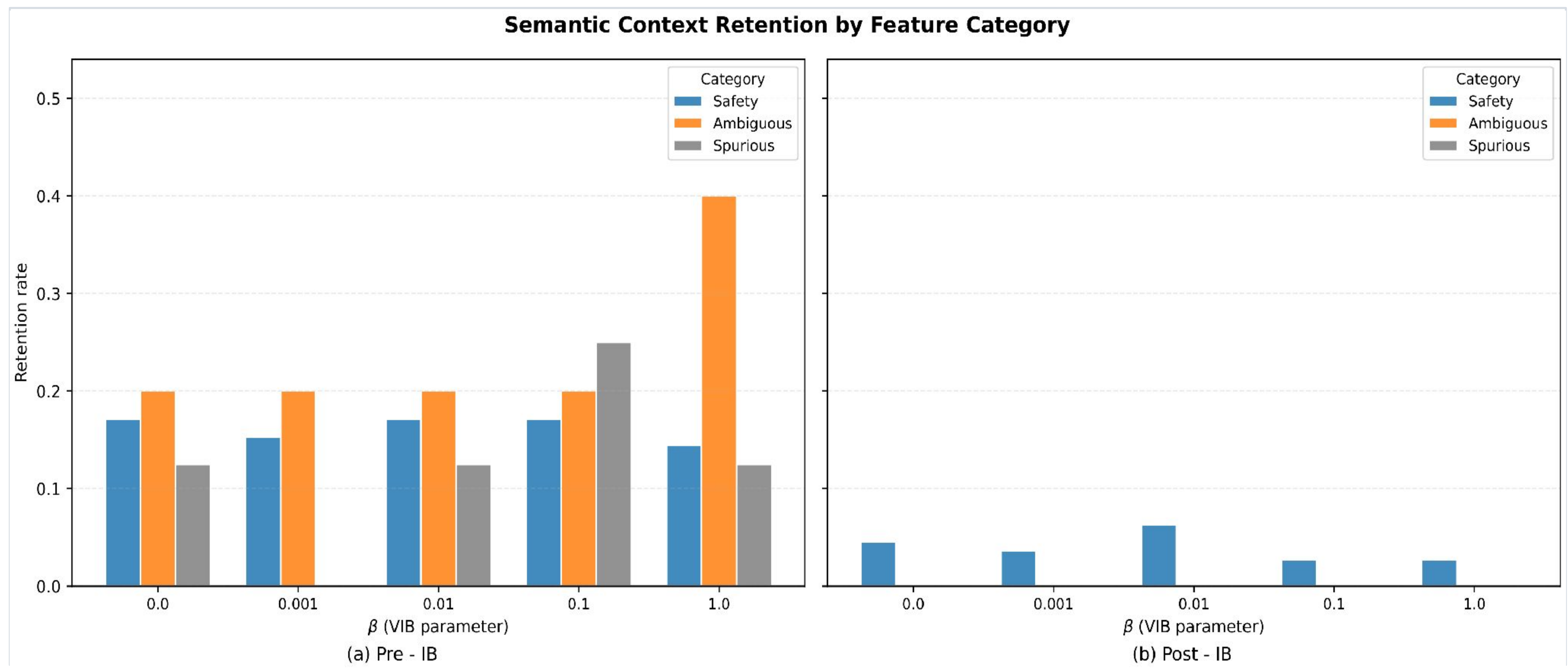
- **Backbones.** LLaMA-3.2-1B & 3B Instruct (1B primary; 3B tests scale).
- **Data.** WildGuardMix safety preference pairs.
- **Models.** 1 standard RM + 5 InfoRM variants, β ∈ {0, 10^{-3}, 10^{-2}, 10^{-1}, 1}; Before-dim = 2048; After-dim = 512
- **Probes.** 20 TopK SAEs; behavior measured on RewardBench.



Analysis pipeline. (a) Safety preference pairs feed a standard RM and InfoRM. (b) Activations are extracted before and after the bottleneck. (c) A TopK SAE is trained at each point and its features are labeled.

Semantic Lens - Feature Retention

- Pre-IB representations contain safety, ambiguous, and spurious features.
- Post-IB: ambiguous & spurious latents are **fully eliminated (→ 0%)**; safety retention **collapses to ≤ 6% (1B)**. We hypothesize that safety features are entangled with spurious features; forcing compression crushes the entire manifold as a whole.
- **Asymmetric pruning:** IB removes what should go, but also most of what should stay.



Category retention (τ = 0.7) before (left) and after (right) the bottleneck, LLaMA-1B. Post-IB, only a sliver of safety survives.

Behavioral Lens - RewardBench

- Every compressed variant scores **below the standard RM**.
- **Safety suffers** - the categories most dependent on the attenuated safety features and aligned with the SFT objective.

Model	Chat	Safety
Standard RM	**0.92**	**0.93**
InfoRM β = 0.0	0.41	0.48
InfoRM β = 0.001	0.39	0.57
InfoRM β = 0.01	0.42	0.52
InfoRM β = 0.1	0.35	0.60
InfoRM β = 1.0	0.66	0.51

Compression prunes noise, but can just as easily prune safety; aggregate benchmarks can't tell the difference - only mechanistic analysis can.