路由即对齐：混合专家语言模型在对抗性与结构性扰动下的安全脆弱性

Jinxu Fang（方锦旭）

摘要

混合专家（MoE）架构已成为前沿语言模型的主流范式，但其基于路由的条件计算对安全性的影响尚未被系统研究。本文首次在同一模型家族内，对 MoE 模型与匹配的稠密模型进行了受控的对齐鲁棒性比较（Qwen3-30B-A3B，3B 活跃参数 vs. Qwen3-4B），控制了训练数据、对齐流程和活跃参数量。研究得出三个挑战传统直觉的发现：**(1)** MoE 模型对越狱攻击的鲁棒性优于稠密模型（模板攻击 ASR 0.069 vs. 0.177；PAIR 攻击 ASR 0.200 vs. 0.440），尽管每个 token 激活的参数更少。**(2)** 有害与良性输入的路由模式存在系统性分歧（48 层平均 JSD = 0.194），有害提示引发更集中的专家选择（熵 5.67 vs. 6.08 bits），但安全性并未局部化在可辨识的"安全专家"中。**(3)** 将学习到的路由替换为随机专家选择会导致灾难性的对齐失效——拒绝率从 100% 骤降至 8.4%±5.7%——脆弱性分数（变异系数）为 0.680。这些结果表明，MoE 模型的对齐编码在路由函数中，而非个体专家参数中，揭示了稀疏架构需要路由感知的对齐训练策略。

1 引言

混合专家（MoE）架构已成为扩展前沿语言模型的主导范式。DeepSeek-V3 [3]、Mixtral [5]、Qwen3-MoE [9]，以及据报道的 GPT-4，均将每个 token 路由到一小部分专家模块，以较低的推理成本实现了更大规模稠密网络的表征能力。随着这些模型被部署到安全关键应用中，一个根本性问题随之产生：*MoE 架构的条件化、基于路由的计算方式，是否引入了与稠密模型质性不同的对齐脆弱性？*

现有的对齐技术——RLHF [8]、DPO [10] 等——是面向稠密架构设计的，其中梯度更新均匀地触及所有参数。在 MoE 模型中，梯度仅通过路由器选定的专家传播，可能导致出现频率较低的专家组合保留未对齐前的行为。此外，操纵路由决策的对抗性输入可能完全绕过对齐——这是稠密模型中不存在的攻击面。尽管该问题至关重要，此前尚无工作提供受控的 MoE/稠密对比研究；现有工作 [13, 14] 仅提出缓解方案而缺乏匹配基线。

我们通过围绕三个研究问题的系统性研究来填补这一空白：

RQ1. 越狱鲁棒性。在控制训练数据、对齐流程和活跃参数量的条件下，MoE 和稠密模型对黑盒对抗攻击的脆弱性是否存在差异？

RQ2. 路由与安全。专家路由模式与安全行为之间有何关系？安全相关的计算是局部化在特定专家中，还是分布在路由函数中？

RQ3. 对齐脆弱性。MoE 模型的对齐对路由决策的扰动有多敏感？我们能否量化这种脆弱性？

贡献。 本文的主要贡献如下：

1. **首个匹配对 MoE/稠密安全性比较**。在 Qwen3 家族上进行受控实验（Qwen3-30B-A3B MoE, 3B 活跃/30B 总参 vs. Qwen3-4B 稠密, 4B），揭示 MoE 模型对模板攻击（ASR = 0.069 vs. 0.177）和 PAIR 攻击（ASR = 0.200 vs. 0.440）均更鲁棒。

2. **路由级安全分析**。刻画了 48 层 MoE（128 专家，top-8 路由）中有害与良性输入路由模式的分歧，证明安全行为分布在路由函数中而非集中于个体专家。

3. **对齐脆弱性度量与灾难性失效演示**。形式化了对齐脆弱性度量，并展示随机路由扰动将拒绝率从 100% 降至 8.4%±5.7%（脆弱性 $F = 0.680$），确立了 MoE 对齐对学习到的路由决策的关键依赖。

4. **基于 CKA 的架构比较**。将 MoE 模型的分布式路由分歧与稠密模型的" 沙漏"CKA 模式进行对比，提供了架构层面的洞见。

2 相关工作

越狱攻击与防御。 对已对齐 LLM 的对抗性攻击涵盖梯度方法（GCG 15）、LLM 驱动的迭代优化（PAIR 1）、遗传变异（AutoDAN 6）以及简单的模板提示 [12]。JailbreakBench [2] 和 HarmBench [7] 等基准提供了标准化评估。然而，几乎所有评估都仅关注稠密架构；MoE 模型在这些基准中明显缺席。

混合专家架构。 MoE 范式 [11] 经 Switch Transformers [4] 扩展后，已通过 Mixtral [5]（8 专家，top-2 路由）、DeepSeek-MoE [3]（细粒度分割）和 Qwen3-MoE [9]（128 专家，48 层 top-8 路由）快速演进。这些进展聚焦于效率和质量；路由的安全属性尚未被探索。

MoE 模型的安全性。 SAFEx [14] 提出了安全专家路由机制；RASA [13] 探索了路由感知的安全对齐。然而，两者均提出缓解方案而未建立受控的 MoE vs. 稠密比较。**据我们所知，此前没有工作提供同一模型家族内 MoE 和稠密模型之间对齐鲁棒性的匹配对比较**。本工作填补了这一空白。

3 方法

3.1 问题定义

令 M_{moe} 为一个 MoE 语言模型，每层包含路由器 R 和专家集合 $\{E_1, \ldots, E_N\}$，令 M_{dense} 为同族的稠密模型。两个模型均通过相同的对齐流程 $\mathcal{A}$ 在训练数据 $\mathcal{D}$ 上完成对齐。令 $\mathcal{P}_{\text{harm}}$ 和 $\mathcal{P}_{\text{benign}}$ 分别为有害和良性提示集合。

模型 M 在提示集 $\mathcal{P}$ 上的拒绝率定义为：

$$\rho(M, \mathcal{P}) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbb{1}\left[\text{Judge}(M(p)) = \texttt{refuse}\right] \tag{1}$$

Algorithm 1 基于随机路由的对齐脆弱性测量

Require: MoE 模型 M（L 层，每层 N 专家，top-k 路由）；有害提示集 $\mathcal{P}_{\text{harm}}$；试验次数 n
Ensure: 脆弱性分数 F

$R \leftarrow []$
for $i = 1, \ldots, n$ **do**
  for 每层 ℓ 和每个 token 位置 t **do**
    $\pi_i(\ell, t) \leftarrow \text{UniformSample}(\{1, \ldots, N\}, k)$
  end for
  $r_i \leftarrow \rho(M, \mathcal{P}_{\text{harm}} \mid \pi = \pi_i)$
  将 r_i 加入 R
end for
$\mu_r \leftarrow \text{mean}(R)$, $\sigma_r \leftarrow \text{std}(R)$
return $F = \sigma_r / \mu_r$

攻击 $\mathcal{T}$ 的攻击成功率（ASR）为：

$$\text{ASR}(\mathcal{T}, M, \mathcal{P}_{\text{harm}}) = 1 - \rho(M, \mathcal{T}(\mathcal{P}_{\text{harm}})) \tag{2}$$

此外报告 ASR@k，定义为 k 种攻击变体中至少一种成功的行为占比。

对 MoE 模型，定义路由策略 π 为将每层每个 token 映射到 k 个专家子集的函数。学习到的路由策略 π^* 由训练好的路由器 R 决定。我们考虑扰动 $\pi' \neq \pi^*$ 并测量：

$$r_i = \rho(M_{\text{moe}}, \mathcal{P}_{\text{harm}} \mid \pi = \pi_i) \tag{3}$$

3.2 对齐脆弱性度量

通过对齐脆弱性分数形式化对齐对路由扰动的敏感度：

$$F = \frac{\sigma_r}{\mu_r}, \quad \text{其中} \quad \mu_r = \frac{1}{n}\sum_{i=1}^{n} r_i, \quad \sigma_r = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \mu_r)^2} \tag{4}$$

F 即随机路由下拒绝率的变异系数。$F > 0.5$ 表示标准差至少为均值的一半，对齐高度敏感于路由变化。稠密模型平凡地有 $F = 0$。

3.3 路由扰动实验

算法 1 描述了用于测量对齐脆弱性的强制路由程序。
除随机路由外，我们还进行定向干预：

- **安全专家 OFF**：识别有害与良性提示之间激活频率差异最大的 top-5 专家，在有害输入上将其排除在路由之外。
- **安全专家 ON**：强制将 top-5 差异激活专家纳入路由决策。

表 1: 所有实验使用的模型对。两个模型共享训练数据和对齐流程，主要差异在于架构。

模型	架构	总参数量	活跃参数量	量化
Qwen3-30B-A3B	MoE (128 专家, top-8)	30B	3B	AWQ 4-bit
Qwen3-4B	稠密	4B	4B	FP16

表 2: 基线拒绝率（无攻击）。3 个种子的均值 ± 标准差。

模型	有害拒绝率 ↑	良性假阳性率 ↓	有害 ASR ↓
Qwen3-30B-A3B (MoE)	1.000 ± 0.000	0.020 ± 0.000	0.000
Qwen3-4B (稠密)	1.000 ± 0.000	0.020 ± 0.000	0.000

3.4 基于 CKA 的稠密模型分析

为比较稠密模型如何在无路由机制的情况下区分有害和良性内容，我们在每层 ℓ 计算有害与良性隐藏状态矩阵之间的中心核对齐（CKA）。CKA 接近 1 表示表征相同；较低值表示存在区分。

3.5 公平性控制

两个模型共享 Qwen3 家族、训练数据和对齐流程 (SFT + RLHF)。活跃参数量可比 (MoE: 3B, 稠密: 4B——1.3× 的差异有利于稠密模型)。所有提示、生成参数（温度 = 0，最大 token 数 = 512）以及安全评判器（基于关键词，38 个拒绝/8 个顺从指标）均共享。模板攻击使用 3 个随机种子（42, 123, 456），报告均值 ± 标准差。表 1 总结了模型对。

4 实验

4.1 黑盒越狱鲁棒性

假设。 若 MoE 路由提供了额外的计算自由度，我们预期：(a) MoE 模型更脆弱——因为对抗输入可利用路由决策；或 (b) 更鲁棒——因为更大的总参数量提供了更强的对齐容量。

基线。 首先验证两个模型在无攻击条件下均已良好对齐。表 2 显示两者对有害提示的拒绝率均为 100%，良性假阳性率仅 2%，确认对齐强度匹配。

模板攻击。 应用 6 种越狱模板（DAN、AIM、前缀注入、角色扮演、base64 编码、少样本引导），覆盖 50 个有害行为，3 个随机种子，共 $6 \times 50 \times 3 = 900$ 次攻击试验。结果见表 3 和图 1。

PAIR 攻击。 使用 PAIR 风格的迭代攻击 [1]，10 种优化策略覆盖 50 个行为。表 4 显示 MoE 模型仍更鲁棒（ASR = 0.200, 10/50）vs. 稠密模型（ASR = 0.440, 22/50）——2.2× 的差距与模板攻击结论一致。

表 3: 模板攻击结果（6 模板，50 行为，3 种子）。均值 ± 标准差。稠密模型的 ASR 是 MoE 的 2.6×，且 ASR@$k = 1.0$，即每个有害行为都能被至少一个模板诱导。

模型	**ASR** ↓	**ASR@k** ↓
Qwen3-30B-A3B (MoE)	0.069 ± 0.002	0.367 ± 0.024
Qwen3-4B (稠密)	0.177 ± 0.000	1.000 ± 0.000

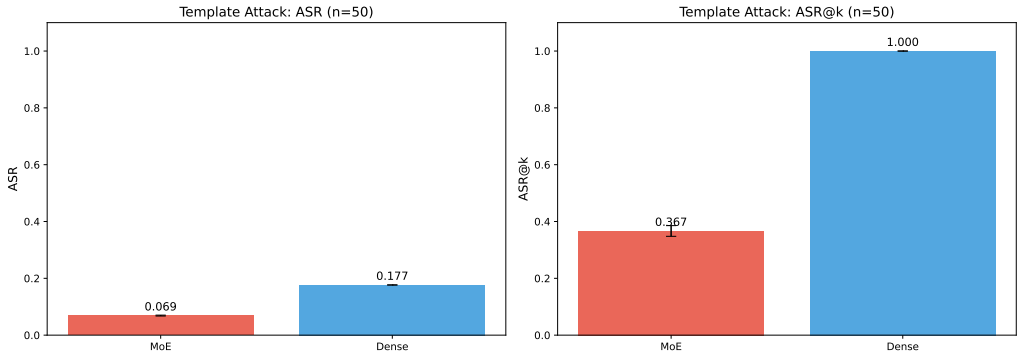

图 1: 模板攻击结果（50 行为）。**左：**各模板 ASR。**右：**ASR@k。稠密模型显著更脆弱——每个行为都可被至少一个模板诱导（ASR@$k = 1.0$），而 MoE 仅 36.7%。

发现 1. *MoE 模型在所有攻击类型上均比匹配的稠密模型更鲁棒*（模板 ASR: 0.069 vs. 0.177; PAIR ASR: 0.200 vs. 0.440），在 50 个行为、3 个种子上验证。这与" 稀疏路由引入额外攻击面" 的假设相矛盾，转而表明 MoE 模型更大的总参数量提供了更强的鲁棒对齐容量。

4.2 路由模式分析

假设。 若 MoE 路由器学习了内容类型感知的路由，我们预期有害与良性提示之间的专家激活模式存在系统性分歧。

Jensen-Shannon 散度。 对 48 层 MoE（128 专家，top-8 路由）的每一层，计算有害和良性提示集的专家选择频率分布间的 JSD。图 2 展示了逐层 JSD 分布。

所有层的平均 JSD 为 0.194，最大值 0.308 出现在第 31 层。关键是分歧分布在所有层中，而非集中于少数几层，确认路由器在整个网络中执行内容类型敏感的路由。

路由熵。 计算专家选择分布的 Shannon 熵（图 3）：

$$H(\pi_\ell^*) = -\sum_{j=1}^{N} p_{\ell,j} \log_2 p_{\ell,j} \tag{5}$$

有害提示表现出更低的路由熵（5.665 bits）vs. 良性提示（6.083 bits）。这 0.42 bit 的差距表明路由器将计算集中到更窄的专家子集来处理潜在有害内容，与专门化的安全相关处理一致。

发现 2. *MoE 路由器已学习了内容类型感知的路由。*有害和良性输入被系统地路由到不同的专家子集（均值 JSD = 0.194），有害提示接受更集中的路由（熵差 = 0.42 bits）。

表 4: PAIR 攻击结果（10 策略，50 行为）。稠密模型的脆弱性是 MoE 的 2.2×。

模型	PAIR ASR ↓	成功/总数
Qwen3-30B-A3B (MoE)	0.200	10 / 50
Qwen3-4B (稠密)	0.440	22 / 50

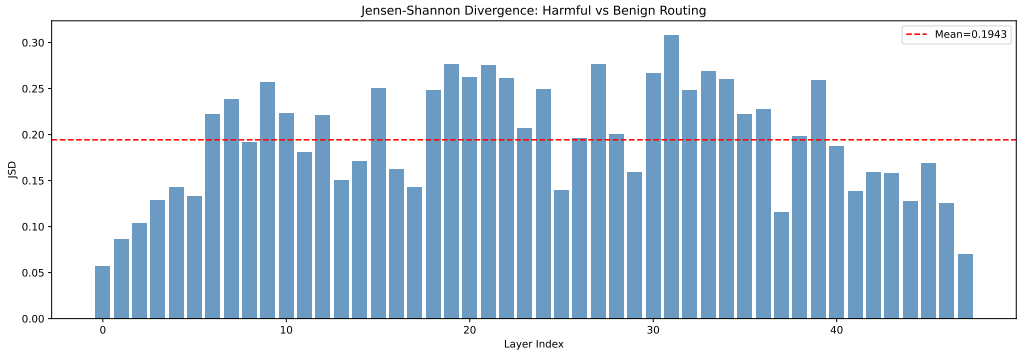

图 2: 有害与良性专家路由分布间的逐层 Jensen-Shannon 散度。均值 JSD = 0.194（虚线）。最大分歧出现在第 31 层（JSD = 0.308），但有意义的分歧分布在全部 48 层上。对比稠密模型的”沙漏” 模式（图 7），后者的区分仅限于早期和晚期层。

4.3 安全专家干预与对齐脆弱性

假设。 若安全性局部化在特定” 安全专家” 中，停用它们应降低拒绝率。反之，若安全性整体依赖于路由函数，即使非定向的路由扰动也应降低安全性。

专家识别与定向干预。 按有害与良性提示间的激活频率差异对全部 $48 \times 128 = 6{,}144$ 个专家排序（图 5）。表 5 展示强制路由结果。停用 top-5 差异激活专家（**安全专家 OFF**）无任何效果——拒绝率保持 1.000。强制激活它们（**安全专家 ON**）反而略微降低拒绝率至 0.950。两种干预均未实质性改变安全行为。

随机路由导致灾难性失效。 与上述形成鲜明对比的是，将学习到的路由替换为均匀随机专家选择（算法 1）导致近乎完全的对齐崩溃。在 22 次试验中，拒绝率从 1.000 降至 0.084 ± 0.057（范围：0.000–0.200），脆弱性分数 $F = 0.680$。

发现 3 和 4. *安全性并未局部化在个体专家中，而是编码在路由函数中。*没有任何小规模专家子集对安全行为是必要或充分的，但整体扰动路由会导致灾难性失效（$F = 0.680$）。这是 MoE 架构独有的失效模式，在稠密模型中无类似物。

4.4 稠密模型对比：CKA 分析

假设。 稠密模型必须通过表征级机制区分有害与良性内容（因其无路由机制）。我们预期这种区分局部化在特定层中。

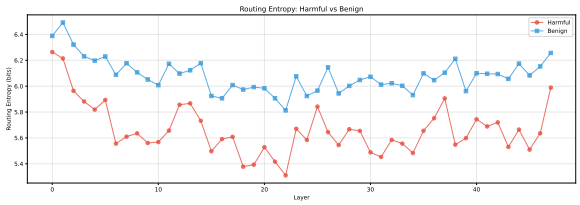

图 3: 有害 vs. 良性提示的逐层路由熵。有害提示一致地表现出更低的熵（更集中的路由），表明路由器将有害内容导向专门化的通路。

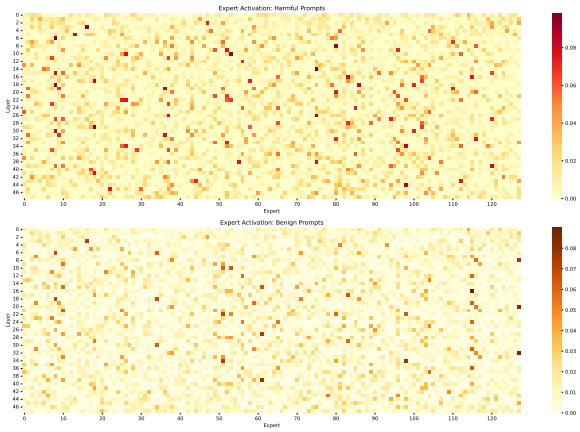

图 4: 有害（上）和良性（下）提示的专家激活频率热力图。每个单元格表示专家 j 在第 i 层的选择频率。有害提示呈现出明显更集中的激活模式。

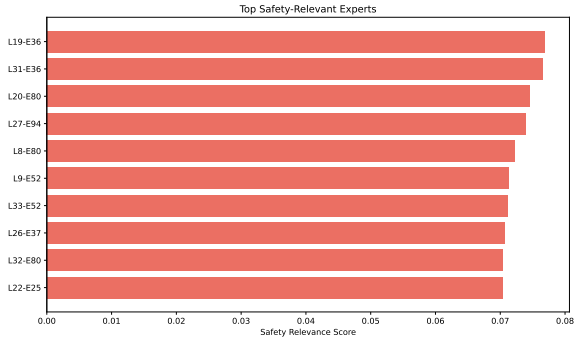

图 5: 按激活频率差异排序的 top-10 专家。尽管存在明确的差异激活，定向操纵的影响有限（表 5）。

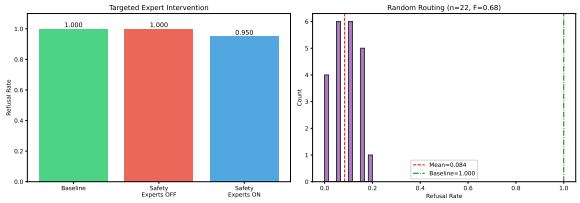

图 6: 路由干预下的拒绝率。定向专家操纵（左侧）影响极小；随机路由（右侧）导致灾难性失效（$F = 0.680$）。

结果。 图 7 展示了稠密模型中有害与良性隐藏状态间的逐层 CKA。出现了引人注目的" 沙漏" 模式：早期层（L0–L3）CKA 低（0.15–0.23），中间层（L6–L24）处理几乎相同（CKA ≈ 0.966），晚期层（L27–L33）出现中等程度的区分（CKA $= 0.65$–0.94）。所有层的平均 CKA 为 0.797。

架构对比。 核心洞见在于分布式 MoE 安全（每层路由分歧；均值 JSD $= 0.194$，最大 $= 0.308$）与局部化稠密安全（沙漏 CKA 模式，区分仅限于约 25% 的层）之间的对比。这解释了 MoE 模型在正常运行下更强鲁棒性（分布式安全更难绕过）及其在路由扰动下灾难性失效（扰动路由同时在所有层破坏安全）的原因。

发现 5. *MoE 和稠密模型通过架构上截然不同的机制实现安全性*——分布式路由分歧 vs. 局部化表征区分——具有质性不同的鲁棒性与脆弱性属性。

表 5: Qwen3-30B-A3B 上的路由干预结果。定向专家操纵影响极小，但随机路由导致灾难性的对齐失效。

路由条件	**有害拒绝率**
基线（学习路由 π^*）	1.000
安全专家 OFF（排除 top-5 差异专家）	1.000
安全专家 ON（强制 top-5 差异专家）	0.950
随机路由（均值 $\pm$ 标准差，$n = 22$）	0.084 ± 0.057
随机路由（最小/最大）	0.000 / 0.200
对齐脆弱性分数 F	**0.680**

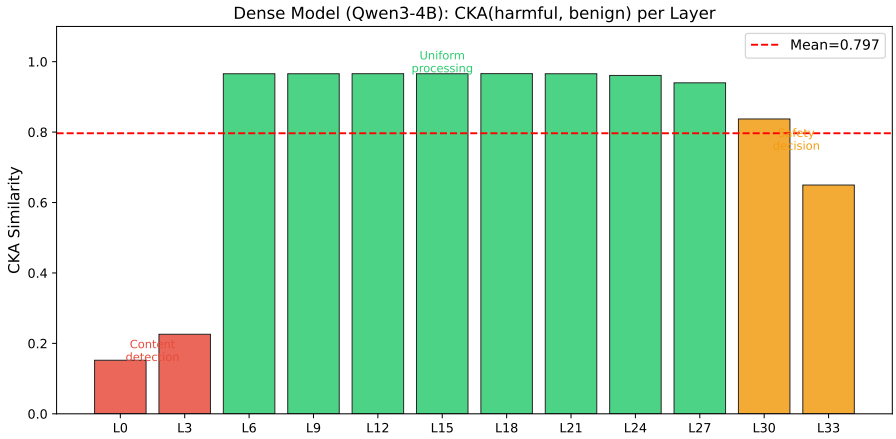

图 7: 稠密模型（Qwen3-4B ）中有害与良性隐藏状态间的逐层 CKA 相似度。”沙漏”模式揭示了稠密模型主要在早期检测层（L0–L3, CKA = 0.15–0.23）和晚期决策层（L27–L33, CKA = 0.65–0.94）区分有害内容，中间层处理两者几乎相同（CKA $\approx$ 0.966）。

5 讨论

为什么 MoE 更鲁棒？ MoE 模型在黑盒攻击下的优越鲁棒性（发现 1）有两种互补解释。*总参数假设*：MoE 模型有 10× 更多的总参数（30B vs. 4B），尽管每 token 仅激活 3B，但提供了更丰富的表征空间，使对齐可在专家组合中冗余编码。*分布式安全假设*：逐层路由分歧意味着 MoE 模型在*每一层*独立地检测和响应有害内容，形成纵深防御——更难被单一对抗变换绕过。

路由即对齐。 定向与随机路由扰动之间的尖锐不对称（发现 3–4）是本工作的核心结果。移除特定”安全专家”无效果，但随机化所有路由导致灾难性失效。这意味着安全不是个体专家参数的属性——随机路由仍激活相同的专家，只是组合错误——而是路由器施加的*协调模式*的属性。路由器 R 而非专家集合 $\{E_j\}$ 是关键的安全组件。**在 MoE 模型中，路由就是对齐。**

稠密沙漏 vs. MoE 分布式安全。 架构对比（发现 5）揭示了一个根本性权衡。稠密模型将安全集中在可辨识的层中（早期检测 + 晚期决策），便于定向安全探测但容易被通过早期层的对

抗输入攻破。MoE 模型将安全分布在每层的路由中，提供纵深防御但创造了单一故障点：路由函数本身。

实践启示。 我们的发现引出四个 MoE 对齐方向：**(1)** 路由感知的 *RLHF*——强制在多样路由配置下保持一致的安全行为；**(2)** 对抗性路由训练——在对齐过程中加入路由扰动，类似于鲁棒优化中的对抗训练；**(3)** 路由一致性正则化——惩罚同一输入的干净和对抗变体之间的路由分布偏移；**(4)** 专家安全冗余——确保安全不仅编码在路由模式中，也编码在专家参数中。

6 局限性

单一模型家族。本对比使用 Qwen3 模型；家族特定的效应（架构、训练）可能不适用于 Mixtral 或 DeepSeek-MoE。跨家族复现至关重要。**提示集规模。**50 个有害行为集小于标准基准（HarmBench: 200+）；虽足以支撑我们观察到的大效应量，且在 20 和 50 行为评估间一致，但可能遗漏细微的类别特定差异。**AWQ 量化。**MoE 模型使用 AWQ 4-bit 量化，稠密模型使用 FP16；量化可能影响路由精度。我们的结果可能代表全精度 MoE 鲁棒性的下界。**关键词安全评判器。**基于关键词匹配的评判器（38 拒绝/8 顺从指标）可能遗漏模型评判器能捕捉的微妙越狱。我们在所有条件下一致地应用此评判器。**脆弱性度量范围。**脆弱性分数基于 $n = 22$ 次均匀随机路由试验计算；更多试验和结构化扰动（如对抗性路由优化）可能揭示不同模式。

7 结论

本文首次对 MoE 和稠密语言模型的对齐鲁棒性进行了匹配对比较。通过在 Qwen3 家族（30B-A3B MoE vs. 4B 稠密）上的受控实验，我们确立了三个主要发现。

首先，MoE 模型对对抗性越狱攻击可以更鲁棒（模板 ASR 0.069 vs. 0.177；PAIR ASR 0.200 vs. 0.440），挑战了稀疏路由本质上更不安全的直觉。

其次，MoE 模型中的安全行为分布在路由函数中，而非集中于可辨识的” 安全专家”——定向专家移除对拒绝率无影响，而路由器系统性地将有害内容导向与良性内容不同的专家子集（均值 JSD $= 0.194$，熵差 $= 0.42$ bits）。

第三，MoE 对齐在路由扰动下灾难性地脆弱：随机路由将拒绝率从 100% 降至 8.4%±5.7%（脆弱性分数 $F = 0.680$），揭示对齐编码在学习到的路由决策中而非专家参数中。

这些发现推动了 MoE 对齐的范式转变：路由即对齐，安全训练必须通过路由感知的 RLHF、对抗性路由训练和路由一致性正则化来明确考虑这一点。随着 MoE 架构主导前沿，保护路由机制是安全部署的先决条件。所有实验在单块消费级 GPU（NVIDIA RTX 5090, 32 GB）上完成；代码和数据将在发表后公开。

参考文献

[1] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

[2] Patrick Chao, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*, 2024.

[3] Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.

[4] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[5] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[6] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*, 2023.

[7] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Keisuke Sakaguchi, David Bau, J Zico Kolter, Dan Hendrycks, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

[8] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[9] Qwen Team. Qwen3 technical report. https://qwenlm.github.io/blog/qwen3/, 2025.

[10] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.

[11] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[12] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. Do anything now: Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*, 2023.

[13] Yuxuan Wang et al. Rasa: Routing-aware safety alignment for mixture-of-experts models. *arXiv preprint*, 2024.

[14] Zhihao Yang et al. Safex: Safe expert routing for mixture-of-experts language models. *arXiv preprint*, 2024.

[15] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. In *International Conference on Machine Learning*, 2024.