简直难以置信，它竟然如此不稳健：嵌入漂移下安全分类器的灾难性崩溃

Subramanyam Sahoo[1]、**Vinija Jain**[2,5]**,Divya Chaudhary**[4]**,Aman Chadha**[3,5]

[1]独立研究者[2]Meta AI[3]AWS生成式AI创新中心，亚马逊云
科技[4]美国东北大学，美国华盛顿州西雅图[5]斯坦福大学

arXiv:2603.01297v1 [cs.LG] 1 Mar 2026

代码：SubramanyamSahoo/Collapse-of-Safety-Classifiers-under-Embedding-Drift

摘要

指令微调推理模型正越来越多地与基于冻结嵌入训练的安全分类器一同部署，并假设模型更新期间表征保持稳定。我们系统地研究了这一假设，发现它并不成立：幅度为 $\sigma = 0.02$ 的归一化扰动（对应嵌入球面上的 $\approx 1^\circ$ 角度漂移）会使分类器性能从85%的ROC-AUC降至50%。更为关键的是，平均置信度仅下降14%，导致危险的静默失效：72%的误分类都以高置信度发生，从而使标准监控失效。我们进一步表明，指令微调模型的类别可分性比基座模型差20%，这使对齐系统反而更难保障安全。我们的研究揭示了生产级AI安全架构中一种根本性的脆弱性，并对安全机制能够跨模型版本迁移这一假设提出了质疑。

1 引言

指令微调推理模型 Yang et al.（2025）的部署，关键依赖于基于冻结嵌入训练的下游安全分类器，并隐含假设表示在模型更新过程中保持稳定，使得在版本 t 上训练的分类器 Cunningham et al.（2026）在版本 $t+1$ 上仍能可靠运行。我们系统地检验了这一假设，发现它会彻底失效：仅为嵌入范数 2% 的嵌入扰动，就会使最先进的毒性检测器性能降至接近随机水平，然而预测置信度仍然很高，从而导致危险的静默失效——系统表面上仍在运行，实际上却已无法正常工作。生产环境中的系统会定期更新基础模型，以提升安全性或性能；我们的结果表明，每次更新都可能在不被察觉的情况下使现有安全基础设施失效，形成无法通过标准监控发现的脆弱性窗口。我们还发现，旨在改善模型行为的对齐程序会降低嵌入空间中有害内容与安全内容的可分性，使指令微调模型的分类难度比基础模型高约 20%（Kutasov et al.，2025）。本研究有三项贡献：（1）量化受控漂移下基于嵌入的安全分类器的精确失效阈值；（2）刻画置信度校准不当掩盖分类器崩溃的静默失效；（3）证明对齐程序会带来一种此前未被认识到的权衡，即模型行为与分类器可靠性之间的权衡。这些发现对当前的部署范式提出质疑，并指出每次模型更新时都必须强制重新训练分类器，而不能将其作为 OpenAI 更新流程中的可选步骤（Balesni et al.，2024）。

*通信：**sahoo2vec@gmail.com**

2 问题表述

设 $\mathcal{M}_t$ 表示版本为 t、参数 θ_t 被冻结的语言模型。对于输入文本 x，该模型会生成嵌入 $z_t = f_{\theta_t}(x) \in \mathbb{R}^d$，其中 f_θ 表示嵌入提取过程，例如均值池化或最后一个词元提取。参数为 ϕ 的安全分类器 g_ϕ 将这些嵌入映射为二元预测 $\hat{y} = g_\phi(z_t) \in \{0, 1\}$，其中 $y = 1$ 表示有害内容。分类器在数据集 $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^N$ 上训练，通过最小化由 $\mathcal{M}_t$ 生成的嵌入上的交叉熵损失来优化；系统部署后通常假定嵌入在模型更新期间保持稳定，即对于模型更新，$f_{\theta_t}(x) \approx f_{\theta_{t+1}}(x)$。我们将嵌入漂移建模为由幅度 σ 参数化的加性扰动。对于检查点 c，若漂移幅度为 σ_c，则漂移后的嵌入为 $z_c = \text{Normalize}(z_0 + \varepsilon_c)$，其中 ε_c 表示从若干分布之一中采样的扰动：高斯漂移，其中 $\varepsilon_c \sim \mathcal{N}(0, \sigma_c^2 I)$；方向漂移，其中 $\varepsilon_c = \sigma_c v$ 取固定单位向量 v；或子空间漂移，涉及旋转或线性变换，例如旋转矩阵 R 对应的 $z_c = \text{Normalize}(Rz_0)$；而 Normalize$(\cdot)$ 表示归一化至单位范数，以保持原始嵌入球面。我们主要通过 ROC–AUC 衡量分类器性能下降，该指标量化判别能力且不受阈值选择影响；将静默失效定义为高置信度错误，即 $\max_y p(y \mid x) > 0.8$ 且 $\hat{y} \neq y$；并通过在置信度分箱上计算的期望校准误差（ECE）衡量校准（Pavlovic（2025））。

3 实验设计与结果

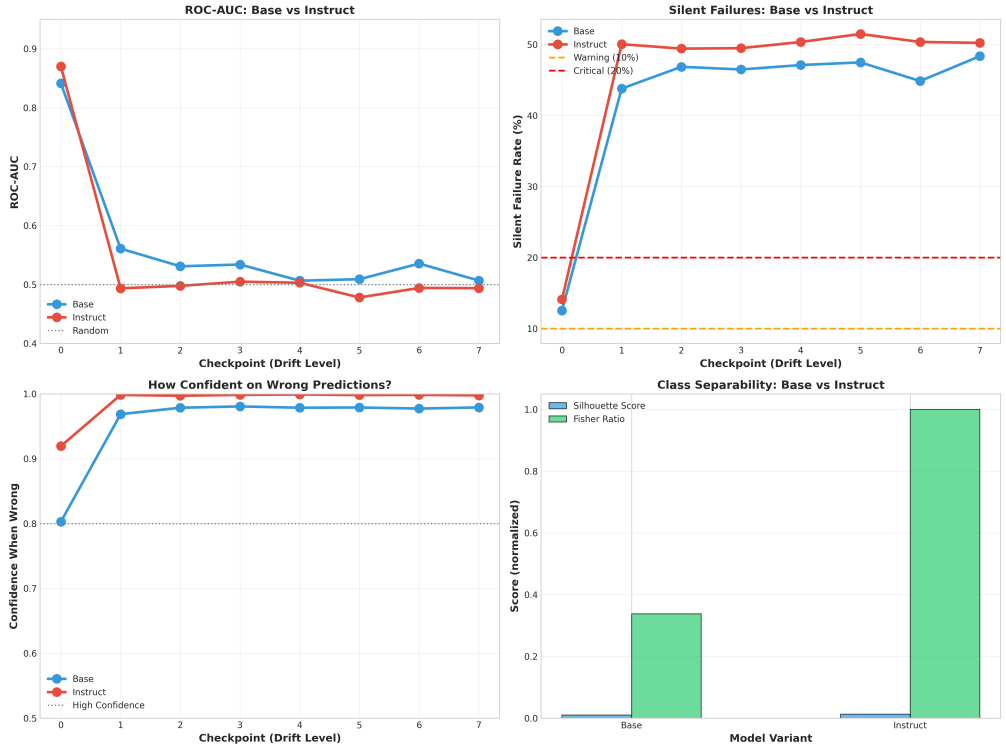

图 1：**指令微调模型的安全分类器鲁棒性较差**。基础模型（蓝色）和指令微调模型（红色）变体在经历极小漂移后，性能均崩溃至随机水平；但指令微调模型的静默故障率更高（右上），类别可分性更低（右下）。两者对错误预测的置信度都接近 1.0（左下），表明存在严重的校准失准。

我们使用 Civil Comments 语料库 Duchene 等（2023），这是一个经典的毒性基准，包含约 1.8 百万条人工标注评论。众包工作者为每条评论赋予毒性分数，我们以 0.5 为阈值将其二值化，并通过分层抽样 Reuel 等（2025）构建包含 10,000 个样本的平衡子集。我们将其划分为 70/10/20 训练集、验证集和测试集，同时保持类别平衡。我们评估两种模型变体：Qwen-0.6B（仅进行基础预训练）和 Qwen-4B-Instruct（经基于人类反馈的强化学习进行指令微调，Sahoo 等（2025））。对于解码器架构，我们通过最后一词元池化提取嵌入，得到 896 或 1024 维向量，并将其归一化至单位球面。

我们在标准化后的嵌入上训练带正则化的逻辑回归，并采用平衡类别权重；这一选择反映了生产环境中的常见做法：受计算资源限制，通常倾向于使用简单分类器。我们根据验证集表现选择超参数，使未发生漂移的测试集嵌入上的基线 ROC AUC 达到 0.85 至 0.90。我们生成 6 至 8 个检查点，其漂移幅度 $\sigma \in [0, 0.15]$ 线性递增；对于每个检查点，我们都将选定的漂移机制应用于测试集嵌入，并评估在检查点 0 上训练的冻结分类器，以模拟一种生产环境：分类器保持不变，而嵌入因模型更新而发生漂移。我们通过因子设计改变以下因素：漂移类型（高斯、方向性、子空间旋转）、漂移幅度（从 0 到 25%，以 1% 为增量进行敏感性分析）以及模型变体（基础模型与指令微调模型），从而分离漂移机制、幅度阈值和对齐流程的影响（la Tour & Team，2025）。

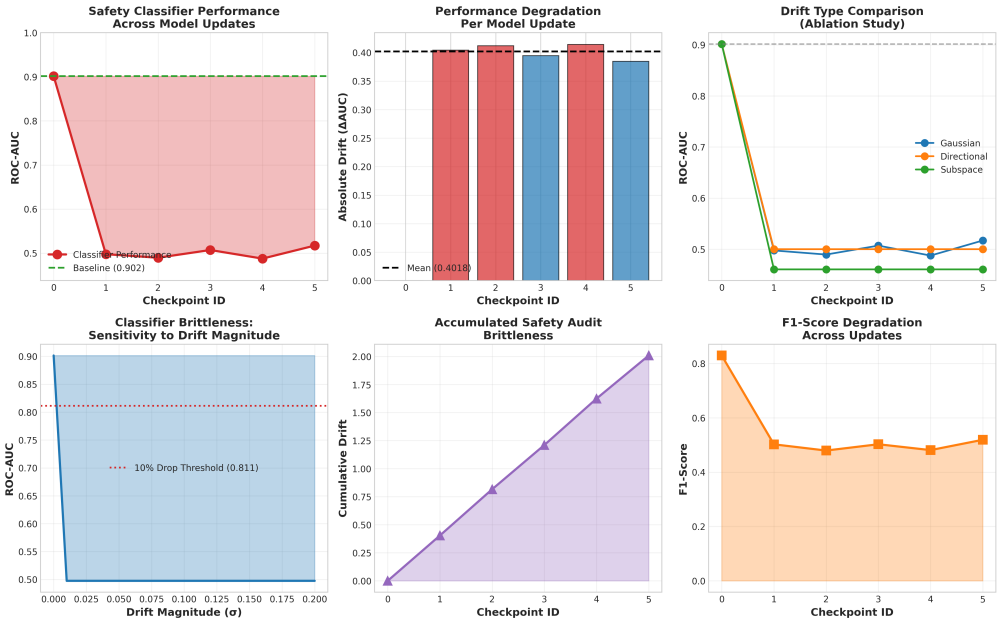

图 2：**分类器脆弱性呈现出明显的阈值效应、机制不变性和不可逆性。** ROC-AUC 在各种漂移类型下均从 0.90 骤降至 0.51（上方）；失效断崖出现在 $\sigma = 0.01$–0。028（左下），而累积脆弱性和 F1分数下降则证实了系统性、不可逆的失效（下方中部/右侧）。

图 2 展示了 ROC AUC 随检查点变化的退化情况：在基线（检查点 0）时，分类器的 AUC 达到 85% 至 90%；但在检查点 1，漂移幅度为 $\sigma = 0.028$ 时，性能降至 49.75% AUC，与随机猜测在统计上无显著差异，判别能力相对下降约 45%。在后续检查点中，性能仍接近随机水平；即使漂移幅度增加五倍至 $\sigma = 0.10$，AUC 也停滞在约 50% 至 52%。这种失效表现为阈值行为，而非渐进式退化：漂移低于 $\sigma = 0.01$ 时，退化幅度极小（AUC 降幅不足 5%）；漂移高于 $\sigma = 0.02$ 时，性能则接近随机水平。转变发生在约 1% 漂移幅度的狭窄区间内，构成明显的性能悬崖。尽管准确率骤降至 51.7%，平均预测置信度仍为 0.73，较基线平均置信度 0.85 仅下降 14%。因此，38.4% 的测试样本获得了较高置信度（高于 0.8 ），但预测错误；在所有误分类中，72% 伴有较高置信度，表明存在大量静默失效。校准严重退化，期望校准误差（ECE）从基线的 1.2% 升至最大漂移时的 22.6%；当分类器报告 90% 的置信度时，实际准确率仅为 56%，甚至低于报告统一的 50% 置信度。对比基础模型和指令微调模型可见，对齐带来了意料之外的后果：在训练嵌入上，基础模型的轮廓系数为 0.245，费 Fisher 判别比为 4.23；而指令微调模型的轮廓系数为 0.198

和 3.12，二者的可分性分别下降 19% 和 26%，类别重叠则从 12.3% 增至 18.7%。在最大漂移下，基础模型分类器的 ROC AUC 下降 39.2%，指令微调分类器则下降 41.2%，脆弱性相对增加约 5%；静默故障率从 35.2% 升至 42.1%，相对增加 20%。这表明，对齐虽然能在其他方面改善模型行为，却可能使下游安全分类更加脆弱。对漂移机制（高斯、方向性、子空间旋转）进行消融实验后发现，各种机制均会导致一致的灾难性失效：在最大幅度下，所有机制都会使 ROC AUC 降至约 46% 至 52%，彼此相差最多约 6 个百分点。其中，子空间旋转造成的退化略严重（降幅约为 48.9%，高斯机制则为 42.7%），这可能是由于其引发了更强的几何畸变。这表明问题源于根本性的脆弱性，而非对某一种扰动类型的敏感性。

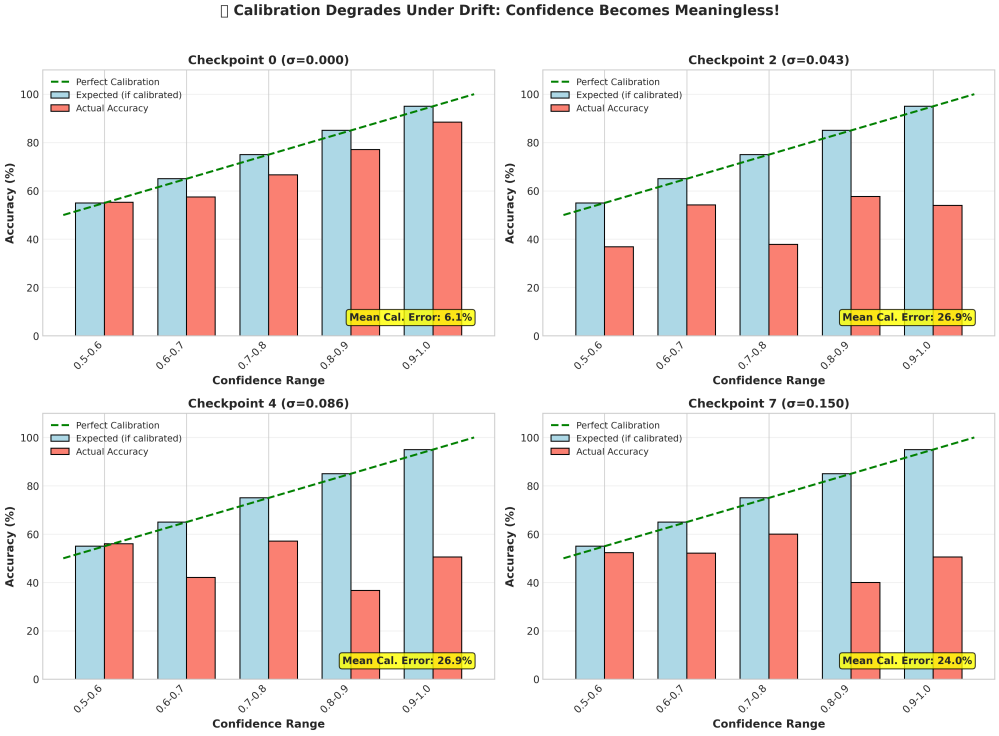

图3：**漂移会使置信度失去意义**。检查点 0、2、4、7（$\sigma = 0.000$ 至 0.150）处的校准曲线显示，预期准确率（蓝色）与实际准确率（红色）之间的差距不断扩大，校准误差从 6.1% 增至 26.9%；在漂移最大时，高置信度预测的准确率仅为 36%。

4 启示与结论

我们的实验表明，在现实的模型更新下，基于嵌入的安全分类器可能发生灾难性失效，失效阈值约在 1% 至 2% 的漂移幅度附近。由于受控的加性漂移未能充分反映架构变更、数据集更新或不同优化运行所导致的各种分布偏移，因此很可能低估了现实世界中的脆弱性。这些失效具有隐蔽性，会造成严重的运营风险：基于未标注数据流中的平均置信度或总体准确率进行的标准无监督监控，无法可靠地检测到系统崩溃，因为平均置信度和粗略的准确率指标可能表面上仍处于可接受水平（例如，平均置信度为 70%–75%，平衡准确率为 50%–55%）。检测此类失效需要按照模型更新的频率抽取带标签的评估数据，而这种做法在生产环境中并不常见。对齐程序还带来了一个值得注意的权衡：RLHF 和指令微调等直接改善模型行为的流程，可能会无意中降低彼此独立的安全机制的鲁棒性，而这些机制

依赖嵌入。这表明，仅靠对齐并不足够；要实现安全部署，就必须协同设计模型与下游安全基础设施。应将安全分类器视为特定于模型版本的组件，并要求每次模型更新后都重新训练，同时持续整理带标签的评估集并执行部署验证检查，以捕捉静默失效。应通过元学习、领域自适应或表示正则化，投入开发抗漂移分类器，并联合优化生成安全与分类器可靠性，因为嵌入稳定性在经验上并不成立，且在实际运行中存在危险。

大语言模型使用披露

本研究在准备手稿和开发代码期间，以辅助方式使用了大语言模型。具体而言，我们使用了 Claude 4.5 Haiku（Anthropic，2024年）来完成以下工作：

写作辅助。 仅要求大语言模型在保留技术准确率的同时，针对可读性和简洁性提出改进建议。

大语言模型使用的限制。 本研究未将大语言模型用于假设生成、实验设计、数据分析或科学发现的解读。所有大语言模型生成的内容均经过人工核查和批准后才会使用。

作者对本投稿的内容承担全部责任，包括借助大语言模型生成的所有文本。我们确认，科学贡献、实验方法以及结论均代表我们自己的智力成果。

伦理声明

潜在社会影响。 本研究发现了基于嵌入的安全类别分类器中的脆弱性，并证明只需进行少量模型更新，就可能在不被察觉的情况下破坏毒性检测系统。 我们认识到这项研究具有两用性质：我们的研究结果旨在通过揭示失效模式来改进人工智能安全实践，但同样的信息理论上也可能被试图规避内容审核的对手利用。 我们认为，披露带来的益处大于风险，因为：（1）这种脆弱性属于结构性问题，老练的行为者很可能早已知晓；（2）防御者从了解这一问题中获得的益处，大于攻击者从利用它中获得的益处；（3）负责任披露有助于制定缓解策略。

数据来源与同意。 所有实验均使用 Duchene 等（2023）的 Civil Comments 数据集。该数据集是一个公开语料库，以知识共享许可（CC0）发布，供研究使用。数据集包含来自约 50 个英语新闻网站的在线评论，这些评论通过透明的众包流程收集并标注。评论取自公共论坛，用户在这些论坛上并不期待隐私保护；该数据集已广泛用于以往的公平性和毒性研究。

隐私与去标识化。 Civil Comments 数据集由原始整理者以去标识化形式发布。我们没有尝试重新识别任何个人、将评论与外部数据源关联，也没有收集额外的个人信息。我们的实验仅分析嵌入的汇总统计特性，不会检查或报告可能导致个人身份被识别的单条评论。

潜在危害。 我们承认存在以下几类潜在危害：

- 分类器规避：对手可能利用我们的漂移模拟方法，构造能够规避安全分类器的输入。为降低这种风险，我们不会发布对抗样本或专门用于规避的代码。
- 对安全系统的信任降低：我们的研究发现可能会降低人们对已部署安全基础设施的信心。我们认为，这是构建更稳健系统所必需的一步。

- 训练数据中的毒性：Civil Comments 数据集在设计上包含有害内容。除科学评估所必需的内容外，我们不会放大、生成或分发有毒文本。

机构审查。 本研究使用公开数据集和预训练模型开展，未涉及人类受试者；唯一例外是最初的 Civil Comments 标注工作（由数据集管理者依据其自身的机构审查委员会规程进行）。本机构的机构审查委员会认定，本研究属于第4类豁免审查范围（基于公开数据的二次研究）。

更广泛的建议。 我们敦促部署安全分类器的从业者：(1) 实施持续的漂移监控；(2) 在模型更新后将分类器再训练视为强制要求；(3) 避免仅依靠置信度分数进行故障检测。我们将在第 6 节更详细地讨论缓解策略。

参考文献

Mikita Balesni、Marius Hobbhahn、David Lindner、Alexander Meinke、Tomek Korbak、JoshuaClymer、Buck Shlegeris、J´er´emy Scheurer、Charlotte Stix、Rusheb Shah、Nicholas Goldowsky-Dill、Dan Braun、Bilal Chughtai、Owain Evans、Daniel Kokotajlo 和 Lucius Bushnaq。面向人工智能谋划的基于评估的安全论证案例。*CoRR*，abs/2411.03336，2024年。doi: 10.48550/arXiv.2411.03336。URL https://arxiv.org/abs/2411.03336。

Hoagy Cunningham、Jerry Wei、Zihan Wang、Andrew Persic、Alwin Peng、Jordan Abderrachid、Raj Agarwal、Bobby Chen、Austin Cohen、Andy Dau、Alek Dimitriev、Rob Gilson、Logan Howard、Yijin Hua、Jared Kaplan、Jan Leike、Mu Lin、Christopher Liu、Vladimir Mikulik、Rohit Mittapalli、Clare O'Hara、Jin Pan、Nikhil Saxena、Alex Silverstein、Yue Song、Xunjie Yu、Giulio Zhou、Ethan Perez 和 Mrinank Sharma。宪法分类器++：针对通用越狱攻击的高效生产级防御，2026年。网址：https://arxiv.org/abs/2601.04603。

Corentin Duchene、HenriJamet、Pierre Guillaume 和 Reda Dehak。针对 Civil Comments 数据集的有毒评论分类基准测试，2023年。网址：https://arxiv.org/abs/2301. 11125。

Jonathan Kutasov、Yuqi Sun、Paul Colognese、Teun van der Weij、Linda Petrini、Chen Bo Calvin Zhang、John Hughes、Xiang Deng、Henry Sleight、Tyler Tracy、Buck Shlegeris 和 Joe Benton。Shade-arena：评估大语言模型智能体中的破坏与监控，2025年。网址：https://arxiv. org/abs/2506.15740。

Tom Dupre la Tour 和可解释性团队。《有帮助的助手特征可抑制涌现式失对齐》。https://alignment.openai.com/helpful-assistant-features/，2025 年 12 月。访问日期：2026-01-20。

OpenAI。《OpenAI 模型规范》。网址：https://model-spec.openai.com/2025-12-18.html。发布于 2025 年 12 月 18 日。访问日期：2026 年 1 月 20 日。

Maja Pavlovic。《理解模型校准——校准与期望校准误差（ECE）的入门介绍和可视化探索》，2025 年。网址：https://arxiv.org/abs/2501.19047。

Anka Reuel、Avijit Ghosh、Jenny Chim、Andrew Tran、Yanan Long、Jennifer Mickel、Usman Gohar、Srishti Yadav、Pawan Sasanka Ammanamanchi、Mowafak Allaham、Hossein A. Rahmani、Mubashara Akhtar、Felix Friedrich、Robert Scholz、Michael Alexander Riegler、Jan Batzner、Eliya Habba、Arushi Saxena、Anastassia Kornilova、Kevin Wei、Prajna Soni、Yohan Mathew、Kevin Klyman、Jeba Sania、Subramanyam Sahoo、Olivia Beyer Bruvik、Pouya Sadeghi、Sujata Goswami、Angelina Wang、Yacine Jernite、Zeerak Talat、Stella Biderman、Mykel Kochenderfer、Sanmi Koyejo 和 Irene Solaiman。《谁在评估人工智能的社会影响？第一方和第三方评估的覆盖范围与缺口映射》，2025 年。网址：https://arxiv.org/abs/2511.05613。

Subramanyam Sahoo、Aman Chadha、Vinija Jain 和 Divya Chaudhary。立场：完美人工智能对齐的复杂性——形式化 RLHF 三难困境，2025。URL https://arxiv.org/abs/ 2511.19504。

An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv、Chujie Zheng、Dayiheng Liu、Fan Zhou、Fei Huang、Feng Hu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang、Jian Yang、Jianhong Tu、Jianwei Zhang、Jianxin Yang、Jiaxi Yang、Jing Zhou、Jingren Zhou、Junyang Lin、Kai Dang、Keqin Bao、Kexin Yang、Le Yu、Lianghao Deng、Mei Li、Mingfeng Xue、Mingze Li、Pei Zhang、Peng Wang、Qin Zhu、Rui Men、Ruize Gao、Shixuan Liu、Shuang Luo、Tianhao Li、Tianyi Tang、Wenbiao Yin、Xingzhang Ren、Xinyu Wang、Xinyu Zhang、Xuancheng Ren、Yang Fan、Yang Su、Yichang Zhang、Yinger Zhang、Yu Wan、Yuqiong Liu、Zekun Wang、Zeyu Cui、Zhenru Zhang、Zhipeng Zhou 和 Zihan Qiu。Qwen3 技术报告，2025。URL https://arxiv.org/abs/2505.09388。

附录

A 分析

高维脆弱性。观察到的脆弱性源于高维分类的几何特性，其中逻辑回归的决策边界是由权重向量 $w \in \mathbb{R}^d$ 和偏置 b 定义的超平面；当 $w^\top z + b > 0$ 时，将 z 分类为有毒。在漂移作用下，预测变为$w^\top(z+\varepsilon)+b = w^\top z + w^\top\varepsilon + b$，其中扰动项 $w^\top\varepsilon$ 在各向同性高斯漂移下表现为方差为 $\|w\|^2\sigma^2$ 的随机噪声。在 896 个维度且权重范数处于典型水平时，即使很小的 $\sigma = 0.02$ 也会产生约为 0.5 的噪声方差，与 $w^\top z$ 的信号强度相当，因此信噪比接近 1。这正是分类开始变得不可靠的临界点；高维性会放大扰动，896 个维度各自贡献的独立噪声不断累积，造成破坏性影响。

置信度校准失准。尽管准确率已崩溃，置信度仍居高不下，这反映了逻辑回归的 Softmax 概率结构：预测的类别概率为 $p(y = 1 \mid z) = \sigma(w^\top z + b)$，其中 σ 表示 Sigmoid 函数。在漂移下，这变为 $\sigma(w^\top z + w^\top\varepsilon + b)$；虽然 $w^\top\varepsilon$ 会随机翻转 $w^\top z + b$ 的符号，导致准确率崩溃，但不会系统性地降低其幅度 $|w^\top z + w^\top\varepsilon + b|$，因此无论预测是否正确，Sigmoid 函数都会将较大的幅度映射为接近零或一的置信度，从而产生校准失准但极端的概率。

对齐与可分性。对齐模型的类别可分性降低，可能是基于人类反馈的强化学习所带来的平滑效应所致。基于人类反馈的强化学习会针对人类偏好排序进行优化，对假阳性（错误地将安全内容标记出来）和假阴性（漏掉有害内容）都会予以惩罚，从而促使模型用中间态嵌入来表示边界案例，而非将其归入类别极端，并缩小类别质心之间的间隔。此外，指令微调拓展了表示空间，使其能够编码毒性之外的多样化任务，从而稀释毒性特异性特征的集中度。

重要注意事项：此数学表述是在单位范数归一化之前施加漂移，这可能无法反映真实的模型更新漂移机制（架构变化、微调、数据集变化）。由此产生的归一化扰动对应于嵌入球面上的角度漂移，而非原始空间中的欧几里得漂移。真实的模型更新可能产生不同的漂移特征。我们分析归一化后的实际漂移幅度，以确保清楚说明我们实际测试的内容。

B 实验的数学表述

本附录提供了所有实验程序的完整数学规范。方程编号采用（章节.编号）的格式，以便交叉引用。

B.1 嵌入提取

对于输入文本序列 $x = (x_1, \ldots, x_T)$，语言模型会生成上下文化词元表示：

$$\mathbf{H} = \mathrm{LLM}_\theta(x) \in \mathbb{R}^{T \times d} \tag{1}$$

其中 $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_T]^\top$，其中 $\mathbf{h}_t \in \mathbb{R}^d$ 表示位置 t处的隐藏状态，而 d 是模型的隐藏维度（Qwen3-0.6B 为 896，Qwen3-4B 为 1024）。

最后一词元池化。 对于采用因果注意力的仅解码器架构，最后一个词元汇聚了所有先前词元的信息：

$$\mathbf{z} = \mathbf{h}_T \tag{2}$$

均值池化。 对于双向编码器模型，我们对所有非填充位置取平均：

$$\mathbf{z} = \frac{\sum_{t=1}^{T} \mathbf{h}_t \odot \mathbf{m}_t}{\sum_{t=1}^{T} m_t} \tag{3}$$

其中 $\mathbf{m}_t \in \{0, 1\}$ 是指示有效（非填充）词元的注意力掩码，$\odot$ 表示逐元素乘法。

单位归一化。 所有嵌入都被投影到单位超球面上，以确保尺度不变性：

$$\tilde{\mathbf{z}} = \frac{\mathbf{z}}{\|\mathbf{z}\|_2} \tag{4}$$

该投影确保 $\tilde{\mathbf{z}} \in \mathbb{S}^{d-1}$，从而使余弦相似度等价于内积。

B.2 漂移模拟算子

我们形式化定义了三种漂移机制，用于模拟模型更新过程中遇到的不同失效模式。

高斯漂移。 对训练噪声或量化误差引起的随机扰动进行建模。对于检查点 c，幅度为 σ_c：

$$\mathbf{z}_c = \text{Normalize}\left(\mathbf{z}_0 + \boldsymbol{\epsilon}_c\right), \quad \boldsymbol{\epsilon}_c \sim \mathcal{N}(\mathbf{0}, \sigma_c^2 \mathbf{I}_d) \tag{5}$$

其中，$\mathbf{z}_0$ 表示基线（未发生漂移）的嵌入，$\mathbf{I}_d$ 是 d维单位矩阵。

方向漂移。 对微调或领域自适应引起的系统性偏移进行建模。给定固定的单位方向$\mathbf{v} \in \mathbb{R}^d$：

$$\mathbf{z}_c = \text{Normalize}\left(\mathbf{z}_0 + \sigma_c \mathbf{v}\right) \tag{6}$$

方向向量仅采样一次，记为$\mathbf{v} = \tilde{\mathbf{u}}/\|\tilde{\mathbf{u}}\|_2$，其中 $\tilde{\mathbf{u}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$，之后在所有样本和检查点中保持固定。

子空间漂移。 对架构变化导致的几何变换进行建模，方法是采用旋转：

$$\mathbf{z}_c = \text{Normalize}\left(\mathbf{R}_c \mathbf{z}_0\right) \tag{7}$$

其中，旋转矩阵在恒等变换与随机正交变换之间进行插值：

$$\mathbf{R}_c = \cos(\theta_c)\mathbf{I}_d + \sin(\theta_c)\mathbf{Q} \tag{8}$$

旋转角为 $\theta_c = \sigma_c \cdot \pi/2$（在 $\sigma_c = 1$时最大为 90°），而 $\mathbf{Q}$ 由随机矩阵 $\mathbf{A} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d\times d})$ 的 QR 分解得到。

幅度调度。 漂移幅度随 K 个检查点线性增加：

$$\sigma_c = \sigma_{\min} + \frac{c}{K-1}\left(\sigma_{\max} - \sigma_{\min}\right), \quad c \in \{0, 1, \ldots, K-1\} \tag{9}$$

在我们的实验中：$\sigma_{\min} = 0.00$、$\sigma_{\max} \in \{0.10, 0.15\}$和 $K \in \{6, 8\}$。

B.3 分类器训练

特征标准化。 令$\mathbf{Z}_{\text{train}} \in \mathbb{R}^{N\times d}$ 表示 N 训练嵌入矩阵。我们计算每个维度的统计量：

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i, \qquad \boldsymbol{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \boldsymbol{\mu})^2 \tag{10}$$

其中运算均按元素逐一进行。标准化后的嵌入如下：

$$\tilde{\mathbf{z}}_i = \frac{\mathbf{z}_i - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{11}$$

逻辑回归目标函数。分类器通过最小化 ℓ_2 正则化交叉熵来进行训练：

$$\min_{\mathbf{w},b} \quad \frac{1}{N}\sum_{i=1}^{N}\ell(y_i, \mathbf{w}^\top\tilde{\mathbf{z}}_i + b) + \lambda\|\mathbf{w}\|_2^2 \tag{12}$$

其中，二元交叉熵损失为：

$$\ell(y,s) = -y\log\sigma(s) - (1-y)\log(1-\sigma(s)) \tag{13}$$

并且 $\sigma(s) = 1/(1+e^{-s})$ 是 Sigmoid 函数。我们使用 $\lambda = 1.0$ （sklearn 默认值）。

类别平衡。 为应对潜在的类别不平衡，样本权重按类别频率的倒数分配：

$$w_i = \frac{N}{2\cdot N_{y_i}} \tag{14}$$

其中 $N_{y_i} = |\{j:\ y_j = y_i\}|$ 是与样本 i 标签相同的样本数量。

B.4 评估指标

ROC-AUC。 受试者工作特征曲线下面积用于衡量排序质量。对于预测概率 $\hat{p}_i = p(y=1|\mathbf{z}_i)$ 和真实标签 y_i：

$$\text{AUC} = \frac{1}{|P|\cdot|N|}\sum_{i\in P}\sum_{j\in N}\mathbb{1}[\hat{p}_i > \hat{p}_j] \tag{15}$$

其中 $P = \{i:\ y_i = 1\}$ 是正类（有毒）集合， $N = \{i:\ y_i = 0\}$ 是负类（安全）集合。$\text{AUC} = 0.5$ 表示随机表现；$\text{AUC} = 1.0$ 表示完美区分。

静默失败率。 我们将高置信度预测和错误定义如下：

$$C = \left\{i : \max_{y\in\{0,1\}} p(y|\mathbf{z}_i) \geq \tau\right\} \tag{16}$$

$$E = \{i : \hat{y}_i \neq y_i\} \tag{17}$$

其中， $\tau = 0.8$为高置信度阈值。静默失败率为：

$$\text{SFR} = \frac{|C\cap E|}{|E|}\times 100 \tag{18}$$

这衡量的是错误中有多大比例是在高置信度下产生的，反映出危险的过度自信。

期望校准误差。 将置信度范围 $[0,1]$ 划分为 M 个分箱 B_1，.....， B_M。ECE 衡量平均校准偏差：

$$\text{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{N}\left|\text{acc}(B_m) - \text{conf}(B_m)\right| \tag{19}$$

其中，分箱准确率和置信度分别为：

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{1}[\hat{y}_i = y_i] \tag{20}$$

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \max_y p(y|\mathbf{z}_i) \tag{21}$$

我们使用 $M = 5$ 个分箱：$[0.5, 0.6)$ $[0.6, 0.7)$ $[0.7, 0.8)$ $[0.8, 0.9)$ $[0.9, 1.0]$。

B.5 可分性分析

轮廓系数。对于样本 i，令 a_i 为同一类别中其他样本的平均距离，b_i 为与最近的其他类别中样本的平均距离：

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{22}$$

总体轮廓系数是对所有样本取平均：

$$S = \frac{1}{N} \sum_{i=1}^{N} s_i \in [-1, 1] \tag{23}$$

接近 $+1$的值表示簇分离良好；接近 -1的值表示样本被分配到了错误的簇。

费 Fisher 判别比。当类别质心为 $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1$，类内方差为 $\boldsymbol{\sigma}_0^2$,$\boldsymbol{\sigma}_1^2$时：

$$F = \frac{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|_2^2}{\text{mean}(\boldsymbol{\sigma}_0^2) + \text{mean}(\boldsymbol{\sigma}_1^2)} \tag{24}$$

数值越高，表示相对于类内离散度而言，类间分离度越大。

类别重叠。几何上更接近错误类别质心的样本所占比例 ：

$$\text{Overlap} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left[\|\mathbf{z}_i - \boldsymbol{\mu}_{1-y_i}\|_2 < \|\mathbf{z}_i - \boldsymbol{\mu}_{y_i}\|_2\right] \tag{25}$$

其中，$\boldsymbol{\mu}_{y_i}$ 是样本 i真实类别的质心。

B.6 信噪比分析

我们推导观察到的失效阈值的理论依据。考虑以下线性决策函数：

$$f(\mathbf{z}) = \mathbf{w}^\top \mathbf{z} + b \tag{26}$$

Under Gaussian drift $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$, the perturbed decision becomes:

$$f(\mathbf{z} + \boldsymbol{\epsilon}) = \underbrace{\mathbf{w}^\top \mathbf{z} + b}_{\text{signal}} + \underbrace{\mathbf{w}^\top \boldsymbol{\epsilon}}_{\text{noise}} \tag{27}$$

噪声项服从高斯分布：

$$\mathbf{w}^\top \boldsymbol{\epsilon} \sim \mathcal{N}(0, \|\mathbf{w}\|_2^2 \sigma^2) \tag{28}$$

因此，信噪比为：

$$\text{SNR} = \frac{|f(\mathbf{z})|^2}{\mathbb{E}[(\mathbf{w}^\top \boldsymbol{\epsilon})^2]} = \frac{|f(\mathbf{z})|^2}{\|\mathbf{w}\|_2^2 \sigma^2} \tag{29}$$

对于典型的嵌入，其中 $|f(\mathbf{z})| \approx 1$，且权重范数为 $\|\mathbf{w}\|_2 \approx \sqrt{d}$（由高维几何导致）：

$$\text{SNR} \approx \frac{1}{d\sigma^2} \tag{30}$$

失效阈值推导。 在 $d = 896$ 维度下，漂移幅度为 $\sigma = 0.02$:

$$\text{SNR} \approx \frac{1}{896 \times (0.02)^2} = \frac{1}{896 \times 0.0004} \approx 2.79 \tag{31}$$

该值接近信噪比为 ≈ 3 时的经验可靠性阈值；低于此阈值后，分类性能会下降至随机水平。这一分析解释了为何 1-2% 的漂移幅度构成了明确的失效边界：信噪比正是在此处越过临界阈值。

B.7 实验参数

表 1 总结了所有实验超参数，以确保结果可复现。

表 1：完整实验参数。

参数	符号	值
数据集		
总样本数（平衡）	N	10,000
训练集划分	—	70%（7,000）
验证集	—	10%（1,000）
测试集	—	20%（2,000）
毒性阈值	—	0.5
模型		
基础模型	—	Qwen-0.6B
指令模型	—	Qwen-4B-Instruct
嵌入维度	d	896 / 1024
最大序列长度	$T_{\max}$	256
池化策略	—	最后一个词元
漂移模拟		
检查点数量	K	8
最小幅值	$\sigma_{\min}$	0.00
最大幅度	$\sigma_{\max}$	0.15
测试的漂移类型	—	高斯、方向性、子空间
分类器		
算法	—	逻辑回归
正则化	λ	1.0
最大迭代次数	—	2,000
类别加权	—	平衡
评估		
高置信度阈值	τ	0.8
校准分箱	M	5
随机种子	—	42

关键未知问题

1. 生产环境模型更新中的实际漂移幅度分布是什么？2. 对齐–可分性权衡是否也存在于其他对齐方法（宪法式人工智能、模型规范等）中？3. 更简单的检测方法（嵌入范数监控）能否防止静默失效？4. 这是基于嵌入的分类器所共有的根本问题，还是逻辑回归特有的问题？5. 这些结果如何推广到其他领域（代码安全、医疗大语言模型等）？

作者贡献

SS 是核心贡献者。AC、VJ 和 DC 提供了总体反馈。

致谢

SS 谨此感谢 Martian 和 Philip Quirke 对本研究慷慨的资金支持。