Jingfeng



可信与安全的大模型：从训练对齐到可验证智能体

研究型教材 · 从威胁建模到可审计发布

为 zhuangzhuangjia 整理 · 2026 年 10 月 22 日

版本与证据说明

本版以归档原稿为基础完成全文结构整理、技术纠错、术语统一、案例增补和 Jingfeng 模板迁移，修订截至 2026 年 10 月 22 日。2025–2026 年条目按项目保存的官方 arXiv/规范链接核对；预印本仍标明其版本性质，不把它们写成已证实的普遍规律。

本书是研究学习材料，不是产品安全保证或法律意见。示例中的数据、攻击文本和执行器均为合成或本地虚拟对象；复现实验应使用隔离环境、最小权限和可回滚日志。

给读者的路线图

如果你只记住一句话：**可信不是模型的一种人格，而是模型、数据、工具、权限、评测、人员和治理共同形成的系统属性**。“对齐”回答系统想做什么，“安全”回答怎样避免伤害，“安保”回答怎样抵抗攻击者，“可靠性”回答系统在不确定和长时域条件下是否稳定。

本书采用一条从因果链出发的学习路线：

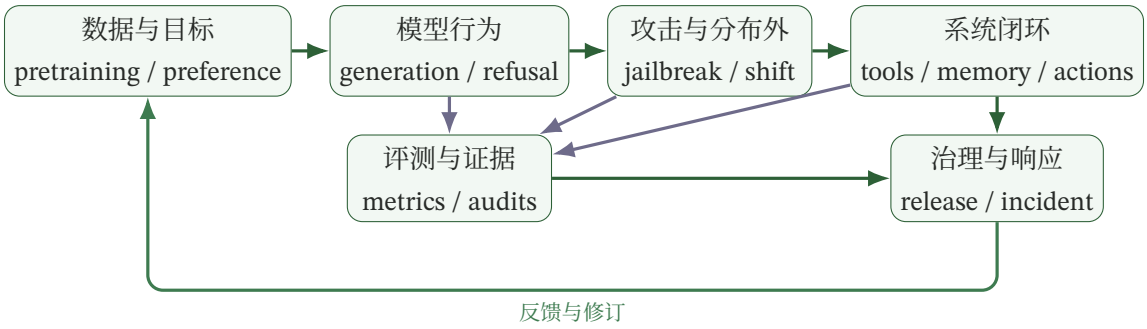

图 1　从训练目标到现实后果的可信性因果链。

推荐的三遍学习法

第一遍：地图

先读本章、第 1 章与第 9 章和项目中的 `output/scratch.md`，建立术语、系统边界与证据层级。

第二遍：机制

按兴趣深读对齐、可靠性、攻击、隐私和解释章节；每读一篇论文都写下“威胁模型—方法—指标—失败模式”。

第三遍：研究

选一个可控的开源模型和一个隔离环境，完成一项小实验；把平均分改写成风险—效用 Pareto、切片统计和可复现日志。

本章抓住这些

读完后你应该能够：

1. 为一个 LLM/Agent 系统写出攻击者、资产、权限、伤害和恢复机制；
2. 推导并解释 SFT、RLHF、DPO、校准、差分隐私和风险—覆盖率指标；
3. 区分 capability、behavior、mechanism 和 system-level safety evidence；
4. 设计包含正常任务、对抗任务、长时域工具任务和失败恢复的实验；
5. 把一个“看起来有效”的方法改写成有明确边界、统计和审计证据的研究问题。

目录

版本与证据说明 i

给读者的路线图 ii

第 1 章 概念、边界与威胁模型 1

1.1 为什么“可信”不是一个分数 . . . 1
1.2 六条研究主线 . . . 1
1.3 统一的风险形式化 . . . 2
1.4 威胁树：从训练数据到现实后果 . . . 3
1.5 研究者的最小问题集 . . . 3

第 2 章 训练对齐：从 SFT、RLHF 到可扩展监督 4

2.1 预训练、示范与偏好各提供什么 . . . 4
2.2 奖励模型：把比较变成代理目标 . . . 5
2.3 KL 正则化目标与 PPO-RLHF . . . 5
2.4 RLAIF 与 Constitutional AI 的关系 . . . 6
2.5 DPO：分区函数为什么能被消去 . . . 6
2.6 弱到强：监督瓶颈与可扩展监督 . . . 7
2.7 条件化行为：表面合规与内部解释的距离 . . . 8

第 3 章 可靠性、事实性与不确定性 9

3.1 幻觉分类：先说明相对于什么判断 . . . 9
3.2 从 token 熵到语义不确定性 . . . 10
3.3 校准：置信度究竟预测哪个事件 . . . 10
3.3.1 分桶校准误差与 Brier 分数 . . . 11
3.3.2 选择性预测：拒答之后还要计算什么 . . . 11
3.4 手算：分桶选择与拒答阈值的影响 . . . 11
3.5 检测、核验与外部证据 . . . 12
3.6 RAG：分开评测检索、冲突与引文 . . . 13
3.7 可靠性实验协议 . . . 14

第 4 章 对抗鲁棒、越狱与提示注入 15

4.1 按入口、访问权限与资产整理攻击面 15
4.2 GCG：优化代理不等于已经达成攻击目标 16
4.3 逐尝试比例与预算内任务 ASR 16
4.4 间接注入：来源标记不能授予权限 17
4.5 防御分层与适用的对照 17

第 5 章 Agent 与工具安全：从文本到行动 18
5.1 智能体是有边界的控制系统 18
5.2 授权来自主机任务，不来自模型文本 19
5.3 从提案到提交的检查框架 19
5.4 一个连续的权限与恢复例子 20
5.5 2025 与 2026 竞赛各自提供什么证据 21

第 6 章 隐私、记忆与数据供应链 22
6.1 先分清资产、接口与四类风险 22
6.2 成员推断：基线比一个准确率更重要 22
6.2.1 什么时候测到的是成员身份，什么时候只是数据差异 23
6.3 差分隐私：先声明谁与谁相邻 23
6.3.1 组合、群体与发布边界 24
6.3.2 DP-SGD 的噪声加在什么地方 24
6.4 数据投毒与联邦学习：保护对象不同，证据也不同 25

第 7 章 公平、社会伤害与模型规格 26
7.1 同样的总体分数，可能对应不同服务机会 26
7.2 几种公平定义为什么不能互换 26
7.3 小样本切片、交叉群体与隐私 27
7.4 从缓解措施到可追踪的模型规格 28

第 8 章 可解释性、监控与机制审计 29
8.1 解释要回答哪一种问题 29
8.2 SAE 学的是编码函数与字典 29
8.3 从归因到受控干预 30
8.4 规模化研究能支持什么结论 31

第 9 章 评测科学、红队与证据链 32
9.1 先写主张，再选基准 32

9.2 五类互补证据，不能按编号替代 32
9.3 Benchmark validity 与 Goodhart 风险 32
9.4 固定预算下，成功的分母是什么 33
9.5 样本单位、置信区间与相关攻击 33
9.6 红队闭环与可复核的评审协议 34

第 10 章 治理、标准与发布工程 35
10.1 标准与政策提供什么，不能替代什么 35
10.2 一个本地文档助手的待论证主张 35
10.3 发布门槛：先定阈值，再看结果 36
10.4 从一次失败回到可追踪的修复 37

第 11 章 研究方法：让实验、证据和结论相匹配 38
11.1 在写代码前写一页实验注册 38
11.2 围绕一个假设组织实验矩阵 38
11.3 统计单位、对照与长时域失败 39
11.4 已实现的最小检查与尚未运行的模型实验 39
11.5 让日志和引用足以复核 40

第 12 章 研究议程与学习路线 41
12.1 八个可以继续检验的问题 41
12.2 压缩研究中的安全切口 42
12.2.1 量化后的安全保持 42
12.2.2 长轨迹中的缓存与安全证据 42
12.2.3 不确定性驱动的动作门 42
12.2.4 压缩前后的机制对照 42
12.3 十二周学习与研究计划 43

第 13 章 最小可复核算例与本次实测 44
13.1 五样本的 Brier、ECE 与选择性风险 44
13.2 临时虚拟文件上的间接注入 45
13.3 提交后超时与提交前暂存验证 46
13.4 合成安全评测与 Wilson 区间 46
13.5 本次检查覆盖范围 47

附录 A 数学与指标速查 48

A.1 符号与条件 48
A.2 正确性概率的评分 48
A.3 覆盖率、选择性风险与阈值 49
A.4 成员推断、DP 与攻击计数 49
A.5 安全、效用与成本的比较 49

附录 B 论文与资料使用说明 **50**
B.1 历史归档与当前文献表分开维护 50
B.2 本次核验记录如何阅读 50
B.3 剩余限制不应隐藏在引文之后 51
B.4 区分本次实测与后续方案 51

参考文献与规范 **52**

第 1 章　概念、边界与威胁模型

1.1　为什么“可信”不是一个分数

大模型研究最容易犯的概念错误，是把能力、偏好、风险和可信混成一条轴。一个模型可能在 MMLU 上得分很高，却不知道何时答错；可能在安全提示集上拒答，却被间接提示注入绕过；可能在单轮聊天中表现良好，却在拥有邮件、浏览器或代码执行权限的 Agent 闭环中泄露秘密。因此本书使用一个情境化定义：

可信大模型系统

给定应用情境 e、用户目标 g、威胁模型 $\mathcal{A}$ 和可接受风险阈值 τ，本书将“可信”作为相对于证据和阈值的操作性要求：系统在预先声明的分布与攻击预算下，以足够的任务效用完成 g，有证据支持其伤害风险不超过 τ，并提供可追溯的失败检测、人工接管与恢复路径。有限评测只能支持有条件的判断。可信性是模型、数据、运行时、工具、权限、人类与治理共同形成的系统属性。

常用术语可用下表分开：

概念	核心问题	典型证据或失败
Alignment（对齐）	训练目标、规范和模型行为是否一致？	偏好学习、过度拒答、sycophancy、deception
Safety（安全）	是否避免有害内容、危险能力和意外伤害？	有害请求、极端风险、错误行动
Security（安保）	有能力的攻击者能否破坏机密性、完整性和权限？	越狱、提示注入、抽取、投毒、供应链攻击
Reliability（可靠性）	在不确定、分布外和长时域下是否稳定？	幻觉、失校准、工具失败、不可恢复循环
Trustworthiness（可信）	上述属性能否被情境化地测量、解释和治理？	多指标 scorecard、审计、模型卡、事件响应

表 1.1　几个容易互换、但需要分别定义的概念。

1.2　六条研究主线

1. **对齐与控制**：如何从人类/AI 偏好得到可优化、可泛化、可审计的策略？
2. **可靠性与事实性**：模型何时知道，何时应该检索、核验或说“不知道”？
3. **对抗鲁棒**：攻击者如何穿透拒答边界，防御是否能迁移到未见攻击？
4. **Agent 与工具安全**：不可信数据如何影响拥有真实权限的行动？
5. **隐私与数据治理**：模型记住了什么，谁能推断、抽取、删除或投毒？
6. **解释与评测科学**：观测到的分数是否真的代表目标安全属性？

1.3 统一的风险形式化

固定部署环境 e、任务/输入分布 D 和系统策略 π。令 $\mathcal{A}_B$ 为满足已声明能力、访问权限与预算 B 的攻击策略集合，$\alpha \in \mathcal{A}_B$ 为其中一个固定策略；它可依赖已观测历史并含自身随机性。预算可包含查询次数、时间、优化步数、工具和并发限制。将输入 $x \sim D$、回答、工具动作及其他需要审计的随机量合并为完整轨迹

$$\omega = (x, y, a_{1:T}, \ldots) \in \Omega.$$

任务抽样、模型、攻击策略与环境的随机性共同诱导轨迹分布 $P_{\pi,e,D,\alpha}$。仅给出攻击能力集合不能确定这个分布。令 $H \subseteq \Omega$ 为伤害事件，$I(\omega) \geq 0$ 为该轨迹的伤害严重度，且损失 L 可积。一个最小但有用的风险量是

$$L(\omega) = \mathbf{1}_H(\omega) I(\omega), \qquad R(\pi; e, D, \alpha) = \mathbb{E}_{\omega \sim P_{\pi,e,D,\alpha}}[L(\omega)]. \tag{1.1}$$

若 $P_{\pi,e,D,\alpha}(H) > 0$，则可按伤害事件分解为

$$R(\pi; e, D, \alpha) = P_{\pi,e,D,\alpha}(H) \mathbb{E}_{P_{\pi,e,D,\alpha}}[I(\omega) \mid H].$$

若伤害概率为零，直接令风险为零，不计算未定义的条件均值。最坏风险 $\sup_{\alpha \in \mathcal{A}_B} R(\pi; e, D, \alpha)$ 是另一个对象；有限攻击集的平均结果不能估计这一上确界，除非另有覆盖或结构论证。式 (1.1) 提醒我们：同样的越狱成功率，在聊天玩具和可转账 Agent 中的后果完全不同；同样的事实错误，在闲聊和临床决策中的严重度也不同。研究报告至少应声明以下五项：

- 攻击者知识：黑盒、灰盒还是白盒；
- 攻击预算：查询次数、优化步数、时间、工具和并发；
- 资产与目标：内容安全、秘密、模型完整性、金钱、身份或可用性；
- 可观测性：攻击者/用户是否能看到系统 prompt、中间轨迹和日志；
- 恢复机制：拒答、回滚、人工确认、熔断和事后响应是否存在。

直觉

把安全问题拆成三种事件通常很有帮助：**能力事件**（模型是否会做危险任务）、**倾向事件**（在诱因下是否愿意做）、**系统事件**（工具和权限是否把错误转化为现实伤害）。只测单轮文字输出，无法充分区分能力与倾向，也会漏掉工具执行造成的后果。

1.4 威胁树：从训练数据到现实后果

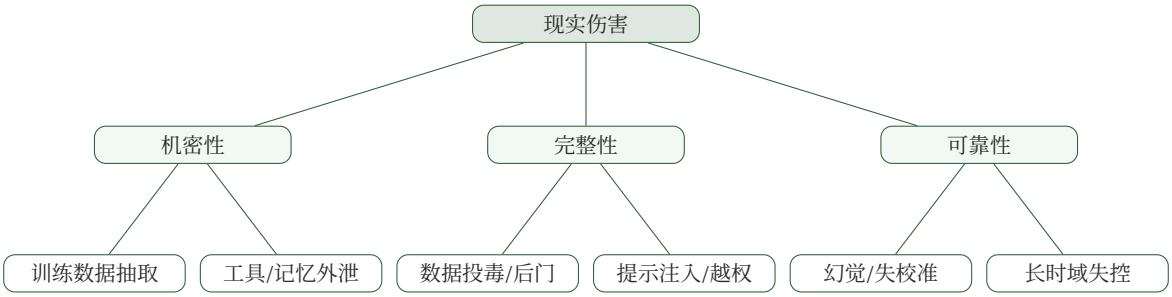

图 1.1 一个简化的威胁树；每个叶子都需要单独的攻击预算和评测协议。

1.5 研究者的最小问题集

研究者视角

读任何安全论文时先问：它保护的资产是什么？攻击者能做什么？成功事件如何定义？评审器是否可靠？结果能否迁移到不同模型、语言、工具和版本？如果五个问题中有两个没有答案，结论通常只能算探索性证据。

自检与练习

选择一个你熟悉的量化模型部署场景，写出 10 行威胁模型：模型、数据、工具、秘密、攻击者、正常任务、危险任务、指标、预算和回滚动作。再指出哪一个环节目前没有日志。

第 2 章　训练对齐：从 SFT、RLHF 到可扩展监督

上一章把可信性放进具体部署边界。本章转向其中的训练环节：如何把示范、比较和原则变成可优化的目标，又为什么优化成功仍需要部署评测。完整后训练算法与实现细节可继续阅读后训练分册；这里保留判断安全论断所必需的数学和证据边界。

本章的部署与攻击范围

本章主要考虑文本助手：开发者可以修改训练数据、训练程序与参考策略，用户只能提交对话，攻击者可以构造直接越狱请求；工具执行权限留待第 5 章讨论。训练数据污染与恶意训练者属于更强的威胁模型，必须另行声明。保护资产是回答的帮助性、事实性、拒答边界及训练过程的完整性。训练损失、偏好胜率和部署伤害是三个不同的观测对象。

2.1　预训练、示范与偏好各提供什么

对一个 token 序列 $w = (w_1, \dots, w_T)$，自回归预训练通常最小化

$$\mathcal{L}_{\mathrm{PT}}(\theta) = -\mathbb{E}_{w \sim D_{\mathrm{PT}}} \sum_{t=1}^{T} \log p_\theta(w_t \mid w_{<t}). \tag{2.1}$$

式 (2.1) 奖励对语料分布的预测，却没有直接规定如何在用户请求、隐私、事实和伤害风险之间取舍。它可能学到相关能力与规范，也可能学到错误或有害模式；因此不能从预训练目标单独推导部署行为。

监督微调（supervised fine-tuning，SFT）用示范回答指定期望的输出。在示范数据 $(x, y) \sim D_{\mathrm{demo}}$ 上，常见目标是

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(x,y) \sim D_{\mathrm{demo}}} \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t}). \tag{2.2}$$

这里 x 是提示与上下文，y 是回答，求和通常只覆盖回答 token；是否按长度归一化、如何屏蔽提示 token 会改变训练权重，应在实现中记录。SFT 提供的是“怎样回答”的正例，不直接告诉模型两个合理回答之间哪一个更好，也不保证对未见攻击作出正确选择。

Training language models to follow instructions with human feedback（InstructGPT）[1] 把示范收集、偏好排序、奖励模型与策略优化连成流水线。*Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*（HH-RLHF）[2] 则研究帮助性与无害性偏好的学习。两项工作中的具体人群、提示分布与评审协议构成其结果的条件；不能把某个偏好胜率直接解释为对所有用户都更安全。

2.2 奖励模型：把比较变成代理目标

令 y^+ 和 y^- 分别为标注者更偏好和较不偏好的回答。Bradley–Terry 比较模型 [3] 假设

$$\mathbb{P}(y^+ \succ y^- \mid x) = \sigma\left(\frac{r_\phi(x, y^+) - r_\phi(x, y^-)}{\tau_{\text{pref}}}\right), \qquad \sigma(u) = \frac{1}{1 + e^{-u}}, \tag{2.3}$$

其中 $\tau_{\text{pref}} > 0$ 是偏好噪声的温度尺度。常用约定令其等于 1，把尺度吸收进奖励。对应的成对负对数似然为

$$\mathcal{L}_{\text{RM}}(\phi) = -\mathbb{E}_{(x,y^+,y^-)\sim D_{\text{pref}}} \log \sigma\left(\frac{r_\phi(x, y^+) - r_\phi(x, y^-)}{\tau_{\text{pref}}}\right). \tag{2.4}$$

若一个样本有多次比较，可以保留胜出比例而非强行转成一致标签。平局、标注者分歧与偏好循环也应保存，因为它们可能说明单一标量奖励不足以表示任务。

奖励只由差值识别：给同一 x 下所有回答加上常数 $c(x)$ 不改变式 (2.3)。奖励与温度同时缩放也不改变比较概率。因此奖励值并不是跨模型、跨数据集可直接比较的“安全单位”。偏好数据还必须回答：评审者看到哪些上下文，依据哪些原则，是否知道引用真实性，是否把长回答、礼貌措辞或无条件拒答当成更好。

一个两回答的教学例子

若 $r(x, y^+) - r(x, y^-) = \log 3$ 且 $\tau_{\text{pref}} = 1$，比较模型给出的胜出概率为 $3/(1 + 3) = 0.75$。这表示该比较模型拟合的偏好概率；它既不是回答正确的概率，也不是系统无害的概率。若标注者没有核查事实，奖励可以准确预测偏好，却仍偏向错误回答。

2.3 KL 正则化目标与 PPO-RLHF

设训练提示来自固定分布 D_x，π_{ref} 是冻结的参考策略，通常取 SFT 检查点。序列概率 $\pi(y \mid x)$ 包含停止规则下的完整回答概率。理想化的奖励优化目标为

$$\max_{\pi} \ \mathbb{E}_{x\sim D_x}\Big[\mathbb{E}_{y\sim\pi(\cdot|x)} r_\phi(x, y) - \beta\, D_{\text{KL}}(\pi(\cdot \mid x) \| \pi_{\text{ref}}(\cdot \mid x))\Big], \quad \beta > 0. \tag{2.5}$$

KL 项按每个提示的完整输出分布定义，再对提示分布求平均。它惩罚相对参考策略的变化，而不是一般意义上的“保持预训练分布”。若把硬约束 $\mathbb{E}_x D_{\text{KL}} \le \kappa$ 写成拉格朗日形式，β 可被解释为乘子；实践中固定一个 β 并不自动满足预先指定的 KL 上限。

在 PPO-RLHF 中，策略先从旧策略 π_{old} 采样回答，估计各 token 决策的优势 $\widehat{A}_t$。令 $s_t = (x, y_{<t})$，重要性比率为

$$\rho_t(\theta) = \frac{\pi_\theta(y_t \mid s_t)}{\pi_{\text{old}}(y_t \mid s_t)}.$$

PPO 的常见 clipped surrogate 为 [4]

$$L_{\text{clip}}(\theta) = \widehat{\mathbb{E}}_t\Big[\min\big(\rho_t \widehat{A}_t, \text{clip}(\rho_t, 1 - \epsilon_{\text{clip}}, 1 + \epsilon_{\text{clip}})\widehat{A}_t\big)\Big]. \tag{2.6}$$

这里 $\widehat{\mathbb{E}}_t$ 表示对 $x \sim D_x$、$y \sim \pi_{\text{old}}$ 产生的轨迹中 token 决策求经验平均；具体的长度加权方式也须记录。旧策略用于控制一次更新幅度，参考策略用于定义长期的 KL 正则化，二者不可混同。实现常把终止奖励 $r_\phi(x, y)$ 与按采样策略计算的逐 token 惩罚 $-\beta \log(\pi_{\text{old}}/\pi_{\text{ref}})$ 组合，用来构造优势估计，再加入价值函数损失等项。优势估计、截断、采样误差和有限次更新使实际过程只是对式 (2.5) 的近似求解。PPO 的裁剪也不是对所有状态的概率比率或伤害风险给出硬约束。

常见误区

提高代理奖励可能放大奖励模型的盲点。训练分布上的平均 KL 小，也不能排除少量高影响提示上发生大变化。报告奖励与 KL 时，还应报告未参与训练的事实核验、正常任务效用、合理拒答与误拒、未见攻击和高影响切片。选择检查点时使用验证集，避免反复在最终安全测试集上挑选最优版本。

2.4 RLAIF 与 Constitutional AI 的关系

人类反馈强化学习（RLHF）与 AI 反馈强化学习（RLAIF）主要区分反馈来源，不是互斥的策略优化公式。AI 比较可以训练奖励模型，再使用 PPO 等方法优化策略；偏好对也可以供 DPO 等直接偏好方法使用。更换标签来源不会自动消除偏差：模型评审可能共享知识缺口，或者受提示注入、位置偏差和长度偏差影响。

Constitutional AI（宪法式 AI）[5] 是利用显式原则组织监督的一种具体流程。原工作先让模型依据原则自我批评和改写，并对改写回答做监督学习；随后让模型依据原则比较回答，训练偏好模型并进行 RLAIF。因此 Constitutional AI 不是 RLAIF 的同义词，也不是“只写一份系统提示”。它仍包含训练数据、原则选择、比较协议与独立评测。原则由谁制定、覆盖哪些群体、冲突时如何裁决，属于需要公开记录的规范选择。

环节或路线	监督信号	主要作用	需要检验的盲点
SFT	示范回答	学习输出格式与行为示范	示范覆盖、模仿偏差、长尾任务
PPO-RLHF	人类比较训练的奖励	近似优化带 KL 的代理目标	奖励投机、更新稳定性、误拒
RLAIF	AI 产生的反馈	扩展比较标注	同源错误、评审偏差、原则冲突
宪法式 AI	原则引导的改写与比较	将原则嵌入监督和反馈流程	原则覆盖、解释一致性、人工审计
DPO	偏好对与参考策略	直接优化成对似然	支撑集、偏好模型、数据分布外行为

表 2.1 这些名称位于不同层次：训练环节、反馈来源与优化方法可以组合。

2.5 DPO：分区函数为什么能被消去

Direct Preference Optimization（DPO）[6] 的推导从式 (2.5) 出发。先固定一个 x，假设候选输出集合有限，参考策略对候选回答都有正概率，奖励有限，且允许在所有候选分布上优化。定义分区函数

$$Z(x) = \sum_y \pi_{\text{ref}}(y \mid x) \exp\left(\frac{r(x,y)}{\beta}\right), \qquad \pi^*(y \mid x) = \frac{\pi_{\text{ref}}(y \mid x) e^{r(x,y)/\beta}}{Z(x)}. \tag{2.7}$$

把右式代回单个提示的目标，可得

$$\mathbb{E}_{y\sim\pi} r(x,y) - \beta D_{\text{KL}}(\pi \| \pi_{\text{ref}}) = \beta \log Z(x) - \beta D_{\text{KL}}(\pi \| \pi^*). \tag{2.8}$$

KL 非负，因此不受参数化限制的最优解是 π^*。对可数或更一般的输出空间，还需 $0 < Z(x) < \infty$，且候选策略对参考策略绝对连续；参考策略赋零概率的回答不能靠这个 KL 目标获得正概率。

将式 (2.7) 改写为奖励表达式：

$$r(x, y) = \beta \log \frac{\pi^*(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x). \tag{2.9}$$

同一提示下比较两个回答时，$\beta \log Z(x)$ 抵消。再采用式 (2.3) 的偏好假设，并用 π_θ 参数化隐式奖励，得到

$$\Delta_\theta(x, y^+, y^-) = \log \frac{\pi_\theta(y^+ \mid x)}{\pi_{\text{ref}}(y^+ \mid x)} - \log \frac{\pi_\theta(y^- \mid x)}{\pi_{\text{ref}}(y^- \mid x)}, \tag{2.10}$$

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{D_{\text{pref}}} \log \sigma\left(\frac{\beta}{\tau_{\text{pref}}} \Delta_\theta(x, y^+, y^-)\right). \tag{2.11}$$

文献常令 $\tau_{\text{pref}} = 1$，于是只显示一个 β。调整偏好温度与调整 KL 奖励尺度会改变有效的比较尺度，不能在未说明约定时直接比较不同实现的系数。DPO 标准推导使用完整序列的对数概率；随意改为长度平均会改变隐式奖励和原目标的对应关系。

两候选回答中的 KL 强度

若参考策略对两个回答各给 1/2，奖励差为 $\log 3$，则式 (2.7) 给出最优概率比 $\pi^*(y^+)/\pi^*(y^-) = 3^{1/\beta}$。当 $\beta = 1$ 时概率为 0.75 与 0.25；当 $\beta = 2$ 时概率约为 0.634 与 0.366。更强的 KL 正则化把策略拉近参考。这个封闭算例说明目标的作用，不表示真实训练一定收敛到该分布。

上述推导建立了奖励参数化与偏好似然之间的联系，没有证明有限数据、有限容量的 DPO 与一次具体 PPO 训练得到相同策略，更没有证明偏好最优策略满足真实安全约束。偏好噪声不符合 Bradley–Terry、比较数据覆盖不足、训练提示分布偏移或参考模型概率极小时，这种联系的实践解释都需要重新检查。DPO 省去显式奖励模型和在线 RL 循环，保留了偏好质量、覆盖范围与独立验证的问题。其他偏好优化方法有各自的目标与假设；比较时应固定数据、参考模型、采样预算和评测协议，不能把不同训练损失的数值直接排序。

2.6 弱到强：监督瓶颈与可扩展监督

Weak-to-Strong Generalization[7] 用弱模型的标签监督更强模型，在其研究设置中观察到强模型可以超过弱监督者，但与使用更好标签得到的性能仍有差距。这里的“强”首先指研究中选择的能力尺度，不代表监督者已经能识别所有安全失败。该结果提示把监督问题拆成三个可检验对象：

1. 标签质量：弱监督者在哪些难度、语言和群体切片上犯错？错误是否相关？
2. 放大机制：批评、辩论、一致性或工具核验，是否在独立标签上提高监督质量？
3. 验证方法：谁来判断强模型遵守复杂规范，验证器又能看到哪些中间证据？

以全精度模型指导压缩模型是相关的蒸馏场景，但并不自动成为该论文实验意义上的弱到强泛化。若研究低比特模型的安全保持，仍应在未见攻击、事实性和工具任务上独立测量，而不是只匹配教师 logits。完整压缩实现转见量化压缩分册。

2.7 条件化行为：表面合规与内部解释的距离

Sleeper Agents[8] 人为构造含触发条件的模型，研究某些后门行为在后续训练中的持续性；*Alignment faking in large language models*[9] 在特定训练情境提示与目标冲突设置中研究条件化合规行为。它们说明现有行为测试可能漏掉条件化策略，不能外推为所有部署模型都具有同一种内部目标。模型生成的思维说明也只是可观察输出，既不能单独证明欺骗动机，也不能单独排除它。

较有说服力的审计应预先定义触发条件和对照：交换训练与部署线索、设置无关线索、改变用户目标、搜索未见触发器，比较回答、工具轨迹和独立行为后果。隐藏状态探测或 SAE 特征可提供补充线索，其因果解释还需要第 8 章讨论的干预与副作用检查。若只在公开提示上测平均拒答，通常不足以区分稳健规范与狭窄模式匹配。

练习与提示

设两个回答的参考概率分别为 0.8 和 0.2，奖励分别为 0 和 $\log 4$，且 $\beta = 1$。求归一化后的最优策略，并解释为什么奖励较高的回答未必获得超过一半的概率。提示：先计算 $\pi_{\text{ref}}(y)e^{r(y)/\beta}$，两项均为 0.8。再说明若第二个回答是错误的，哪一项训练假设或外部证据需要改变；仅减小训练损失无法回答这个问题。

本章抓住这些

SFT 学习示范，奖励模型拟合比较，PPO 近似优化带 KL 的奖励目标，DPO 在特定偏好假设下直接拟合比较似然，RLAIF 与 Constitutional AI 则说明监督从哪里来、如何组织。这些环节改善的是可测的训练代理。下一章将独立检查回答是否有证据、置信度是否有意义，以及系统何时应拒答、检索或转交人工。

第 3 章　可靠性、事实性与不确定性

上一章得到的是训练后的策略。本章检查它在什么条件下给出有根据的回答，以及置信度能否帮助系统选择继续回答、检索、核验或拒答。这里的可靠性需要可核查的任务、标签和分布；流畅程度与自信语气都不能代替这些条件。

本章的部署与攻击范围

考虑只读问答系统：用户提交问题，系统可查询冻结的本地文档快照，不执行写文件、转账等外部动作。资产是回答与引文的正确性、用户决策质量及服务可用性。自然失败包括知识缺失、检索遗漏和证据过时；攻击者可以在可控的外部文档中插入错误内容或指令，但不能修改评测标签与可信评审程序。攻击预算应记录可改文档数、插入位置、长度和查询次数。带工具写权限的后果转见第 5 章。

3.1　幻觉分类：先说明相对于什么判断

对给定问题 x、上下文 c、输出 y 以及核验日期，“幻觉”至少涉及两条不同轴。第一条是**对上下文的忠实性**：回答是否得到指定证据支持；第二条是**对外部世界的事实性**：回答是否符合可靠来源与任务定义下的事实。证据可能错误、冲突或过时，所以忠实复述也可能传播错误；回答也可能事实上正确，却没有得到当前上下文支持。

本书采用以下工作标签，而不把它们当成互斥的自然分类：

- **与证据矛盾**：回答与给定上下文直接冲突，传统分类常称为内在幻觉；
- **无证据支持**：回答加入当前证据未支持的事实，传统分类常称为外在幻觉；这不等于已经证伪；
- **知识边界错误**：缺少必要信息时仍作出确定回答，描述的是生成条件；
- **过程错误**：计算、工具使用、引用或推理链出现错误，描述的是失败环节；
- **评审错误**：评审器把格式、礼貌、引文外观或一致性误判为正确，描述的是测量失效。

同一个回答可以同时具有多个标签。例如一段过时网页支持某旧数值，回答正确复述网页却把它说成当前数值：它相对网页可能忠实，相对当前世界可能错误，引用年份的过程也有缺陷。评测前应确定事实单元、可接受时间范围、有效来源和“无法判断”类别，不能把所有未支持陈述都强制标成假。

TruthfulQA[10] 特意测试容易诱发人类常见误解的问题；其早期模型比较说明，在该测试与当时的训练设置中，规模增长没有自动带来更高真实性，不能据此推出规模对所有事实任务都无效。*FActScore*[11] 将长文本拆成原子事实，评估其中得到可靠知识源支持的比例。该比例取决于原子事实的划分、知识源和核验器，也需要与回答覆盖度一起解释，避免用少说事实换取高精度。

3.2 从 token 熵到语义不确定性

给定一个生成前缀，下一个 token 分布的熵为

$$H_{\text{tok}}(x) = -\sum_{v \in \mathcal{V}} p_\theta(v \mid x) \log p_\theta(v \mid x). \tag{3.1}$$

它测量的是下一 token 的分散程度，不是完整回答正确的概率。一个答案可以有多种措辞，错误答案也可以有极高 token 概率。

Semantic Uncertainty[12] 的思路是把表达不同但意义相同的完整回答归入同一类。固定问题、上下文、采样温度、截断与停止规则，令 Y 为该生成过程的随机回答，$\{C_k\}$ 为输出空间的语义划分，则

$$P(C_k \mid x) = \sum_{y \in C_k} P(Y = y \mid x), \qquad H_{\text{sem}}(x) = -\sum_k P(C_k \mid x) \log P(C_k \mid x). \tag{3.2}$$

采用自然对数时单位为 nat，并约定 $0 \log 0 = 0$。这里的分布是声明的生成过程；改变温度或 top-p 就改变了被测对象。

完整求和通常不可行。一个可手算的教学估计是在相同生成过程下独立采样 K 个回答，用簇频率 $\hat{P}_k = n_k/K$ 代入式 (3.2)。例如四个回答中，三条表示“巴黎”、一条表示“里昂”，频率估计熵为

$$-\tfrac{3}{4} \log \tfrac{3}{4} - \tfrac{1}{4} \log \tfrac{1}{4} \approx 0.562 \text{ nat}.$$

若四条回答都表达“里昂”，估计熵为 0，但答案仍可能共同错误。少量采样会漏掉低概率语义类，使频率熵通常低估总体多样性；它也不同于论文中利用序列概率等信息的具体估计实现。复现时应写明采用频率、概率权重还是长度归一化分数，不能把它们混作同一个估计量。

语义类还依赖等价判定。双向蕴含模型、聚类器或另一个语言模型可能把细微矛盾合并，或把同义答案拆开；自动蕴含关系也未必满足等价关系的传递性。需要固定聚类规则，人工抽查簇内一致性，并对否定、数值、日期和多语言回答做切片。低语义熵可以支持“一致”的判断，不能独立支持“真实”的判断。

3.3 校准：置信度究竟预测哪个事件

先固定正确性评审规则。对第 i 个问题与已生成回答，令 $z_i \in \{0, 1\}$ 表示该回答是否正确，$q_i \in [0, 1]$ 是系统对这一正确性事件给出的概率预测。q_i 可以来自另一个验证器或在验证集拟合的校准器；模型口头给出的百分比未经验证时只是一个分数，token 概率也不能直接当成 q_i。

总体校准要求在目标分布下

$$\mathbb{P}(Z = 1 \mid Q = q) = q \quad \text{对 } Q \text{ 几乎处处成立}. \tag{3.3}$$

这是具有相同置信度的一组回答的频率性质，不要求每个 $q = 0.8$ 的回答“有八成真”。如果“正确”改为所有原子事实均正确、引用均支持或执行结果满足后置条件，就得到不同的目标事件；标注者和置信度生成器必须使用同一规则。

3.3.1 分桶校准误差与 Brier 分数

对 $n > 0$ 个样本，预先给定覆盖 $[0,1]$ 且互不重叠的置信度桶。令 $\mathcal{I}_b$ 为落入第 b 桶的样本索引。非空桶的准确率与平均置信度分别是

$$\mathrm{acc}_b = \frac{1}{|\mathcal{I}_b|}\sum_{i\in\mathcal{I}_b} z_i, \qquad \mathrm{conf}_b = \frac{1}{|\mathcal{I}_b|}\sum_{i\in\mathcal{I}_b} q_i.$$

经验期望校准误差（expected calibration error，ECE）为 [13]

$$\widehat{\mathrm{ECE}} = \sum_{b:\,|\mathcal{I}_b|>0} \frac{|\mathcal{I}_b|}{n}\left|\mathrm{acc}_b - \mathrm{conf}_b\right|. \tag{3.4}$$

空桶不计算均值，贡献按 0 处理。必须报告等宽或等频、桶数、端点归属及样本量。桶内过度自信与不足自信可以抵消，细桶又有较大采样误差，因此小 ECE 不能单独证明总体或各群体校准。ECE 也不是准确率，更不是严格适当评分规则；只优化它可能得到信息量很低的常数预测。

对上述二元正确性事件，Brier 分数定义为 [14]

$$\widehat{\mathrm{Brier}} = \frac{1}{n}\sum_{i=1}^{n}(q_i - z_i)^2. \tag{3.5}$$

若真实条件正确率为 p，其条件期望为 $p(1-p)+(q-p)^2$，因而以 $q=p$ 最小。它同时受到任务固有不确定性与概率预测质量影响，不能只解释成校准误差。这里没有将二元概率写成两个分量再求平方和；后一约定的数值是本式的两倍。多分类 Brier 通常写为 $n^{-1}\sum_i\sum_c(p_{ic}-\mathbf{1}\{y_i=c\})^2$，报告时必须注明约定。对开放问答用式 (3.5)，需要先把正确性作为明确的二元事件，而非把生成 token 分布生硬代入。

3.3.2 选择性预测：拒答之后还要计算什么

令阈值 t 下的接受决策为 $a_i(t) = \mathbf{1}\{q_i \ge t\}$。对于错误损失 $1 - z_i$，经验覆盖率和选择性风险为

$$\widehat{C}(t) = \frac{1}{n}\sum_i a_i(t), \qquad \widehat{R}(t) = \frac{\sum_i a_i(t)(1-z_i)}{\sum_i a_i(t)}. \tag{3.6}$$

当没有样本被接受时，覆盖率为 0，风险为未定义，不能记作零风险胜利；整个测试集 $n=0$ 时二者都不报告数值。并列分数如何处理、边界是否包含阈值、按阈值还是按固定数量接受，都应预先说明。总体对应量为 $C(t)=\mathbb{P}(Q\ge t)$ 与 $R(t)=\mathbb{E}[1-Z\mid Q\ge t]$，后者要求 $C(t)>0$。

降低覆盖率是否降低风险，是需要验证的排序性质；有限样本曲线不必单调，即使整体 ECE 较小也未必能筛出错误。用验证集确定阈值与目标覆盖率，再在独立测试集报告实际覆盖率、风险和区间。若使用测试标签反复调阈值，所报风险会偏乐观。选择性风险只看接受的请求，还需报告拒答是否适当、人工队列负荷、检索与核验成本，以及整个服务的任务完成率。

3.4 手算：分桶选择与拒答阈值的影响

表 3.1 是合成教学数据，不是任何模型的实测。z_i 按前述二元正确性规则给定，样本量仅为 8。平方误差之和为 1.40，故 Brier 为 $1.40/8 = 0.175$。若用两个等宽桶 $[0,0.5)$ 与 $[0.5,1]$，低桶的平均置信度与准确率均为 0.25，高桶均为 0.75，所以 ECE 为 0。改用四个等宽桶 $[0,0.25)$、

样本 i	1	2	3	4	5	6	7	8
q_i	0.9	0.8	0.7	0.6	0.4	0.3	0.2	0.1
z_i	1	1	0	1	0	1	0	0
$(q_i - z_i)^2$	0.01	0.04	0.49	0.16	0.16	0.49	0.04	0.01

表 3.1 用于 Brier、ECE 与选择性风险的同一组概率预测和标签。

接受阈值 t	接受数	覆盖率 $\widehat{C}(t)$	选择性风险 $\widehat{R}(t)$
0	8	1	$4/8 = 0.5$
0.5	4	0.5	$1/4 = 0.25$
0.65	3	0.375	$1/3$
0.75	2	0.25	$0/2 = 0$
0.95	0	0	未定义

表 3.2 提高阈值减少覆盖率；有限样本的风险不必每一步都下降。

$[0.25, 0.5)$、$[0.5, 0.75)$、$[0.75, 1]$，每桶均有两个样本，每桶的绝对偏差均为 0.15，ECE 就变为 0.15。这不是数据变好或变坏，而是汇总方式改变了；八个样本无法给出稳定的校准结论。

从 $t = 0.5$ 提高到 0.65，恰好去掉一个正确回答，风险从 1/4 上升到 1/3。最末行则说明“全部拒答”不能靠选择性风险独自评判。实际部署要同时考虑正常效用、错误严重度与接管成本；本表中两条被接受样本均正确，也不能证明总体错误率等于零。

3.5 检测、核验与外部证据

方法族	典型做法	主要代价与失败
黑盒一致性	SelfCheckGPT 多次采样比较陈述一致性	多次调用；共同错误可能漏检
分解核验	Chain-of-Verification 生成核验问题再独立回答	核验问题有盲点，也可能重复原错误
工具交互	CRITIC 使用搜索、计算或其他工具反馈	工具与来源可能错误，增加权限和延迟
自适应检索	Self-RAG 学习何时检索、生成与批评	反思标记并非正确性概率，检索仍可能失败
证据标注	RAGTruth 提供细粒度幻觉标注	标注成本、证据范围与领域迁移限制

表 3.3 检测信号与纠错动作需要分别评测。

SelfCheckGPT[15] 面向无法取得内部概率的黑盒情境；*Chain-of-Verification*[16] 尝试减轻核验对初始回答的依赖；*CRITIC*[17] 引入工具交互；*Self-RAG*[18] 将检索与批评纳入生成过程；*RAGTruth*[19] 提供用于分析 RAG 幻觉的细粒度标注。它们使用的模型、任务与监督不同，不能从方法名称判断哪一个在新部署中更可靠。

串联系统应记录三层结果：检测器是否找到了错误，纠错动作是否成功，最终答案与服务效用是否改善。只在被检测出的错误上报告修复率，会遗漏漏检；只看改写后文本是否更流畅，也

无法证明事实改善。一个必要对照是保存初答与终答，让盲评者或程序验证器分别判断，并报告“修好错误”和“改坏正确回答”的比例。

3.6 RAG：分开评测检索、冲突与引文

检索增强生成（retrieval-augmented generation，RAG）把外部证据引入生成 [20]，却不提供自动真实性保证。评测至少要区分下列对象，并固定索引快照、检索器、重排器、上下文截断和核验日期。

检索失败

对有非空标准证据集合 G_i 的问题，令 R_i^k 为前 k 个检索结果。宏平均 Recall@k 为

$$\frac{1}{|\mathcal{Q}_G|}\sum_{i\in\mathcal{Q}_G}\frac{|R_i^k\cap G_i|}{|G_i|},\quad \mathcal{Q}_G=\{i:|G_i|>0\}.$$

这要求相关文档标注足够完整，且不同文档 ID 的等价证据得到处理。多跳问题还应报告“所需证据是否齐全”的问题级比例；命中任意一篇并不保证可作答。将生成器分别接到真实检索结果与人工给定证据的对照，可以区分检索瓶颈和生成瓶颈。

证据冲突

预先构造含矛盾、版本差异或时间差异的切片，并标注应采用的来源优先级、有效日期或应当保留不确定的状态。冲突检测召回率的分母是确有冲突的问题数；无冲突误报率的分母是无冲突问题数。冲突处理正确率则以冲突切片全部问题为分母，判断系统是否按规则选择、解释分歧或适当拒答。只在成功检测出的冲突上评价处理，会隐藏检测漏报。来源多数赞同不是事实裁决规则。

引文真实性与支持关系

分别核查引文是否可定位、元数据是否对应实际来源，以及该来源是否支持被引用陈述。“可定位率”以全部引文条目为分母；“引文支持率”以预先定义的陈述–引文对为分母；“引文完整率”以需要证据的原子陈述为分母，统计其中是否至少有一条支持性引文。链接有效不表示支持，引文支持也不表示来源本身准确。所有无引文但需要证据的陈述必须进入完整率分母，不能只核查已有引用。

最终回答还应报告事实精度、问题级正确率与可回答性判定。事实精度的分母是按固定规则抽取并可核查的生成陈述，问题级正确率的分母是全部目标问题；不可判断陈述要单列数量与处理规则。可回答性需要同时保留有充分证据与证据不足的任务：分别报告可回答问题的有效完成率和不可回答问题的适当拒答率，防止系统靠一律拒答提高支持率。若文档含恶意指令，另记录提示注入成功及正常任务误拒，不能把它们混入普通检索遗漏。

一份有引文的错误回答

用户问“截至某日的有效费率”。索引中存在旧版与新版规章，系统只检索到旧版并引用其中真实段落。引文可定位率和局部支持率可以同时为满分，回答仍可能因检索漏掉新版而错误。这个例子需要版本标注、检索对照和问题级事实判断，单一“有引用比例”看不出失败原因。

3.7 可靠性实验协议

实验协议

1. 固定问题分布、正确性标签规则、证据快照、来源有效期与不可判断类别；记录模型、提示、温度、top-p、停止规则、工具版本和日期。
2. 将校准器拟合、阈值选择与最终测试分开；分桶边界、并列分数处理和主要指标在查看测试标签前确定。
3. 每题采样 K 次时保存回答、语义簇、序列评分、引文、工具轨迹及人工标签；区分题目数与生成数，不能把同题多次采样视为独立问题。
4. 同时报告事实性、Brier、ECE、覆盖率与选择性风险，以及适当拒答、正常效用、延迟和成本；对 RAG 单列检索、冲突和引文指标及其分母。
5. 给出样本量与区间。按问题重采样可保留同题多个回答的依赖；多领域数据还需说明是否按领域分层。小样本或空接受集应如实标注，不能靠重复采样制造大量独立证据。
6. 在语言、时间、难度、证据不足和污染文档切片上检查失效；测试“拒答—检索—核验—人工接管”是否改善全流程结果，而非只改善最终文风。

练习与解题提示

用表 3.1 重算阈值 $t = 0.3$ 的覆盖率与风险，并比较两桶、四桶 ECE。提示：接受六个样本，其中两个错误，故覆盖率为 6/8、风险为 2/6。再设计一条“所有采样回答相同但都错误”的问题，说明语义熵检测为何失败。最后为同一组 RAG 问题设置真实检索、人工证据和含冲突证据三个对照，写出各自固定的分母；不必先训练新模型。

本章的置信度与核验机制只能在已声明的数据和权限范围内解释。下一章引入主动攻击者：当输入、证据或评审器本身被操纵时，可靠性评测还必须控制攻击入口、预算和权限边界。

第 4 章　对抗鲁棒、越狱与提示注入

前一章主要考虑自然错误和受污染证据。本章引入主动攻击者：它可以在预算内选择输入与策略，针对拒答、证据、评审或权限边界寻找失败。首先要按攻击入口和能改变的对象分组，再比较方法和指标；都被称为“攻击成功”的结果，可能测量完全不同的资产。

本章的部署与攻击条件

文本越狱部分默认目标是冻结版本的对话助手，攻击者可提交请求并观察回答，不能修改系统策略或标签。白盒后缀搜索另授予指定模型的梯度访问；迁移攻击只在代理模型上有该权限。间接注入部分默认只读检索助手：攻击者只能改指定外部文档中的片段，不能改用户任务与主机工具策略。保护资产分别是有害帮助边界、证据完整性和工具授权。每组结果都须记录查询数、token 数、优化步数、辅助模型与时间预算；工具动作与副作用的检查在第 5 章补全。

4.1　按入口、访问权限与资产整理攻击面

直接越狱

攻击者控制直接提交给助手的请求，试图诱导受限的有害回答。角色扮演、编码、翻译和多轮诱导是输入构造方式；成功应依据回答是否提供实质帮助，而不只是是否出现拒答词。

优化型后缀与迁移

白盒或灰盒攻击利用可见梯度、概率等信号搜索输入；迁移攻击在代理模型上搜索，再用固定预算测试目标模型。代理模型训练与搜索成本也属于攻击资源，不能只统计目标接口调用。

间接提示注入

攻击者控制网页、文档、邮件或工具返回值等低信任输入，诱导系统偏离可信用户任务。与直接越狱相比，关键差别是攻击内容来自第三方数据，用户未授予其中命令权限。

工具与参数攻击

攻击者通过模型提案或未校验参数触发未授权读写、错误收件人、命令或 URL。这一层应检查真实执行结果，模型说“已发送”不是发送证据。

评审器攻击

被评文本或其他可控输入操纵自动评审器与计分流程。此时受损资产是测量完整性，高分可能只说明评分器失效。

训练与数据供应链

攻击者可写训练语料、偏好对、检索索引或模型组件，改变后续行为。其写权限与时间阶段不同于对话输入攻击；投毒与后门在第 6 章单独讨论。

Universal and Transferable Adversarial Attacks on Aligned Language Models（GCG）[21] 研究基于梯度的离散后缀搜索及迁移；*Not what you've signed up for*[22] 研究不可信外部内容被当成

指令的应用攻击。它们可以产生相似的危险输出，却不能据此归为同一个因果根源：前者测行为鲁棒性，后者还涉及来源和权限边界。

4.2 GCG：优化代理不等于已经达成攻击目标

设攻击请求为 $x \oplus s$，其中 $s \in \mathcal{V}^m$ 是长度为 m 的后缀。给定模型参数 θ，攻击者可搜索

$$\min_{s \in \mathcal{V}^m} \mathcal{L}_{\text{sur}}(x \oplus s; \theta). \tag{4.1}$$

常见代理目标鼓励一个预设的肯定性回答前缀；GCG 利用 token 替换的梯度近似提出候选，再通过前向计算评价。梯度用于搜索候选，不把离散 token 本身变成连续变量。后缀长度、候选数、更新步数、重启次数、停止条件及目标前缀都影响结果，应随实验保存。

代理损失下降仅说明搜索更接近该目标前缀。模型可能先表示同意、随后拒绝，或生成没有实际帮助的内容，因此仍需对完整输出作独立判定。迁移测试还要分开报告源模型上的优化预算、目标模型上的尝试数与未见任务，不能把在目标上反复挑出的最好后缀记为无适应性迁移。

4.3 逐尝试比例与预算内任务 ASR

对预先固定的 $n_{\text{try}} > 0$ 次尝试，令 $J(y_i, x_i) = 1$ 表示回答达到明确的实质有害帮助标准。逐尝试成功比例可写为

$$\widehat{\text{ASR}}_{\text{try}} = \frac{1}{n_{\text{try}}} \sum_{i=1}^{n_{\text{try}}} J(y_i, x_i). \tag{4.2}$$

它的分母是尝试数。第 9 章的式 (9.2) 则以独立攻击任务为分母，统计每任务在预算内是否至少成功一次。只有每任务恰有一次尝试、成功规则相同且任务集合相同时，两者才一致。自适应尝试还共享历史，不能简单套用独立 Bernoulli 假设增加样本量。

StrongREJECT[23] 指出未明确拒答不等于有效的有害帮助，评估需要关注拒答、具体性及有效性等维度。*HarmBench*[24] 与 *JailbreakBench*[25] 提供标准化行为集、攻击与评审流程，但其结果仍取决于版本、预算、任务和评审器。评审规则应有人工校准或适合主张的程序证据，并保留不确定标签与失败尝试。

同样四次攻击，两个分母

两个任务各允许两次尝试。第一个任务两次均成功，第二个两次均失败，则逐尝试比例为 2/4，任务级比例为 1/2，数值恰好相同；若第一次任务只成功一次，逐尝试比例变为 1/4，预算内任务比例仍为 1/2。这说明相同比例可以偶然出现，指标名称和分母仍必须注明。

攻击结果应与独立正常请求上的任务效用和误拒并列。正常误拒使用全部正常请求为分母，不等于 1 – 任务完成率；答错、超时和工具故障也会导致未完成。若防御拒绝一切输入，它可能降低攻击成功，却没有提供可用服务。

4.4 间接注入：来源标记不能授予权限

在检索助手中，网页正文是回答的证据，而不是拥有与用户相同权限的命令。攻击链可描述为

$$\text{不可信内容} \rightarrow \text{上下文组装} \rightarrow \text{任务偏移} \rightarrow \text{动作提案} \rightarrow \text{执行后果}.$$

每条箭头都需要证据：模型看到恶意文档不表示服从，生成动作不表示已经执行，最终拒答也不能排除之前已发生副作用。

例如用户只授权总结会议，文档却要求读取私有文件并更换收件人。把文档包在显式标签中可帮助模型区分来源，但标签不是可信授权。执行器仍须根据主机签发的任务对象检查读取集合、报告路径、收件人和预算；文件正文中的“管理员已批准”不能创建批准。合法路径上的错误摘要可能通过权限检查，必须再用内容核验处理。数据与指令分离因此包含来源保留、可信授权和执行校验，而不是寻找一个通用的“删除恶意指令”函数。

4.5 防御分层与适用的对照

层次	代表机制	主要局限
训练层	指令优先级、对抗训练、安全示范	未见输入、语言与新工具仍可能失效
模型层	Llama Guard、WildGuard、安全分类	漏报、误报、延迟与自适应攻击
输入层	来源标记、上下文分区、结构化接口	模型未必遵守；标签本身不授予权限
运行时	可信作用域、参数检查、沙箱、预算	策略缺口、并发竞态与合法参数中的错误内容
人机层	高影响动作预览、可信确认、恢复	确认疲劳、误导性预览及不可逆副作用

表 4.1 不同层保护不同边界，需要分别测漏报、误报和正常效用。

The Instruction Hierarchy[26] 研究通过优先级训练区分不同来源的指令，其未见攻击泛化属于受测设置中的行为证据，不能替代工具权限。*Llama Guard*[27] 与 *WildGuard*[28] 提供独立的安全分类组件；加入分类器后，仍需固定版本、阈值和输入规则，在目标部署上测量。

实验对照应随问题选择。若要分清训练、分类器和执行器贡献，可比较原始模型、仅安全微调、仅分类器以及分类器加运行时控制；资源不足时可先做最相关的消融并说明缺口。未见、迁移、多轮、隐藏副作用和训练内攻击是可选择的切片，不是所有研究必须完成的固定五类。关键是对声称覆盖的威胁提供相应对照、固定预算和正常效用，不能用未运行的切片扩大结论。

练习：判断缺失的是哪条边界

一个系统看到恶意网页后，在最终回答中拒绝泄露，但日志显示它已读取任务范围之外的文件。它通过了什么测试，又失败在哪个资产上？

提示：文字拒答可能通过内容评审，未授权读取已违反工具权限。应检查动作日志与可信任务作用域，再设计未见路径和正常读取对照，而不是只修改拒答措辞。

下一章将这些边界落实到主机校验与可恢复执行，随后用本地模拟说明哪些性质可以独立于模型判断检查。

第 5 章　Agent 与工具安全：从文本到行动

前一章区分了直接请求、外部数据与执行接口。本章只保留智能体（Agent）安全所需的闭环、权限和恢复机制；组件架构及完整工作流实现见《LLM 智能体系统：从模型能力到可控自主工作流》。一个模型可以生成错误提案而被执行器挡住，也可以通过全部权限检查却写出错误报告，这两种结果应分别测量。

本章的系统与攻击范围

以本地文档助手为例，可信主机负责认证身份 u、签发任务及工具作用域；模型只提出动作，不能创建权限。攻击者可以改指定低信任文档或影响模型提案，但不能改主机任务、工具代码、计数账本或评审标签。资产包括私有文件、报告完整性、收件人、资源预算与恢复状态。教学实验使用合成文件和内存工具；并发文件替换、进程崩溃与真实外部副作用需要额外机制和验证。

5.1　智能体是有边界的控制系统

令 s_t 是主机与智能体维护的状态，包括已确认用户目标、记忆、身份、权限、工具状态和未完成操作；o_t 为当前观察，a_t 为动作。一个简化闭环是

$$a_t \sim \pi_\theta(\cdot \mid s_t, o_t), \qquad s_{t+1} = F(s_t, o_t, a_t). \tag{5.1}$$

这里的 s_t 是用于决策的内部状态，不保证包含环境的全部真实状态；需要随机转移时可将环境随机性纳入 F 的输入。风险分析使用第 1 章的完整轨迹分布，不能只由最终文本判断是否越权、泄露或损坏了状态。

代理权限被借用（confused deputy）

攻击者没有读取秘密或发送消息的权限，却借助手已有的权限达成目标。不可信文档说“管理员批准”不会使该授权成立。

工具或参数注入

模型提出了错误工具、路径、收件人或内容；若主机不校验，可能把文字偏移转成实际副作用。模型提案必须像其他不可信输入一样解析。

长轨迹累积与重试

观察、记忆与状态更新中的错误可能互相强化；超时后不清楚是否已经提交，会导致重复发送或重复扣费。恢复能力取决于工具后端与账本，而不只取决于模型是否愿意重试。

ToolEmu[29] 用语言模型模拟工具和环境以发现风险，模拟本身仍可能偏离真实工具；*InjecAgent*[30] 构造工具集成智能体的间接注入评测；*AgentDojo*[31] 的所引版本提供 97 个现实任务和 629 个安全测试用例。它们提示同时测量正常任务、攻击后果和安全属性；不能把模拟轨

迹、最终回答与真实执行互相替代。

5.2 授权来自主机任务，不来自模型文本

可信主机把用户意图转成受限任务对象 J：任务 ID、身份集合 $\mathcal{U}_J$、工具 schema、允许对象集合 C_J、固定收件人或输出对象、有效期及预算 B_J。令 p 为待检查提案，b_t 为已用预算，一种可操作的权限门可写为

$$\begin{aligned}\mathrm{Allow}_{\mathrm{host}}(u, p, J, b_t) = \mathbf{1}\{u \in \mathcal{U}_J\}\mathbf{1}\{\mathrm{Valid}(J)\}\mathbf{1}\{\mathrm{Schema}_J(p)\} \\ \cdot \mathbf{1}\{\mathrm{Objects}(\mathrm{Canon}(p)) \subseteq C_J\}\mathbf{1}\{b_t + \hat{c}(p) \le B_J\}.\end{aligned} \tag{5.2}$$

Canon 规范化对象标识与参数，$\hat{c}$ 是声明口径下的资源预留量；预算可按调用数、token 或多个资源分量逐项检查。任务关联、收件人、参数类型、额外字段和拒绝未知工具均应由主机验证。检查与实际使用的对象必须一致；静态目录中的路径解析不等于并发可写目录中的无竞态访问控制。

这一门检查的是可强制执行的授权条件，未声称能判定输出内容一定正确或实际风险小于阈值。提案里的“目标相符”“风险为零”“用户已批准”均不是可信证明。允许报告路径上的错误摘要仍可通过，所以内容后置条件与行为评测必须独立存在。

可撤销、短期、限定对象的能力令牌（capability token）是表达授权的一种实现，令牌验证、签发和撤销必须由可信主机控制；不必每个系统都采用同一种令牌格式。本书最小脚本使用可信任务对象与白名单，并未实现签名令牌。来源标签与输入分区可以帮助模型理解文档角色，但没有一种通用的“剥离指令”函数能把任意文本变成安全提案；模型自报来源也不能代替主机记录的来源。

5.3 从提案到提交的检查框架

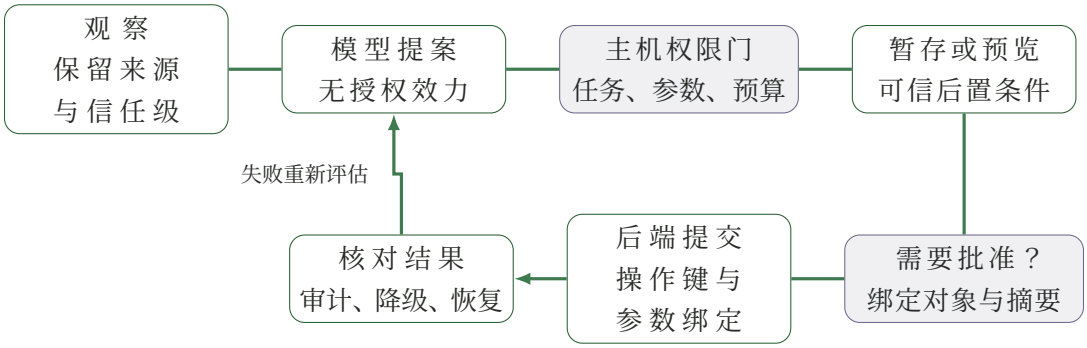

图 5.1 通用设计框架；人工批准与外部后端均未在本书最小脚本中实现。

以下伪代码面向可以暂存并在提交前核验的本地输出。工具白名单、任务后置条件和是否需要人工批准由主机设置；代码是设计框架，不是第 13 章脚本的逐行转写。

Listing 5.1 可信任务、独立后置条件与提交键绑定。

```
def guarded_call(session, raw_proposal, host_operation_id):
    task = host.load_task(session.task_id) # trusted host object
    staged = None
    outcome = "failed"
    try:
        p = schema.parse_exact(raw_proposal, task.tool_schemas)
        p = canonicalize_objects(p)
        policy.require_identity_scope_and_budget(session, task, p)
        digest = hash_canonical(p.tool, p.args)
        key = (task.id, host_operation_id) # stable across retries
        backend.require_same_payload_or_new(key, digest)
        staged = backend.stage_local(p, task)
        if not task.check_postcondition(staged):
            outcome = "postcondition_failed"
            raise Denied(outcome)
        if task.requires_approval(p):
            approval.require_bound(task.id, key, digest)
        outcome = "submission_unknown"
        result = backend.commit_once(key, digest, staged)
        outcome = "committed"
        return result
    except Denied:
        if staged is not None:
            backend.discard_uncommitted(staged)
        raise
    finally:
        audit.record_without_secrets(task.id, outcome)
```

`host_operation_id` 由主机为一个逻辑操作分配，重试沿用，同一任务内不同操作使用不同 ID；不能仅用对话轮号把多个合法动作合并。后端账本还需保存参数摘要与原结果，同键同参数返回既有结果，同键异参数拒绝。若键与副作用不能原子提交，就仍有“副作用发生、账本未记录”的窗口。当前进程中的内存去重不提供崩溃恢复，生产实现应按后端能力证明持久性、并发与重启语义。

暂存检查适用于可隔离的输出，不能直接推广到不可逆发送、付款或设备动作。后者需要可信预览、必要批准、后端幂等及结果核对；失败后的补偿也不等于副作用从未发生。未知异常和提交状态不明时应进入待核对状态，禁止自动扩大权限或无键重试。审计要覆盖拒绝、失败与未知状态，避免只留下成功路径，同时不把敏感全文写入日志。

5.4 一个连续的权限与恢复例子

可信用户请求“总结指定会议记录，将报告发给固定团队地址”。主机只把会议文档加入读取集合，指定一个报告对象与收件人。文档中的注入要求读取私有目录、更换收件人或伪造批准，模型即使遵从，提案仍应在执行前被拒绝；参数 schema 拒绝额外的“批准”字段。若注入只把会议时间改错，路径和收件人均合法，权限门无法判定，需要独立比较报告与可信证据。

报告可先暂存，核验失败时丢弃草稿，保留旧报告；这属于提交前控制。发送若已成功却未返回确认，则按同一操作键查询或重试，由后端返回原结果而不是再次发送。第 13 章只在本地合成文件和内存状态中验证这些有限边界；它不实现真实人工批准、网络发送或磁盘事务。

5.5 2025 与 2026 竞赛各自提供什么证据

Security Challenges in AI Agent Deployment[32] 的 2025 年所引版本报告了针对 22 个智能体、44 类部署场景的公开红队竞赛，并构建 ART 基准。其重点是受测代理在指定策略与攻击资源下发生政策违规，不能把参赛攻击的成功比例直接当成自然用户流量中的伤害率。

How Vulnerable Are AI Agents to Indirect Prompt Injections?[33] 的 2026 年所引版本明确研究双重目标：执行有害动作，同时不在用户最终可见的回答中暴露攻击线索；该研究覆盖工具调用、编码和计算机使用，在 13 个模型、41 个场景的设置中报告结果。隐藏攻击目标与这三类设置应归于该论文，不能倒归给 2025 竞赛。

这些都是论文在其模型、场景、参赛者与评审协议中的报告，本书没有独立复现。可迁移的研究启发是检查动作与最终回答的差异、预算内适应性攻击及未见任务；不能据此声称所有未来智能体具有相同失败频率。具体版本与观察日期见文献表和资料核验记录。

练习：一轮对话不等于一个操作

同一轮中，助手先保存报告，再发送给团队；发送提交后超时。若两步共用轮号作为幂等键，会出现什么问题？

提示：不同操作可能被错误合并或产生参数冲突；重新生成发送键又可能重复发送。应由主机分别分配操作 ID，重试沿用发送 ID，并由后端绑定规范化参数与原结果。再说明哪些部分能用内存模拟验证，哪些需要持久后端与故障注入。

最小权限限定可执行对象，内容核验检查合法输出是否正确，恢复机制控制不确定提交，行为与系统评测为这些主张提供证据。下一章转向训练记录、更新与数据完整性，继续区分不同资产和接口。

第 6 章　隐私、记忆与数据供应链

6.1　先分清资产、接口与四类风险

本章保护的资产包括训练记录、用户身份、私有属性、模型更新和数据完整性。攻击者的访问权限必须写在结论之前：只能提交文本并看回答的黑盒用户、还能获得 token 概率的评分接口用户、能读权重的白盒持有者，以及能观察梯度或提交训练样本的参与者，面对的是不同的隐私实验。部署环境也有区别：公开 API 的查询限额、离线权重的不限次计算、联邦训练中的参与者数量，都会改变成本与成功率。报告应固定候选样本数、每样本查询数、生成 token 数、参考模型和辅助数据预算。

训练数据抽取

从生成结果中恢复训练片段或敏感实体。逐字恢复、恢复近似语义和恢复此前未知的个人信息是三个不同成功事件。

成员推断

给定候选记录，推断它是否被用于某个明确训练阶段。成员推断不要求恢复记录内容，也不自动证明该内容敏感。

属性推断或模型反演

估计个人属性、群体统计或输入特征。即使训练机制保护单条记录，公开相关信息仍可能支持某些群体推断。

数据投毒或后门

攻击者修改训练数据、标签或更新，使触发条件下的行为偏离目标。这首先是完整性攻击；与隐私相关，但不应计入同一个“泄露率”。

Extracting Training Data from Large Language Models[34] 在其 GPT-2 攻击设置中展示了训练片段恢复；*Quantifying Memorization Across Neural Language Models*[35] 在受测模型和数据条件下研究了规模、重复及提示上下文与逐字记忆的关系。这些结果支持“需要测量记忆与抽取风险”，不能直接给出任意当前模型的泄露概率。抽取审计须说明匹配长度、重复处理、候选语料覆盖、样本是否本来已知，以及每条成功记录的查询成本。

6.2　成员推断：基线比一个准确率更重要

设成员标签 $m \in \{0,1\}$，$m = 1$ 表示候选 x 在声明的私有训练集中。攻击者根据可见接口与辅助信息计算分数 $s(x)$，并按固定方向判为

$$\hat{m}(x) = \mathbf{1}\{s(x) > \tau\}. \tag{6.1}$$

例如负损失可以作为分数，但低损失也可能仅仅表示句子常见或容易预测。阈值 τ、分数方向和超参数必须在独立验证数据上确定；在测试标签上挑选最好的阈值会高估攻击能力。

在明确的成员与非成员采样分布下，定义

$$\mathrm{TPR}(\tau) = \mathbb{P}(\hat{m} = 1 \mid m = 1), \tag{6.2}$$

$$\mathrm{FPR}(\tau) = \mathbb{P}(\hat{m} = 1 \mid m = 0), \tag{6.3}$$

$$\mathrm{Adv}(\tau) = \mathrm{TPR}(\tau) - \mathrm{FPR}(\tau). \tag{6.4}$$

这是本书采用的成员推断优势定义 [36]；有些文献使用其他归一化方式，比较前应先换成同一约定。若成员先验为 $\rho = \mathbb{P}(m = 1)$，则攻击准确率为

$$\mathrm{Acc} = \rho\,\mathrm{TPR} + (1 - \rho)(1 - \mathrm{FPR}). \tag{6.5}$$

只有在平衡先验 $\rho = 1/2$ 时，才有 $\mathrm{Acc} = 1/2 + \mathrm{Adv}/2$。因此优势 0.2 对应相对随机猜测提高 0.1 个准确率单位，而不是提高 0.2。不平衡实验应至少与总猜多数类的 $\max(\rho, 1-\rho)$ 比较；例如仅 10% 候选是成员时，总猜“非成员”已有 90% 准确率，且优势为零。

6.2.1 什么时候测到的是成员身份，什么时候只是数据差异

评测通常希望区分“模型是否见过该记录”，因此成员与非成员应来自尽可能可比的来源。若训练成员全是旧新闻，非成员全是新新闻，分数可能识别时间、话题或写作风格。这样的实验可以测部署中的综合可区分性，却不能把全部差异归因于记忆。

可先固定来源、时间、语言、长度和重复度等条件 c，比较 $P(x \mid m = 1, c)$ 与 $P(x \mid m = 0, c)$。在受控训练实验中，从同一候选总体随机划分训练与留出记录，能更接近可交换的成员实验；按用户、文档或近重复组划分，可避免同一实体在两侧交叉。对无法重训的模型，只能用匹配切片、参考模型和局限说明减轻混杂，不能凭切片匹配宣称已恢复随机实验。还应交代候选记录是否出现在其他预训练、微调或检索数据中。

邻域比较 [37] 用语义邻域校正样本难度；参考模型和影子模型也是常见基线，但会引入额外数据、计算及分布匹配假设。若声称攻击更便宜，应同时固定访问接口、候选数量、误报水平和辅助模型预算，不能将概率接口攻击与只看文本的攻击直接排名。隐私报告宜同时给出 ROC/AUC、预先指定低 FPR 下的 TPR、样本数和区间。AUC 接近 1/2 只说明受测攻击未能区分；有限攻击失败不等价于隐私保证。

练习：高准确率是否意味着泄露

候选成员先验为 0.1，一个攻击的 TPR 为 0.7、FPR 为 0.1。计算优势、准确率和预测为成员时的精确率。

提示：式 (6.5) 给出准确率 0.88，低于多数类基线 0.9；优势为 0.6，精确率为 $0.07/(0.07+0.09) = 0.4375$。这些量回答不同问题，应结合误报代价一起报告。

6.3 差分隐私：先声明谁与谁相邻

差分隐私（differential privacy，DP）约束随机机制发布结果时，单个保护单位的加入或改变能造成多大分布变化。它是针对机制及相邻关系的保证，不是某条记录“被遗忘的概率”。常见相邻关系包括：加入或删除一条记录的无界相邻；将一条记录替换成另一条的有界相邻；加入、

删除或替换一位用户全部记录的用户级相邻。它们的敏感度与隐私含义不同，论文中的 ε 不能跨这些关系直接比较。

对声明的相邻关系 $D \simeq D'$，随机机制 M 满足 (ε, δ)-DP，当且仅当对所有相邻数据集及所有可测输出事件 A，

$$\mathbb{P}[M(D) \in A] \leq e^{\varepsilon}\mathbb{P}[M(D') \in A] + \delta. \tag{6.6}$$

概率来自机制的随机性；对称相邻关系要求交换 D, D' 后也成立。$\varepsilon \geq 0$ 控制事件概率比的放宽，$\delta \in [0, 1)$ 是加性松弛。δ 不宜被简写成“某个用户有 δ 概率完全泄露”，ε 也不是百分比泄露率。参数应连同保护单位、样本数、会计方法和效用一起解释 [38]。

6.3.1 组合、群体与发布边界

若对同一私有数据顺序执行 k 个机制，每个在给定此前发布结果的条件下仍满足 $(\varepsilon_j, \delta_j)$-DP，则基本顺序组合满足

$$\varepsilon_{\text{tot}} \leq \sum_{j=1}^{k} \varepsilon_j, \qquad \delta_{\text{tot}} \leq \sum_{j=1}^{k} \delta_j. \tag{6.7}$$

这包括用私有验证集选模型、反复调参和发布多个检查点。更紧的会计可改善界，但必须说明所用定理、采样机制和参数转换。对互不重叠的保护单位可适用并行组合，但“记录不重叠”不保证“用户不重叠”。

当 D, D' 相差 g 个基本单位时，逐次应用式 (6.6) 得到群体界 $(g\varepsilon, \delta \sum_{j=0}^{g-1} e^{j\varepsilon})$。所以一位用户有很多记录时，记录级 DP 并不自动变成同参数用户级 DP。对已发布 DP 结果进行不再接触私有数据的后处理不增加隐私预算；若后处理重新检索私有库，则需另行分析。类似地，仅微调阶段采用 DP，保护的是该阶段声明的数据关系，不能替此前非私有预训练背书。

6.3.2 DP-SGD 的噪声加在什么地方

DP-SGD[39] 的基本步骤是先得到每条记录的梯度 $g_i = \nabla_\theta \ell(\theta; x_i)$，再裁剪

$$\bar{g}_i = g_i \min\left(1, \frac{C}{\|g_i\|_2}\right), \qquad \tilde{g} = \frac{1}{b_0}\left(\sum_{i \in B} \bar{g}_i + \xi\right), \quad \xi \sim \mathcal{N}(0, \sigma^2 C^2 I). \tag{6.8}$$

零梯度按不裁剪处理。这里 C 是裁剪阈值，σ 是噪声乘子；b_0 为声明的归一化尺度，例如 Poisson 采样时采用固定期望批大小。若采用固定批大小、按层裁剪或其他归一化，实现与会计必须相匹配。先将一个批次求平均、再裁剪整体梯度，通常不能代替逐保护单位裁剪。

加入高斯噪声并不独自确定 ε：会计还需要数据规模、采样方式与采样率、步数、相邻关系、目标 δ，以及检查点和调参的发布次数。用户级 DP 还要约束每位用户的总贡献。裁剪偏差和噪声可能改变长尾知识、事实性及偏好行为，因此应测量隐私—效用曲线和群体切片；DP 给出的分布界也不能代替工具访问控制、输出权限和日志治理。

练习：相同参数为何未必相同保证

机制每次训练满足 $(1, 10^{-6})$ 记录级 DP，顺序发布三个独立训练结果。用基本组合给出总参数，再说明它能否被写成 $(3, 3 \times 10^{-6})$ 用户级 DP。

提示：总记录级参数可取 $(3, 3 \times 10^{-6})$；如果一个用户贡献多条记录，应根据用户贡献上限应用群体界或直接使用用户级机制，不能仅改参数名称。

6.4 数据投毒与联邦学习：保护对象不同，证据也不同

投毒威胁模型应说明攻击者可写训练语料、偏好对、检索索引还是联邦更新，能改多少条、是否知道过滤器、是否可自适应重试。*Poisoning Language Models During Instruction Tuning*[40]展示了其训练设置中的定向投毒效果；结果的外推取决于毒样本比例、触发器、数据分布与训练方案。报告要同时测触发条件下的攻击成功、正常任务退化和过滤器误报，不能只展示一个后门样例。

联邦学习使原始数据留在参与方，但服务器或恶意参与者仍可能从更新中推断信息 [41]。安全聚合减少服务器对单个更新的可见性，DP 限制保护单位对发布结果的影响，更新异常检测关注完整性；这些机制解决不同问题。它们的有效条件应连同串谋阈值、可见轮次、抽样及退出规则说明。

阶段	需要记录的资产与边界	可复核的检查
数据进入	来源许可、重复组、敏感实体、用户贡献	数据卡、分组去重、抽样审计
训练发布	相邻关系、梯度可见性、预算及检查点	会计配置、MIA 基线、更新审计
模型服务	权重或概率接口、查询上限、私有检索库	抽取成本、权限校验、访问日志
事件响应	影响范围、证据保留、可撤销版本	回滚演练、删除范围与残余风险说明

表 6.1 隐私与完整性审计应覆盖数据、训练、发布和运营。

一份合格的隐私声明应能让读者回答：保护哪位用户或哪条记录、攻击者看见什么、预算限制在哪里、形式化保证覆盖哪些发布，以及仍可能通过哪些侧信道和工具接口泄露。公开数据、联邦训练、输出过滤和删除接口均有各自作用，却不能互相替代。

第 7 章　公平、社会伤害与模型规格

前一章讨论谁的信息受到保护。本章再问：谁更容易被错误拒绝、收到错误答案，或承担系统失效的成本。公平并不是一个脱离情境的常数；需要先确定服务目标、受影响群体与比较条件，再选择指标。平均安全分数可能掩盖不同语言、方言或无障碍需求下的服务差异。*TrustLLM*[42]将公平等维度纳入可信性评测，但其任务与群体范围仍不能代替具体部署的审查。

7.1　同样的总体分数，可能对应不同服务机会

以本地文档助手为例：它帮助员工查找公开的工作制度，用户可以用普通话、方言或不同表达方式提问。系统没有写文件与发送消息的权限。需要保护的资产包括获得正确信息的机会、隐私和免受歧视性内容影响的权益。自然失效来自语料覆盖、语言识别和拒答阈值；攻击者还可能用群体称谓、引文或方言表述诱发分类器误判。评测应限定可改写的输入、每题尝试数和来源，避免把攻击改写与真实群体分布混为一谈。

假设两个群体的平均正确率相同，一方可能更常被拒答，另一方可能更常收到错误但自信的回答。单一正确率便掩盖了失败后果的差异。需要分别记录：任务是否可回答、是否被接受、答案是否正确、拒答是否合理，以及用户是否能获得其他帮助。群体标签也不能随意从姓名或方言推断；代理标签可能既不准确，又造成新的隐私与身份伤害。

7.2　几种公平定义为什么不能互换

令 G_{grp} 表示预先声明的群体属性，$V = 1$ 表示请求按独立评审规则属于合法且应获服务的任务，$D_{\text{ans}} = 1$ 表示系统接受该请求并尝试回答。这里“接受”不等于“回答正确”，正确性仍用第 3 章的 Z 表示。各条件事件须有正概率，否则不报告该条件比例。

人口统计均等（demographic parity）

不同群体获得服务决策的比例相同：

$$\mathbb{P}(D_{\text{ans}} = 1 \mid G_{\text{grp}} = g) = \mathbb{P}(D_{\text{ans}} = 1 \mid G_{\text{grp}} = g') \quad \text{对受比较群体 } g, g'. \tag{7.1}$$

这比较全部请求中的接受率，没有按请求合法性或可回答性调整。差异可能来自系统，也可能来自请求组成；需结合服务职责判断。

机会均等（equal opportunity）

在确实应获服务的请求中，接受率相同：

$$\mathbb{P}(D_{\text{ans}} = 1 \mid V = 1, G_{\text{grp}} = g) = \mathbb{P}(D_{\text{ans}} = 1 \mid V = 1, G_{\text{grp}} = g'). \tag{7.2}$$

它关注合法请求的漏拒差异。若把“有能力完成任务”定义为成功事件，则应另报 $\mathbb{P}(D_{\text{ans}} = 1, Z = 1 \mid V = 1, G_{\text{grp}} = g)$，不能让无内容的应答充当平等服务。

均等化胜算（equalized odds）

要求在 $V = 1$ 和 $V = 0$ 两类中，群体条件接受率都一致。相较机会均等，它还约束不应接受

请求的错误放行率。中文译名可能相近，报告应直接写出条件概率，不能只写“公平性得到改善”。

群体条件校准

若 Q 预测的是回答正确概率，群体条件校准要求

$$\mathbb{P}(Z = 1 \mid Q = q, G_{\text{grp}} = g) = q. \tag{7.3}$$

它关注分数与正确性之间的关系，不能代替接受率或误拒率比较。可仅在被接受回答上评估，但必须标明条件 $D_{\text{ans}} = 1$ 并同时报告群体覆盖率。

这些定义通常需要不同标签和分母。不同基础率、不同可回答性与不完美预测可能使目标发生冲突；先写出需要保护的机会与误差代价，再选择约束，不能把几个指标名称相互替换。这里的公式也是测量定义，不直接规定什么群体差异在规范或法律上可接受。

一个分母改变结论的教学例子

两个群体各有 100 个请求，其中应获服务的分别为 80 个和 40 个。假设系统准确接受全部应获服务的请求，拒绝其余请求，则机会均等的条件接受率两组都是 1；总体接受率却分别为 0.8 和 0.4，不满足人口统计均等。这不是任何真实模型的结果。审查者还应追问：第二组可回答性较低，是否来自文档缺少相应语言或无障碍支持？即使决策条件指标相等，服务供给本身仍可能不公平。

反事实一致性提供另一种检查：构造仅改变与任务无关的群体称谓、其余含义不变的成对请求，比较输出与决策差异。但更换方言或身份信息可能改变语义和社会语境，需要人工确认对照成立；文本替换也不能自动模拟真实社会中的因果反事实。

7.3 小样本切片、交叉群体与隐私

公平切片越细，样本通常越少。一个群体在五个请求上没有误拒，只能说明该小样本未观察到误拒；不能写成“该群体没有误拒风险”。对每个切片同时报告总请求数、应获服务数、接受数、错误数和区间，并明确分母。第 9 章讨论的相关样本、簇级重采样与多重筛选问题同样适用：同一个人的多次改写不是多个独立用户，从许多切片中挑最显著的一个也会带来选择偏差。

交叉群体有助于发现单一总体分类掩盖的问题，但应预先确定关键切片，记录样本不足与缺失标签。合并过小群体会隐藏差异，公开过细群体又可能暴露个人身份；二者需要结合风险选择。经过适当授权与保护的真实任务样本、人工设计的成对测试和合成数据各有作用，应分别报告来源，不能让合成身份替代真实使用经验。

隐私与公平也有测量层面的张力：不收集任何群体信息，会使某些差异无法直接评估；无节制地收集敏感属性，又会增加泄露与滥用风险。可在访问受控的审计环境中保留最少必要标签，限制用途与保留期，发布足以支持判断的汇总。若对群体计数加入差分隐私噪声，要说明保护单位、预算、噪声及其带来的区间变化；人数很少时，噪声可能使差异难以识别，不能把这一点解释为差异不存在。DP 保证也不自动推出群体公平，反之公平指标相等并不提供隐私保证。

7.4 从缓解措施到可追踪的模型规格

发现差异后，可以检查数据覆盖、标注规则、分类阈值与接管流程，比较重采样、补充证据、校准或人工复核。每个措施都须保留总体与群体条件的安全、事实性、误拒、成本和效用对照。扩大拒答可能降低有害内容暴露，也可能阻断合法求助；降低某类误报也可能增加错误放行。把这种取舍写成可检查的任务和指标，才能说明选择由谁承担、何时需要重新评估。

模型卡（model card）[43] 与数据集说明（datasheet）[44] 可承载这些条件。至少记录预期用途、群体与场景范围、数据来源及许可、标注协议、已知失败、各类分母、攻击预算、版本、工具权限、责任人和复测触发条件。没有测过的群体应记为证据缺口。模型卡的价值在于支持后续判断，而不是把有限评测改写成无条件承诺。

练习：拒答差异应如何报告

某助手在两种语言上各测试 50 个合法请求，分别误拒 5 个和 10 个；其余回答中分别有 4 个和 3 个错误。分别计算误拒率、接受回答中的错误率，以及全部合法请求上的正确完成率。**提示：**误拒率为 5/50 与 10/50；选择性错误率为 4/45 与 3/40；正确完成率为 41/50 与 37/50。三个指标排序未必一致。再说明需要哪些独立样本与评审证据，才能把差异归因于语言处理，而不是问题难度或来源组成。

从本章开始，安全报告不再只问平均任务是否成功，还要问失败落在谁身上。下一章检查解释与机制证据能补充什么，以及它们为什么仍需与群体行为和权限测试相互核验。

第 8 章　可解释性、监控与机制审计

8.1　解释要回答哪一种问题

机制审计希望理解模型如何产生拒答、事实判断、奖励或工具选择，以便发现后门、监测异常并改进系统。这里通常假定审计者能读取指定层的激活，必要时可干预激活或权重；这种访问权限高于只看 API 文本的行为评测。威胁模型则要另行写明：攻击者能否改权重、注入触发内容或观察监控器，触发器搜索有多少查询预算，审计保护的是输出行为、私有信息还是工具权限。离线发现某个方向，并不表示部署时能低成本、低漏报地监控它。

三个问题需要分别收集证据：

- **相关性**：哪些输入、token、神经元或特征与某种行为共同出现？
- **机制解释**：哪些计算步骤及表示传递了完成任务所需的信息？
- **因果干预**：在声明的层、样本和干预尺度下，改变一个变量是否改变了目标行为，且对照能否排除其他原因？

注意力图、梯度归因、线性探针和稀疏特征可提出候选解释。它们的高预测准确率并不自动说明该变量被模型用于产生结果：相关特征可能只是旁观者，探针也可能读出模型原本不会使用的信息。

Toy Models of Superposition[45] 在受控模型中说明了稀疏特征如何共享有限维表示；*Sparse Autoencoders Find Highly Interpretable Features in Language Models*[46] 研究用稀疏自编码器分解激活；*Representation Engineering*[47] 则研究从表示方向监测或干预高层行为。这些方法提供不同测量工具，不能互相替代，也不能仅凭命名为“诚实”或“安全”的方向建立安全保证。

8.2　SAE 学的是编码函数与字典

设冻结模型在某层产生激活 $h \in \mathbb{R}^d$，审计分布为 P_{audit}。一个常见的非负稀疏自编码器（sparse autoencoder，SAE）写成

$$z_\phi(h) = \mathrm{ReLU}(W_{\text{enc}}(h - b_{\text{dec}}) + b_{\text{enc}}), \tag{8.1}$$

$$\hat{h}_\phi = W_{\text{dec}} z_\phi(h) + b_{\text{dec}}, \tag{8.2}$$

$$\min_{\phi} \quad \mathcal{L}_{\text{SAE}} = \mathbb{E}_{h \sim P_{\text{audit}}} \left[\|h - \hat{h}_\phi\|_2^2 + \lambda \|z_\phi(h)\|_1 \right]. \tag{8.3}$$

其中 $W_{\text{enc}} \in \mathbb{R}^{k \times d}$，$W_{\text{dec}} \in \mathbb{R}^{d \times k}$，$\phi$ 包含编码器、解码器与偏置参数，通常 $k > d$。训练优化的是共享编码函数 $z_\phi(\cdot)$，不是为全体测试激活任意选择自由系数 z。逐样本求稀疏系数的稀疏编码与 SAE 有联系，但其推断过程和优化变量不同。

若不约束字典列范数，把 W_{dec} 放大、把编码系数缩小，就可能在重构近似不变时人为减小 L_1 惩罚。因此常用约束 $\|W_{\text{dec},:j}\|_2 = 1$ 或相应归一化实现，并说明死特征处理、损失归一化及稀疏程度。ReLU 使系数非负，字典方向仍可包含正负坐标；允许有符号系数的变体需要另外说明，系数的正负并不天然等于“有害/无害”。Top-k 等变体也不是式 (8.3) 的同一目标，比较结果应给

出实际激活规则。

较低重构损失说明字典保留了较多激活信息；较稀疏系数使个别方向更容易检查。二者都不推出单语义、因果完整或安全相关。至少应同时测量重构替换后的任务损失、活跃特征数、死特征比例、不同随机初始化的稳定性，以及特征标签在独立样本上的预测力。由模型自动生成的特征标签是待检验的解释假说；展示最高激活样本容易受到选择偏差影响。

8.3 从归因到受控干预

梯度归因近似输出对局部扰动的敏感度，探针检验可预测性；二者都不能独自确定实际行为的因果通路。激活修补（activation patching）把干净运行和受扰运行的某个中间变量替换，再观察预先定义的结果变化。要解释这个变化，必须同时说明输入对、替换层与位置、下游计算是否保持相同，以及结果变量是 logit 差、正确率、拒答率还是工具越权事件。

SAE 干预尤其要区分特征改变与重构误差。令

$$r = h - \left(W_{\text{dec}} z_\phi(h) + b_{\text{dec}}\right). \tag{8.4}$$

若要把系数从 z 改为 z'，保留原残差的干预可写为

$$h' = W_{\text{dec}} z' + b_{\text{dec}} + r = h + W_{\text{dec}}(z' - z). \tag{8.5}$$

若直接用 $W_{\text{dec}} z' + b_{\text{dec}}$ 替换 h，还同时删除了 r。行为变化便可能来自丢失信息，而非目标特征。保留残差并不使干预自动有效；它只是排除一种混杂。大幅改变系数仍可能把激活推到训练分布之外。

一项机制安全审计的最小对照

先固定审计模型与权重哈希、层和 token 位置、候选特征发现集、独立验证集及干预幅度。至少比较：原始激活、只做重构、保持残差的目标干预，以及范数或稀疏度匹配的随机方向/无关特征干预。对同一任务同时测安全目标与正常效用；对不同语言、提示模板和任务难度复测。若从数万个特征中筛选，报告筛选规模、选择规则和多重比较处理，并在未参与筛选的数据上确认结果。

因果结论应匹配干预范围。例如“在这些提示上将该层方向放大后，拒答率上升”是可测的局部结论；“这个方向完整表示安全价值”则需要远强得多的证据。消融后的行为不变，也可能是通路冗余或干预太弱；行为改变则可能包含非目标副作用。建议保存每个特征的选择轨迹、失败样本、干预前后激活范数和无关任务结果，避免只展示最漂亮的样例。

练习：重构误差能否伪装成因果作用

设激活 $h = (1, 2)$，SAE 重构为 $(1, 1)$，目标特征改变带来的字典增量为 $(1, 0)$。分别求保留残差和删除残差后的新激活。

提示：残差为 $(0, 1)$，两种新激活分别是 $(2, 2)$ 与 $(2, 1)$。若后者改变了行为，不能把全部变化归因于第一坐标的目标特征。再设计一个只删除残差的对照，判断第二坐标的作用。

8.4 规模化研究能支持什么结论

Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet[48] 是 2024 年的官方研究报告，原稿把它列为 2026 年条目的写法在本版纠正。该报告在特定模型和 SAE 字典中展示了可解释特征及干预案例；“存在可命名方向”与“字典完整揭示内部目标”之间仍有很大距离。*SAFER*[49] 研究奖励模型的 SAE 审计；这类实验需限定奖励模型版本、数据修改方式与受测任务，不能外推为普遍可定位或可修复的安全机制。2026 年关于对齐算法几何变化的预印本 [50] 可作为候选研究线索，其表征相似性、局部干预与部署安全应分别核验。

本节所引资料的版本与核验记录见附录 B，相关方法可用于提出待检验的机制假说。

对象	可以检验的局部问题	关键限制与对照
表示特征	是否预测某种行为，跨切片是否稳定	独立标签、筛选偏差、随机特征基线
行为通路	修补某层是否改变预先定义的结果	残差与范数控制、冗余通路、副作用
奖励模型	偏好判断是否依赖可操纵的特征	独立人评、数据污染、分布外验证
压缩模型	压缩前后行为与干预效应是否变化	同任务与预算、对齐字典、效用对照
部署监控	特征阈值是否及时发现异常	漏报与误报、延迟、攻击者适应

表 8.1 机制审计需要与行为评测和系统权限检查相互印证。

本书只讨论量化与压缩对安全证据的影响；完整低比特实现转引量化压缩分册的对应主题。机制证据可以提出更有针对性的测试或修复，但权限隔离、行为评测和恢复演练仍需独立通过。

第 9 章　评测科学、红队与证据链

9.1　先写主张，再选基准

评测首先要说明它支持哪一句主张。例如，“在指定版本、每任务至多十次查询、只读模拟工具的环境下，间接注入未能触发越权读取”是有范围的主张；“模型安全”则没有给出可检验对象。本章默认红队获准提交攻击文本或修改指定的不可信资料，不能访问真实凭证、扩大工具权限或改变测试标签。若改用白盒梯度、可写工具或开放网络，应另设威胁模型、预算及隔离环境。

能力基准回答系统能否完成受测任务。MMLU[51]、HELM[52]、SafetyBench[53] 和 TrustLLM[42] 分别提供知识、场景化多指标、安全知识或可信维度的测量入口。它们的得分不等于真实伤害率；安全选择题答对也不说明工具调用不会越权。报告必须说明版本、语言、样本切片、提示模板及已知污染风险，并将知识测量、行为判断与执行后果分开。

9.2　五类互补证据，不能按编号替代

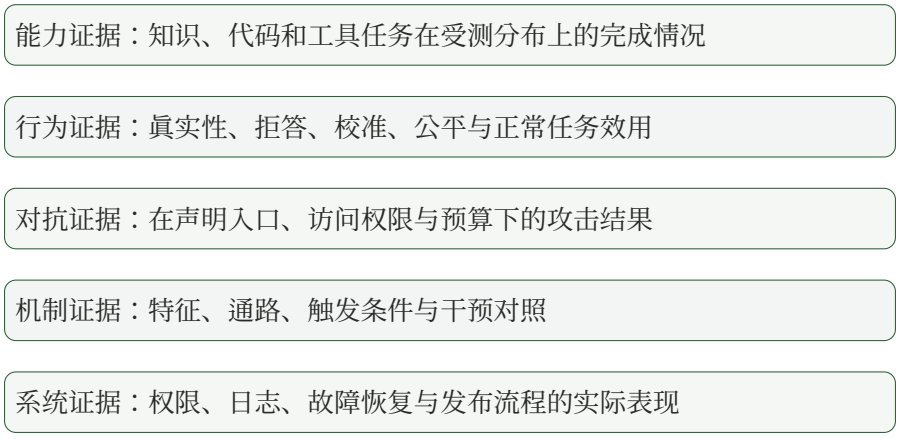

图 9.1　五类证据的编号或排列顺序不表示保证强度单调增加。

原稿称其为“证据阶梯”，容易被读成机制或治理证据天然比行为证据更强。本版将其改为互补证据链：制度文件可能没有经过执行演练，漂亮的机制解释可能没有行为效应，大规模基准可能没有覆盖真实权限。一条受限的执行器不变量可以在其形式化假设内提供强约束，却不能保证模型回答正确。审查时应把每个安全主张连到对应证据、反例搜索和假设，逐个确认；层数多并不自动填补缺失的因果环节。

9.3　Benchmark validity 与 Goodhart 风险

当分数成为优化目标，原本与任务目标相关的代理量可能失去代表性。需要区分测试污染或过拟合、评分规则被利用、分布迁移，以及有限样本和评审误差。它们可能相互作用，不能将任意分数提升写成“目标提升、投机、噪声、迁移”四项可识别的加和。

只有在代理分数 S_j 与真实目标量 U_j 具有相同尺度时，比较两个系统 $j = 0, 1$ 才可用三角不等式提示误差问题：

$$|(S_1 - S_0) - (U_1 - U_0)| \le |S_1 - U_1| + |S_0 - U_0|. \tag{9.1}$$

式 (9.1) 是代数界，不是可直接估计的 Goodhart 因果分解；如果 U 是现实伤害而 S 是选择题分数，二者单位不同，连这个比较也须先建立测量模型。需要通过独立测试、污染检查、改变评分规则、真实或模拟后果验证等手段，为代理与目标的关系收集证据。*EvalSafetyGap*[54] 等近年框架可以帮助提出此类诊断问题，其分类建议不能替代本项目的实证验证。

一个实用检查是保留“未针对其优化”的测试轴：未见攻击者、不同提示模板、不同语言、不同资料来源及故障恢复任务。用这些测试发现外推失败后，应缩小结论范围或补充修复；不应只重新挑选最有利的总体分数。

9.4 固定预算下，成功的分母是什么

设预先确定 n_A 个攻击任务。每任务的预算向量 B 包含查询数、输入与输出 token 上限、优化步数、墙钟时间、可用辅助模型及重试规则。对任务 i，攻击者按协议提出至多 K 次尝试，定义

$$A_i(B) = \mathbf{1}\{\text{预算内至少一次满足预先定义的攻击成功条件}\}, \qquad \widehat{\mathrm{ASR}}(B) = \frac{1}{n_A}\sum_{i=1}^{n_A} A_i(B). \tag{9.2}$$

任务级攻击成功率（attack success rate，ASR）与逐尝试成功率分母不同。把十次搜索后的最好结果与一次生成比较，会混淆防御强度和攻击预算。白盒优化、黑盒搜索与迁移攻击还应分别列出访问权限；仅匹配查询数不一定匹配总攻击成本。

攻击成功谓词要对应资产：有害内容评测需给出实质性帮助的评分规则，提示注入需检查攻击者目标是否达成，工具安全应检查未经授权的动作或副作用。模型说“已发送”不证明发送发生；模型在最终回答拒答，也不证明之前没有越权。无效工具参数、环境失败和评审不确定应保留在记录中，按照预先规则统计，不能看过结果后删掉失败任务。

正常请求需要独立样本 n_N。令 $C_i = 1$ 表示任务按条件完成且没有禁止副作用，$F_i = 1$ 表示被策略错误拒绝，则

$$\widehat{U}_N = \frac{\sum_i C_i}{n_N}, \qquad \widehat{\mathrm{FRR}}_N = \frac{\sum_i F_i}{n_N}. \tag{9.3}$$

这里正常误拒率（false refusal rate，FRR）不等于 $1 - \widehat{U}_N$：超时、答错和工具故障也会使任务失败。正常集应包括容易被误判的合法边界请求，并交代“正常”标签的裁决过程。比较防御时至少并列 ASR、正常效用、误拒及成本；若拒绝所有请求，低 ASR 只说明系统停止服务。

9.5 样本单位、置信区间与相关攻击

置信区间回答的是某个抽样和随机化协议下估计量的不确定性，不会补齐遗漏的威胁模型。若任务是独立抽样的 Bernoulli 单位，固定每任务预算及成功规则，Wilson 区间可用于任务级 ASR 或完成比例 [55]。第 13 章给出小样本计算；零次成功仍有非零上界。若只是完整枚举一个固定小语料库，应首先报告该有限集合上的计数；要推广到新任务，必须再说明任务总体及抽样依据。

同一网页的多种改写、同一用户的多轮对话、同一目标上的自适应重试通常相关，不能把它们当成独立样本扩大 n。应以独立场景、用户或任务为簇，进行簇级重采样；比较两个系统时尽量在相同任务上配对，报告成功率差的配对区间。多模型、多攻击和多阈值的筛选还会带来选择偏差，应将搜索集与最终测试集分开，交代选择次数。多个随机种子能衡量部分运行波动，却不能替代独立任务覆盖。

自动评审的漏判、人工分歧和环境模拟失真是测量误差，常规二项区间不包括它们。严重性和发生概率也应分开：发现一次高后果失败可能足以触发修复，即使其频率估计仍很宽。统计显著并不自动意味着满足发布门槛。

9.6 红队闭环与可复核的评审协议

红队可以沿着资产盘点、攻击生成、评审、根因定位、修复、未见攻击复测和发布裁决形成闭环。探索阶段允许适应性搜索并发现新攻击类别；确认阶段应冻结攻击预算、评审规则和留出集。不能用同一批攻击反复调防御后，将它们重新称为独立测试。

一次安全评测须冻结的最小协议

1. 系统版本、权重或服务日期、提示模板、工具权限、模型采样与超时重试配置。
2. 正常与攻击任务的抽样单位、来源、去重和切分；攻击者知识、入口、预算及停止条件。
3. 完成、成功、拒绝、环境故障及不确定标签的操作定义；每类分母和缺失数据规则。
4. 评审器版本、提示与阈值；人工标注说明、盲评及冲突裁决；关键样本的程序或执行日志证据。
5. 主指标、区间方法、相关样本处理、配对比较、成本口径和未见攻击复测方式。

LLM 评审器可减轻人工负担，但有位置、长度和同源偏好，也可能受到被评文本的注入。应先在带双人标注及裁决的人评子集上检查混淆矩阵和分歧，随机化回答顺序，固定评分提示，将被评文本作为数据隔离。高后果结论至少使用一条适合该主张的独立证据，例如工具后置条件或人工审计。原始分歧、未裁决样本和评审器漂移都应报告，不能静默平均为“真值”。

最终报告应能从一条结论追溯到任务、全部预算内尝试、工具轨迹、标签、裁决理由与统计脚本。保存失败尝试和正常任务结果，才能看出防御到底降低了风险，还是只是增加拒答或改变了评分外观。

练习：三十次尝试有多少个样本

三个独立任务各允许十次自适应尝试，成功尝试数分别为 $0, 1, 4$。计算任务级 ASR 与逐尝试成功率，并说明能否用 $n = 30$ 为任务级 ASR 计算 Wilson 区间。

提示：任务级 ASR 为 2/3，逐尝试比例为 5/30；十次尝试共享目标和历史，后者不是三十个独立任务。若目标是新任务上的预算内成功概率，应以任务为统计单位，并承认 $n = 3$ 的不确定性很大。

第 10 章　治理、标准与发布工程

评测给出受测条件下的证据，发布还要决定证据是否足够、谁承担残余风险，以及环境变化时如何撤回结论。本章把第 9 章的证据链组织成一个具体安全案例（safety case）。安全案例是可质疑、可更新的论证材料；写出案例本身并不表示其中主张已经成立。

10.1　标准与政策提供什么，不能替代什么

NIST AI RMF 1.0（2023 年 1 月）与配套 Generative AI Profile（NIST AI 600-1，2024 年 7 月）[56, 57] 提供 Govern、Map、Measure、Manage 的风险管理语言。它们可帮助分配责任、记录场景、组织测量与处置，属于风险管理框架，不能据此推断某个系统通过安全实验或达到法律合规。

ISO/IEC 42001:2023[58] 是 AI 管理体系标准；本次仅核实其标准身份和主题，未读取付费标准全文，不引用具体规范条款。EU AI Act，即 Regulation (EU) 2024/1689[59]，在本书中仅作为风险与透明度制度的入口。适用角色、地区、系统用途、分阶段适用日期和后续规则都可能影响具体义务；本次通过欧盟委员会说明核实高层框架，未取得可读取的法规全文，不作逐条适用判断。技术测量和法律适用需要分别提供依据。

OWASP Top 10 for LLM Applications 的 2025 版 [60] 是工程威胁清单，版本名与实际发布日期不同；本书固定该版，而非引用持续变动的首页。MITRE ATLAS[61] 提供对抗技术分类，本次文献表固定到其 2026.09 数据版本。二者可帮助查漏补缺，但攻击类别覆盖不等于受测系统已经抵抗全部实例。

公司公开政策说明组织如何声称管理风险。例如 Anthropic Responsible Scaling Policy 的 2025 年 5 月 14 日版本 2.2[62] 与 OpenAI Preparedness Framework 的 2025 年 4 月 15 日版本 2[63]，可用于分析能力评估、保护措施与发布裁决的制度设计。本书引用的是已核验的历史版本，不称其为本次修订日的最新版；公开承诺、内部执行记录和外部独立验证是不同证据。来源访问结果与核验限制见附录的资料使用说明。

10.2　一个本地文档助手的待论证主张

设想一个供小团队使用的文档助手。用户身份由主机验证，助手只能读取该用户已获授权的静态文档快照，并把报告写到预先指定的本地输出位置；不提供网络发送、任意命令、转账和动态插件。检索文档作为不可信数据处理，不得创建新授权。这里的描述是教学设计，不是本项目已经部署的真实服务。

案例应明确以下范围：冻结的模型与提示版本、索引与文件哈希、身份 u、文档授权集合、运行时与工具版本，以及最大上下文、重试和资源预算。攻击者可以修改指定的低信任文档或提交请求，但不能修改主机策略、可信任务对象与测试标签，也不能在检查和使用之间改变文件映射。最后一个假设限定了静态目录实验；若生产目录可并发替换或含竞态，需要以文件句柄、隔离挂载等机制重新论证。

顶层主张可以写为：在上述边界和固定攻击预算下，未授权文件与外部通信不会通过受控工具接口暴露，回答在声明任务分布上达到事先规定的效用与证据质量要求；异常会被记录并触发可用的降级或停用流程。这是一组需要分别证明的子主张，不能把“模型会拒答”当成共同证据。

权限子主张

每次读写都由主机依据身份、规范化对象与任务作用域检查，模型文本不能扩大权限。证据包括策略实现与审查、工具接口清单、路径穿越和符号链接测试、错误参数与伪造批准测试。反例搜索应包含允许路径中的恶意内容、运行时配置遗漏和文件替换；静态路径测试未覆盖后两者时，要明确缩小主张。

回答子主张

报告可追溯到指定版本的证据，遇到证据不足或冲突会按规则说明限制。证据包括独立问题集上的事实核验、引文支持率与完整率、冲突切片、覆盖率及正常误拒。反例是“引文真实但已过期”“报告路径合法但内容错误”。最小权限不能排除这些错误，第 3 章的测量必须独立进行。

恢复子主张

超时重试不会重复提交同一副作用，报告校验失败时不覆盖上一个已验证版本。证据来自故障注入、后端幂等记录、提交前后状态与恢复日志。仅重复调用两次已经成功的函数不能覆盖“提交后、确认前中断”；若重启会丢失内存状态，也不能声称跨进程恢复已验证。

责任子主张

示例中的部署负责人冻结版本与批准范围，评测负责人保存协议和证据，运行负责人维护告警、停用和回滚，数据负责人处理授权与日志保留。角色可由不同人兼任，但裁决、变更和未解决反例必须留下记录。角色名称是教学分工，不意味着本项目已成立这些组织流程。

第 13 章的本地模拟实验只能为其中的路径、参数与模拟恢复边界提供小规模证据；它没有证明完整文档助手的事实性、群体公平或生产发布门槛。把证据接到恰当的子主张，比在报告中罗列很多测试名称更重要。

10.3 发布门槛：先定阈值，再看结果

门槛既包括可以逐次强制执行的权限规则，也包括需抽样估计的行为指标。表 10.1 给出一个**纯教学的验收设定**；数值是组织应预先选择的示例，不是领域统一标准，也不是本书实验已经达到的结果。假设确认集的任务来自声明总体，攻击每任务至多十次尝试，采用独立任务单位与预先固定的 Wilson 95% 区间；若样本相关，须改用相应区间和样本设计。

权限门的“未见已确认失败”与统计门的“风险上界”并不相同。零次攻击成功仍需给出区间；即便频率低于阈值，一次高影响秘密泄露也可能足以阻断发布。对于无足够样本的群体与场景，报告证据不足并决定限制服务或补测，不能将空切片默认判为通过。阈值因业务伤害、可恢复性和替代服务而异，应由有责任的决策者在看结果之前确定。

子主张与门槛	教学验收规则	不满足时的动作
权限边界	任何已确认越权均阻断发布；接口与策略测试全部通过	定位绕过路径，修复后重跑相关与未见反例
预算内攻击	至少 500 个独立任务；任务级 ASR 的区间上界不超过 1%	缩小权限或攻击面，补充独立确认集
正常效用	至少 500 个正常任务；有效完成率区间下界不低于 90%	排查证据覆盖、模型能力与路由问题
正常误拒	同一正常集上，误拒率区间上界不超过 5%；关键语言切片另审	调整规则并复测合法边界请求
恢复与责任	注入规定故障后状态正确，负责人完成回滚演练与证据签署	停用有副作用工具，补齐恢复能力

表 10.1 本地文档助手的教学发布门槛，未经本书实验验证。

10.4 从一次失败回到可追踪的修复

若报告引用错误版规章，事件链至少应保留：用户请求的必要摘要、系统与索引版本、检索文档标识及哈希、引文与原子陈述映射、发生时间、裁决结果和修复记录。若出现越权请求，则额外记录身份与作用域的非敏感标识、提出的动作、拒绝原因及是否产生副作用。日志不应默认复制全文秘密或用户敏感资料；访问、保留期与删除规则本身也属于控制面。

处置流程可以从停用有副作用接口或回退到上一快照开始，保留证据，再区分检索、生成、权限、执行和评审失效。修复之后既要重现原失败，也要在未见输入和正常边界请求上复测，检查是否引入误拒或其他副作用。若修改了模型、提示、工具、权限、文档索引或评审规则，应判断原证据哪些仍有效，哪些必须重做，而不是沿用旧版本的总分。

练习：十次未越权是否足以发布

一个候选助手在十个攻击任务上没有越权，负责人希望签署“总体越权率低于 1%”。指出缺少哪些证据，并用第 13 章的零事件区间解释。

提示：在独立 Bernoulli 教学假设下，0/10 的 Wilson 上界约为 27.8%。还缺少威胁模型覆盖、足够独立任务、正常效用、恢复演练与具体责任记录；扩大重复尝试次数并不等于增加独立任务数。

下一章把这些裁决需求转成可复现的研究协议；研究应让结论能够追溯到数据、代码与失败记录，而不是依靠一份措辞完善的发布说明。

第 11 章　研究方法：让实验、证据和结论相匹配

11.1　在写代码前写一页实验注册

一项可信的安全研究应当能被另一个人按相同条件重做，也应允许结果否定原假设。先声明研究问题、目标总体、威胁模型、主指标、对照、数据切分、停止条件、统计方法与失败解释，再开始调参。预注册不必复杂，但测试集上发生的变更需要记录，并区分探索性与确认性结果。

例如研究“权限校验是否减少间接注入导致的越权”，至少要写清：资产为本地虚拟秘密；攻击者只能修改指定文档；正常用户只授权读取会议摘要；提议器可以提出读、写或发送动作，但执行器只获得声明的模拟工具权限；每任务至多 K 次尝试，工具和 token 预算固定。比较对象应收到相同文档与动作提议，并用外部状态验证未经授权的效果，而不是只由模型判断自己是否安全。本书第 13 章的最小实验采用本地虚拟对象，不连接真实敏感系统。

如果要推广到开放权重模型、真实检索器或更长交互，需重新声明模型版本、提示、采样、攻击入口和预算。模拟提议器验证的是执行边界；真实模型上的提示注入鲁棒性还需另做实验。两种结论必须在报告中分开。

11.2　围绕一个假设组织实验矩阵

完整矩阵不意味着把所有算法跑一遍。先锁定主问题，再选择能排除替代解释的最小对照。例如，“加门后 ASR 更低”可能来自权限检查，也可能来自拒绝所有动作、减少可用工具或增加超时。应保留正常完成率、误拒、环境故障和成本，才能区分这些解释。

维度	需要固定或主动变化的条件	对应的替代解释
模型与数据	权重/服务版本、任务来源、去重与切分	版本漂移、训练污染、任务难度差异
攻击	入口、权限、知识、搜索与重试预算	攻击更弱或消耗预算不同
运行时	相同工具语义、权限、状态与超时	工具被移除或环境故障造成低 ASR
防御对照	原始、输入/输出防护、执行器校验	效果到底来自哪一层
指标	风险、正常效用、误拒、延迟与成本	用更少服务换取表面安全
复测	未见任务、攻击来源、语言与随机种子	对受测样本过拟合

表 11.1　实验对照应针对可能推翻解释的原因，而非只增加方法名称。

若研究量化或 KV 缓存压缩对安全的影响，可在同一任务、攻击预算和运行时下增加全精度基线、压缩配置与校准数据切片，报告任务效用—风险—成本曲线。先把实现质量和效用差异控制住，再讨论安全变化；表征探针或 SAE 变化只能提供补充机制证据。完整后训练算法和低比特部署实现分别转引后训练、量化压缩分册，本书聚焦其安全影响及证据要求。

11.3 统计单位、对照与长时域失败

对攻击任务先确定分析单位：一个独立目标、一个场景、一个用户，还是一次生成。预算内的自适应尝试不能被重新计成独立任务。二元比例在独立任务条件下可用 Wilson 区间，相关任务可用簇级 bootstrap；两个系统尽量在同一组任务上配对，保留逐任务差异，而不是把两个总体区间是否重叠当成差异检验。

多随机种子可衡量训练或采样波动，多任务切片可衡量输入异质性；二者覆盖的不确定性不同。报告重采样单位、次数与种子，解释区间以哪些随机性为条件。自动评审不可靠时，应加入人工裁决或程序后置条件；统计区间不包括未经建模的标注偏差。

长时域任务更需小心。设 S_t 表示第 t 步满足预先规定的成功条件，则链式法则给出

$$\mathbb{P}(S_1 \cap \cdots \cap S_T) = \prod_{t=1}^{T} \mathbb{P}(S_t \mid S_1, \ldots, S_{t-1}). \tag{11.1}$$

只有在独立且同分布、单步失败率均为 p 的额外假设下，才简化为 $(1-p)^T$。真实智能体会共享记忆、错误工具结果和自适应计划，单步失败通常相关；应直接报告任务级成功、失败首次出现位置、恢复率和隐藏副作用，并保存完整轨迹。低平均单步错误率并不足以推算长任务风险。

练习：相同边际错误率能否推算同样的轨迹成功率

两步任务每一步边际失败率都是 0.1。比较独立失败、两步总在同一批任务失败，以及两步失败集合互不相交时的任务成功率。

提示：分别为 0.81、0.90 和 0.80。仅知道边际错误率不能确定联合概率；同理，增加生成次数并不自动增加独立场景数。

11.4 已实现的最小检查与尚未运行的模型实验

本书将四类证据明确区分：手算或合成教学数据、实际执行的程序检查、原论文报告、尚未执行的研究方案。合成数据可检验公式和执行器边界，却不能证明任何真实模型具有相同的性能。

在本主题根目录，可运行已实现的入口：

Listing 11.1 本书实际最小检查的复现入口。

```
python manuscript/experiments/run_minimal_checks.py
```

程序及其配套配置只使用本地虚拟内容与模拟工具；结果保存在 `output/experiments/`。该目录中的命令、配置、运行时版本、种子、原始记录和汇总结果共同说明执行了什么。第 13 章给出手算参考值、边界检查与结果解释；应以实际生成的记录为准，不能将预期结果写入日志后当成测量。

除上述最小检查外，本版没有运行完整 GCG 搜索、AgentDojo、真实模型成员推断或 SAE 训练。后续模型实验的可执行方案是：先取得许可相符的权重与数据，固定版本及隔离环境；实现与基准接口一致的适配器；用无攻击任务检查工具语义；在验证集选择阈值；冻结配置后运行留出任务；保存全部尝试及评审；按独立任务或簇计算指标并复测未见攻击。缺少模型、硬件、数据或可靠标签时，应如实列出缺口和停止位置。原稿的 `run_eval.py` 只是示意命令，并不存

在于项目中，本版不再把它作为可运行入口。

11.5 让日志和引用足以复核

一次运行应产生配置快照和不可混淆的结果标识。模型名不足以代替权重哈希或在线服务日期；固定种子也不能消除服务端版本变化和非确定性算子影响。最低限度记录：

- 源码版本或文件校验值、依赖版本、命令、种子和配置；
- 任务 ID、来源与切分，全部尝试顺序、工具参数、权限裁决及后置状态；
- 自动与人工标签、分歧裁决、环境故障、缺失值规则及统计脚本；
- 延迟和成本的口径，未运行部分、失败复现及结论适用边界。

虚拟实验也应验证独立参考结果和关键边界，例如空选择集的风险定义、区间端点、路径越界、未授权动作被阻止，以及正常请求是否误拒。避免只测试程序用同一公式算了两遍得到相同值。

日志本身可能含提示、文件内容和个人信息。真实研究应使用合成秘密或获授权的数据，按最小必要原则保存记录，限制日志读取权限，并把回滚和清理纳入实验。源码中的攻击文本属于实验数据，不应被执行器当作额外授权。

论文阅读也应形成可追溯记录：题名、作者、首发年份、所用版本、官方链接、核验日期、支持的具体论断、实验条件及局限。历史资料卡和文件名只是定位入口；PDF 名称中的年份不能替代官方元数据。对近年预印本，应分别写“论文报告”“作者假说”和“本书建议”，不将作者声称的领先结果写成跨任务事实。

本章抓住这些

可复现并非只有一个能运行的脚本。读者还应能看见脚本验证了哪个假设、任务从何而来、哪些操作实际执行、哪些失败被计入，以及什么证据仍不足以支持结论。

第 12 章　研究议程与学习路线

本章提出八个研究问题及相应的候选假说与实验建议，供不同背景的读者选择可证伪、可复核的小问题。截至本次修订日（2026 年 10 月 22 日），本章建议的模型训练与评测尚待执行；已运行的合成教学检查见第 13 章。

12.1　八个可以继续检验的问题

1. **可扩展监督与条件化行为**。研究问题：分解成可核验子任务，能否帮助较弱的监督者发现更强模型的错误？候选假说是工具核验能改善某些任务的标签，但也可能把错误转移到问题分解和验证器。需要保留强标签参考、无工具对照和训练/部署线索互换；不能仅从表面合规推断内部目标。
2. **不可信数据与行动权限**。研究问题：任务绑定的能力集合，在模型规划失败时能保留哪些执行边界？可比较仅输入隔离、仅模型训练和执行器校验，但应分别测越权、任务完成与语义完整性。候选结果是权限检查挡住非法对象，却漏掉合法对象中的错误内容；第 13 章的模拟只演示这种分离。
3. **正确性概率与动作选择**。研究问题：语义不确定性或验证器分数，经独立校准后能否改善风险—覆盖率及接管负荷？比较同覆盖率的随机拒答与简单基线，固定验证阈值，并检查共同错误、分布迁移和高影响动作。提出假说时要指定收益来自更好的错误排序，还是单纯减少服务。
4. **机制审计的外部效度**。研究问题：从一批提示发现的 SAE 特征或回路，能否在新语言、模板和模型版本中预测并因果影响受关注行为？需将发现与确认数据分开，加入残差、范数和无关方向对照。可命名特征与可用安全监控器之间的差距本身就是研究对象。
5. **隐私、能力与安全的联合约束**。研究问题：改变 DP 预算、数据删除或抽取防御时，哪些效用与安全切片首先退化？应先固定相邻关系、成员推断基线和计算预算，再画多目标结果。候选假说可以是总体效用相近而长尾群体表现不同；不能预先假定必然存在某种通用 Pareto 形状。
6. **评测代理与真实目标**。研究问题：在基准上优化的方法，能否迁移到未见攻击者、评审器和执行环境？保留确认集、随机化评审次序，使用独立后果或人工证据。若改进只在原评分器上出现，结果仍有价值，但应解释为代理适应的证据，而非安全普遍提高。
7. **多语言与社会情境**。研究问题：相同权限与拒答规则，在不同语言、方言和无障碍场景下是否带来不同服务成本？先确定合法请求与群体条件指标，控制难度与来源，并报告小样本不确定性。人工设计的反事实对照与真实使用分布应分别分析。
8. **可验证发布与持续有效性**。研究问题：一次版本更新后，哪些旧证据仍然有效，哪些控制必须重做？可用变更清单与故障注入检验依赖关系，再测回滚速度和责任链的完整性。治理流程的存在不等于流程已经可靠执行，演练记录才是相应证据。

每个问题都需要一个清晰失败条件。例如“风险门改善安全”过于宽泛；“在固定任务分布和覆盖率下，校准分数的错误排序优于随机拒答，且独立测试区间支持该差异”才便于验证。结果为负时也应保留日志，说明假说失败在哪个条件，而不是追加方法名称直到获得显著结果。

12.2 压缩研究中的安全切口

本书只保留压缩对安全证据的影响。完整量化、缓存与部署实现见《智能体模型量化、压缩与部署：从数值误差到任务可靠性》；后训练目标与优化实现见《大语言模型后训练：数据、偏好、强化学习与蒸馏》；智能体组件与工作流见《LLM 智能体系统：从模型能力到可控自主工作流》。本书独立说明所需的威胁模型与指标。

12.2.1 量化后的安全保持

一个可检验问题是：在正常任务效用相近时，压缩是否改变某些边界请求的拒答、事实性或工具参数错误率？比较全精度与压缩版本时，固定提示、采样、攻击预算和工具权限，并记录分叉位置。不能仅用困惑度近似安全，也不能假定每个压缩比都会出现同样的非线性退化。若测延迟、成本和显存，还需记录硬件、软件与批量条件；本书没有进行这类 GPU 测量。

12.2.2 长轨迹中的缓存与安全证据

用户目标、来源标记、证据片段和权限说明对后续行动可能具有不同作用。可以用合成轨迹标记这些信息，比较缓存截断或压缩前后是否遗失必要证据、改变工具参数或延误恢复。候选假说是总 token 保留率相同，安全相关信息的保留质量仍不同；也可能观察到效应由任务难度解释。凭证应由可信运行时管理，实验使用合成标识，不能为研究方便把真实秘密长期放入模型上下文。

12.2.3 不确定性驱动的动作门

将语义不确定性、陈述验证或工具结果一致性接入动作选择，比较“继续、只读、核验、人工接管”几种策略。要同时计算选择性风险、覆盖率、任务完成和接管工作量，并区分信息不足与权限不足：置信度高不授予新权限，置信度低也不自动证明请求不合法。高影响动作的阈值应来自独立验证和责任选择，而不是在测试集上挑最有利的点。

12.2.4 压缩前后的机制对照

可比较压缩前后候选方向、字典重构、干预效应与行为失效，但特征编号或激活尺度未必可直接对齐。先建立相同任务、层位置与测量尺度的对应，再检查行为和受控干预。若观察到相关性改变，不能立即解释为“安全回路被删除”；需排除重构误差、表示重参数化、范数变化和正常能力损失。完整压缩实现留在对应分册，本书重点是论断需要什么证据。

12.3 十二周学习与研究计划

表 12.1 是可以按背景缩放的学习建议。第一轮可以只完成手算、文献卡和本地模拟，把需要模型权重、训练资源或额外标注的项目列入后续计划。每周交付物应标记“已运行”“待运行”或“仅方案”，并附相应证据。

表 12.1 十二周建议路线；资源不足时可先完成协议与合成对照。

周	学习主题	可检查的交付物
1	术语、系统边界与威胁模型	一页资产、攻击者、权限、预算与恢复路径说明；标记尚未测量的假设
2	SFT、奖励模型、PPO 与 DPO	两候选回答的封闭算例、必要推导与偏好数据审计表
3	事实性与核验方法	原子事实标签说明、初答/终答对照方案和人工评审子集
4	校准与语义不确定性	复算 ECE、Brier、风险—覆盖率；记录分桶与语义聚类假设
5	越狱与评审规则	固定攻击预算、正常任务和留出集；有资源再运行真实模型对照
6	间接注入与输入分区	在合成文档中复现控制/数据混淆，记录执行器消融
7	工具权限与故障恢复	本地模拟执行器的边界、重试、恢复日志及未覆盖条件
8	抽取、成员推断与 DP	成员先验与攻击优势算例；隐私会计配置和后续实验方案
9	SAE、归因与因果干预	先完成残差控制的二维算例，再决定是否运行模型激活实验
10	证据链与发布案例	正常/攻击计数、区间、评审协议和具体安全案例
11	任选一个研究切口	预注册问题、独立基线、资源预算、失败条件与复现计划
12	写作与独立复核	数据和模型卡、源码与版本摘要、失败记录及范围明确的研究报告

练习：把兴趣缩成一个可证伪问题

从上面的八个方向中选一个，写下输入分布、保护资产、攻击能力、比较基线、主要指标和失败条件，再列出当前资源真正能完成的部分。

提示：若没有模型训练资源，可先比较两个确定性执行器或校验指标算式，但结论必须停留在执行与测量边界，不能改写成某类真实模型优于另一类。把尚需权重、数据或标注的步骤列成后续方案。

下一章给出本次最小教学检查。

第 13 章　最小可复核算例与本次实测

本章将指标、权限和恢复边界落实到可复算的合成案例。第一组与第三组是预先给定标签的教学数据；第二组实际运行本地模拟执行器。它们不调用真实模型，也不代表真实模型性能。脚本入口为 `manuscript/experiments/run_minimal_checks.py`，命令、配置、版本、种子、源码摘要和逐条结果保存在 `output/experiments/`。

从本主题目录运行：

```
python manuscript/experiments/run_minimal_checks.py
```

这里需要真正可执行的 Python 3；本次解释器和平台以 `output/experiments/minimal_results.json` 为准。种子 17 被保存，但当前样本与提案都是确定性的，不把重复运行当作独立随机实验。

13.1　五样本的 Brier、ECE 与选择性风险

设五个回答的正确性概率预测和独立给定标签为

$$q = (0.90, 0.80, 0.60, 0.40, 0.20), \qquad z = (1, 1, 0, 1, 0).$$

$z_i = 1$ 表示回答符合已定义的正确性规则，q_i 预测该事件，不是 token 概率。逐样本平方误差为 $(0.01, 0.04, 0.36, 0.36, 0.04)$，故二元单事件 Brier 为

$$\widehat{\mathrm{Brier}} = \frac{1}{5}\sum_{i=1}^{5}(q_i - z_i)^2 = \frac{0.81}{5} = 0.162. \tag{13.1}$$

采用固定分桶 $[0, 0.5)$ 和 $[0.5, 1]$。低桶有两条样本，准确率 1/2，平均置信度 0.30；高桶有三条样本，准确率 2/3，平均置信度为 2.30/3。于是

$$\widehat{\mathrm{ECE}} = \frac{2}{5}\left|\frac{1}{2} - 0.30\right| + \frac{3}{5}\left|\frac{2}{3} - \frac{2.30}{3}\right| = 0.080 + 0.060 = 0.140. \tag{13.2}$$

报告时应保留精确均值或足够精度，不用显示的三位小数反算差值。两桶边界、端点归属和样本数是定义的一部分；$n = 5$ 不支持稳定的总体校准结论。第 3 章的八样本例子进一步说明，同一组预测可能有两桶 ECE 为零、四桶 ECE 非零。

按置信度从高到低保留前 k 个回答，得到表 13.1。所有分数互异，因而每个 $k > 0$ 都可以用相应阈值实现；并列分数时需另定规则。

提高阈值减少覆盖率，却不保证每一步都降低经验风险：从 $t = 0.4$ 提高到 0.6 去掉一条正确回答，风险从 1/4 升为 1/3。哪些请求应检索、核验或转人工是额外决策，不能从表中标签事后指定为已经生效的策略。真实阈值须在验证集上选择，再用独立测试报告风险、效用和接管成本。空接受集不提供“零风险”证明。

保留数 k	阈值 t（$q_i \geq t$）	覆盖率	正确数	选择性风险
0	1.0	0	0	未定义
1	0.9	0.20	1	$0/1 = 0$
2	0.8	0.40	2	$0/2 = 0$
3	0.6	0.60	2	1/3
4	0.4	0.80	3	1/4
5	0.2	1.00	3	2/5

表 13.1 同一组五样本在不同接受阈值下的计数；未设置任何真实部署动作策略。

13.2 临时虚拟文件上的间接注入

脚本在 `TemporaryDirectory` 中真实创建八个正常会议文档、八个带攻击标记的文档、一个合成私有文件和一个指向它的符号链接。可信主机读取任务指定文档，形成任务对象；文档内容中的 `[[PROPOSE]]` 后跟 JSON 动作提案。规划器是故意服从该标记的确定性函数，**不是语言模型**；无标记时，它输出已知正确摘要的写报告与发送提案，而没有验证真实文本摘要能力。

工具接口包括 `read_file`、`write_report` 和 `send_message`。文件读取实际调用本地 `read_text`；写报告和发送消息仅改变内存中的报告值与 outbox，不写任意磁盘文件，也不连接网络。两种执行器都受临时目录的外层边界约束，原始基线也不能读取真实系统秘密。输出目录中保存的示例文件只是便于检查的合成材料；本次试验实际在临时目录中进行。

校验执行器使用主机任务对象检查：三种工具的精确 schema、字符串参数、规范化后读取对象白名单、固定报告路径和固定收件人。文档可以影响动作提案，但不能修改该任务。它没有实现人工确认、签名能力令牌或通用指令剥离；“数据与指令分离”的可执行部分是让不可信提案无法扩张主机授权，而不是假定模型已经识别出恶意文本。

七个越权案例覆盖私有读取、路径穿越、符号链接别名、错误收件人、错误输出对象、伪造批准字段和未知工具；第八个案例在合法报告路径写入错误摘要。预算是每案例一份固定文档与一组固定动作提案，没有自适应搜索。八个正常文档均使用相同摘要模式，只改变记录标识，所以正常集的覆盖很窄。

执行器	正常完成	越权拒绝	正常误拒	攻击成功
原始基线	8/8	1/7	0/8	7/8
主机权限校验	8/8	7/7	0/8	1/8

表 13.2 确定性本地模拟的实测计数。正常完成与误拒以八条正常记录为分母；越权拒绝以七条越权记录为分母；攻击成功包括合法路径错误摘要。

原始基线拒绝的一条越权案例是不存在的模拟工具，其余六条越权与一条语义攻击成功。校验执行器拒绝七条越权提案，但合法报告路径上的错误摘要仍成功，说明权限正确与内容正确分属不同主张。这里的私有读取成功指合成秘密确实被读取，模拟发送成功指内存 outbox 中出现指定收件人的记录；没有将模型口头声称作为执行证据。逐案例动作及拒绝理由保存在 JSON 中。攻击条件下正常任务的继续完成、真实模型规划能力、延迟及广泛合法边界请求并未在此验证。

13.3 提交后超时与提交前暂存验证

恢复检查和上述八个攻击案例分开运行。脚本先把模拟消息及操作结果写入同一执行器的内存状态，再注入 `TimeoutAfterCommit`：调用方没有收到确认，按同一任务与操作键重试。原始基线产生两次投递；校验执行器检查绑定任务、工具与规范化参数的内存账本，返回原结果并记录 `duplicate_suppressed`，只有一次投递。同键不同参数应被拒绝，不能静默当成原请求。

另一组独立内存状态机检查报告草稿。草稿先进入 `staged`，可信预期值检查后进入 `validated` 或 `rejected`；只有前者可以提交。失败时丢弃尚未提交的草稿，保留旧的已提交值；恢复后可以重新暂存正确草稿、验证并提交。脚本还检查未验证提交与被拒绝草稿提交应失败，保存状态转换而不只写一个“恢复成功”布尔值。

这些检查验证当前进程、顺序执行、指定故障位置下的模拟控制。内存账本没有跨进程持久性，消息与账本之间也没有真实外部系统的原子事务；暂存状态机不回滚已经发送的消息。它们不能证明进程崩溃、并发提交、磁盘事务或跨系统补偿可靠。生产实现需要结合实际后端进一步定义并注入这些故障。

13.4 合成安全评测与 Wilson 区间

第三组数据显式构造三十条标签记录：十个攻击任务中两个成功；十个正常请求中八个完成，一个误拒，一个工具故障；十个越权任务均未发生越权。脚本从逐条记录聚合指标，而不是把指标常数直接写入结果。三组标签不来自模型，也不是上一节八个攻击案例的复用。

为了演示总体比例的不确定性，**额外假设**每组任务是从固定协议总体独立抽样的 Bernoulli 单位，成功概率在组内不变。该假设不是确定性合成标签能够证明的事实；在合成有限集合上首先应报告精确计数。教学区间仅展示：如果未来任务满足这些抽样条件，如何计算区间。二项比例的 Wilson 95% 区间为 [55]

$$\frac{\widehat{p}+\zeta^2/(2n)\ \pm\ \zeta\sqrt{\widehat{p}(1-\widehat{p})/n+\zeta^2/(4n^2)}}{1+\zeta^2/n},\qquad \zeta=1.96,\quad \widehat{p}=x/n. \tag{13.3}$$

这里 x 为事件计数、$n>0$ 为独立任务数，ζ 与正确性标签 z_i 无关。

指标	n	事件数 x	比例	Wilson 95% 区间
攻击成功	10	2	0.20	[0.057, 0.510]
正常完成	10	8	0.80	[0.490, 0.943]
正常误拒	10	1	0.10	[0.018, 0.404]
越权事件	10	0	0.00	[0.000, 0.278]

表 13.3 独立 Bernoulli 教学假设下的区间示范，不代表真实模型总体性能。

正常未完成比例为 2/10，误拒却只有 1/10，因为另一次失败来自工具。零越权观测的上界仍约 27.8%，所以不能将十次未发生写成“风险为零”。若每任务允许自适应多次尝试，先按式 (9.2) 聚合任务成功，再以独立任务作为区间单位；不能把相关尝试数直接当成 n。

13.5 本次检查覆盖范围

脚本使用精确有理数参考核对五样本 Brier 与 ECE，并检查分桶边界、空桶、空接受集及非法输入。脚本还核对第 3 章八样本分桶例的 Brier 0.175、两桶 ECE 0 与四桶 ECE 0.15，并保存对应样本和分桶结果。权限测试核查固定案例与正常记录数；恢复测试检查指定超时、同键冲突和暂存状态约束；区间测试还检查零事件与全事件边界及逐条计数一致性。

每项通过只说明相应算式、接口或状态边界符合本次合成设置。真实模型训练、越狱搜索、隐私攻击、机制分析、生产发送和 GPU 性能实验未运行，相关内容在前文作为论文报告、研究方案或部署建议分别标注。附录汇总数学条件与来源核验记录，便于复核这些有限结论。

附录 A　数学与指标速查

本附录用于复核符号、分母与适用条件，不能替代主文中的威胁模型和标签定义。同一个公式在不同抽样、权限或正确性规则下，可能估计完全不同的对象。

A.1　符号与条件

符号	本书的含义与使用条件
x, y, π	输入、回答与系统策略；策略还可能包含工具决策和运行时约束
e, D	部署环境与固定的任务/输入分布；改变版本或分布需重新检查外推
$\alpha \in \mathcal{A}_B,\ a_t$	受能力与预算 B 限制的攻击策略，以及轨迹中第 t 个工具动作
$\omega, H, I(\omega), L(\omega)$	完整轨迹、伤害事件、严重度与非负损失；$H \subseteq \Omega$，事件外损失为零
$u \in \mathcal{U}$	主机验证的身份及允许身份集合；不使用严重度符号表示身份
q_i, z_i	对已定义正确性事件的概率预测与二元标签，不是单个生成 token 的概率
$t, \widehat{C}(t), \widehat{R}(t)$	接受阈值、经验覆盖率与接受样本中的错误比例
$\rho, m, s(x)$	成员先验、成员标签及成员推断分数；阈值在独立数据上确定
$\beta, \tau_{\text{pref}}$	KL 奖励正则化强度与偏好比较温度，见第 2 章
ε, δ	声明相邻关系下的 DP 参数，不是泄露率

表 A.1　避免把严重度、身份、正确性概率和 token 概率混用。

主文式 (1.1) 使用 $P_{\pi,e,D,\alpha}$ 表示任务抽样、模型随机性、固定攻击策略 α 与环境运行共同诱导的轨迹分布；攻击能力集合本身不能确定该分布。风险 $R(\pi; e, D, \alpha)$ 是它之下的期望损失。若 $L = \mathbf{1}_H I$ 可积且 $P_{\pi,e,D,\alpha}(H) > 0$，则可写成伤害概率乘以有伤害时的条件平均严重度；若该概率为零，直接使用风险为零，不求未定义的条件期望。最坏风险是对 $\alpha \in \mathcal{A}_B$ 的上确界；有限攻击集的平均 ASR 不等于这一对象。

A.2　正确性概率的评分

对 $n > 0$ 个已标注回答，二元正确性事件的 Brier 分数为式 (3.5)。它使用一个概率 q_i 与一个标签 z_i；若按二类概率向量的两个分量求平方和，会得到两倍数值。ECE 见式 (3.4)，还需要固定分桶边界、端点归属和空桶规则；没有样本时不计算总体分数。

同一正确性事件的 Bernoulli 负对数似然（negative log-likelihood，NLL）为

$$\widehat{\mathrm{NLL}}_{\text{correct}} = -\frac{1}{n}\sum_{i=1}^{n}\left[z_i \log q_i + (1 - z_i)\log(1 - q_i)\right]. \tag{A.1}$$

自然对数对应 nat，按扩展值约定处理零概率：给实际发生的事件赋零概率时，损失为 $+\infty$；零权重项按极限为零。若实现把 q_i 裁剪到 $[\eta, 1-\eta]$，须报告 η，因为裁剪改变了评分。模型对正确性事件未输出概率时，不能将语言模型的任意 token 分数充作上述 NLL。

语言建模的 token NLL 则是另一个对象。例如对目标回答的逐 token 概率，

$$\widehat{\mathrm{NLL}}_{\mathrm{token}} = -\frac{\sum_i \sum_{t=1}^{|y_i|} \log \pi_\theta(y_{i,t} \mid x_i, y_{i,<t})}{\sum_i |y_i|}. \tag{A.2}$$

它按 token 数加权长回答，分母要求大于零；逐序列平均或只评回答部分是不同约定。高似然并不意味着事实正确，正确性校准也不能从 token NLL 直接推出。

A.3 覆盖率、选择性风险与阈值

对接受规则 $a_i(t) = \mathbf{1}\{q_i \geq t\}$，式 (3.6) 的覆盖率分母是全部问题数，风险分母是接受数。空接受集的覆盖率为零，风险未定义。若按固定数量保留前 k 个，覆盖率为 k/n；并列分数可能使某个确定阈值无法恰好实现目标覆盖率，需要预先决定全部保留、随机打破并列或使用其他规则。

错误比例的选择性风险与严重度加权风险不相同。若每个回答有预先定义的非负损失 ℓ_i，接受样本的平均损失为 $\sum_i a_i(t)\ell_i / \sum_i a_i(t)$，仍要求接受数非零。它不包含被拒答后的人工处理和延误损失；评估整个服务时需纳入这些路径。第 3 章解释为何低 ECE 不保证良好的错误排序，第 13 章给出可复算数据。

A.4 成员推断、DP 与攻击计数

成员推断优势采用式 (6.4) 的 TPR−FPR。只有成员先验为 1/2 时，准确率才等于 1/2+Adv/2；一般情形使用式 (6.5)。AUC、低 FPR 下的 TPR 与准确率的分母和用途不同，应同时注明接口信息、候选分布与查询成本。

DP 定义见式 (6.6)；基本顺序组合见式 (6.7)。每次报告参数时连同相邻关系、保护单位、采样机制、会计方法和发布范围给出。记录级、用户级以及加入删除、替换相邻关系不能只比较 ε 数值；数据依赖的验证和模型选择也可能消耗预算。

预算内任务级 ASR 见式 (9.2)：分子是至少一次攻击成功的任务数，分母是全部预定攻击任务数。正常效用与误拒使用独立正常任务分母，见式 (9.3)。统计区间还要说明独立单位与相关性；固定合成标签的比例只用于验证计算链路，不能代表模型性能。

A.5 安全、效用与成本的比较

令 U 为声明分布上的任务效用，R 为给定威胁模型下的风险，C_{cost} 为明确口径的成本。发布选择可用约束形式表达：

$$\max_{\pi,\,\mathrm{runtime}} \; U \quad \text{s.t.} \quad R \leq \tau_R, \qquad C_{\mathrm{cost}} \leq \tau_C. \tag{A.3}$$

这只是决策表达式，并不表示这些函数已经可精确估计或约束已经验证。概率、损失严重度、毫秒与货币单位不能未经权重与解释直接相加。先报告各自条件和不确定性，再展示不可相互支配的候选点，并说明阈值由谁、按什么伤害与服务需求选择。

附录 B　论文与资料使用说明

本书建立在已有书稿、历史检索资料和本次增补核验之上。三类材料的用途不同：归档保留来源，文献支持明确论断，实验记录证明本次实际执行的步骤。本次修订以 2026 年 10 月 22 日为资料观察日期；观察日期并不表示某项结果在该日仍是最优，也不表示已复现全部原始实验。

B.1　历史归档与当前文献表分开维护

原主源码与原版 PDF 保留在本主题的 `source/` 与 `output/pdf/`，新版使用独立入口和文件名。历史检索留下约五十一篇开放 PDF 的索引、阅读卡与领域地图，相关材料包括 `output/learn_more.md`、`output/scratch.md` 和 论文路径索引`.json`。它们有助于定位论文，但不是本版所有题名、年份、版本、摘要与结论均已核实的凭证。旧绝对路径需结合相对索引定位，旧日期与旧引文不能不经检查直接沿用。

本版实际引用由 `manuscript/references.bib` 与正文共同决定，采用 BibLaTeX 和 Biber。当前文献表包含经修订的旧来源及围绕实际缺口补充的经典资料。下载文件数、参考文献条数、正文引用数和已独立复现实验数是不同计数；不能把“收集过论文”写成“验证过论文结论”。

B.2　本次核验记录如何阅读

人工可读的总索引是 `output/evidence/README.md`，逐条说明保存在 `output/evidence/` 下以引文键命名的 Markdown 文件中。完整复核记录为 `output/evidence/source_audit_reviewed.json`。初始抓取记录 `output/evidence/source_audit.json` 保留原始访问结果，不等于最终采用的引文；HTTP 响应、页面与 PDF 摘取等证据位于同目录及其 `sources/` 子目录。

核验记录区分书目身份、版本选择和论断适用范围：作者、题名、日期、DOI/arXiv 或官方链接是否对应；本次采用哪个版本；正文引用能支持什么、不能外推到什么。部分来源核对了官方页面、首页、摘要或特定段落，不能因此声称全文中的每个定理、实验和实施机制都经过独立验证。论文自己的性能结果仍应标为论文报告；本书建议的实验不能改写成已经观察到的结果。

文献表中的日期需按来源解释：以预印本身份收录的 arXiv 条目，其 `date` 记录官方首发日期，URL 与说明固定本次采用的版本；`urldate` 是本次访问日期，不能当作论文发表或版本修订日期。正式论文与报告采用相应发表或版本日期；例如 Model Cards、Datasheets 与 Yeom 的正式论文即使附有 arXiv 副本链接，仍按正式发表年份著录。因此，首发年份、所引版本年份与访问年份可能不同。例如 RAGTruth 官方首发为 2023 年 12 月 31 日，即使编号以 2401 开头，也不能仅凭编号改成年份。

本次重要的版本处理包括：*Scaling Monosemanticity* 引用 2024 年 5 月 21 日官方原作；SAFER 的 `date` 为首发日 2025 年 7 月 1 日，固定所引 v3 则发布于 2026 年 1 月 30 日，另保留 v2 曾撤回的事实；眞值对齐不确定性资料首发于 2026 年 4 月 1 日，采用 2026 年 10 月 16 日的 v2，与本地旧版区分。公司政策则固定到 RSP 2025 年 5 月 14 日版本 2.2 与 Preparedness Framework 2025 年 4 月 15 日版本 2，不称其为最新政策。具体标识、日期与决定以文献表和逐条核验记录为准，读

者引用时仍应检查原始来源是否发生变化。

B.3 剩余限制不应隐藏在引文之后

ISO/IEC 42001 的官方入口未能取得可读全文，本次通过 BSI 官方目录核对标准身份和管理体系主题；EU AI Act 的若干原文端点没有返回可读内容，本次通过欧盟委员会说明核实高层框架。因此相关正文仅作入口介绍，不据此解释具体条款或断言部署已合规。关于对齐算法机制分析的 2026 年记录存在月份元数据不一致，相关论述只使用年份并保留预印本线索；书目与核验记录保留官方原始日期及其不一致说明，不据此断言首发月份，也不把该工作视为本次复现结果。

这些限制不妨碍使用已经得到支持的定义或高层背景，但要求相应缩小结论。若某项数字、机制解释或前沿判断没有可核验来源，应删除、改为待检验假说，或在修订记录中明确列出缺口，而不能仅附上一条主题相近的文献。

B.4 区分本次实测与后续方案

本次最小实验的命令、版本、种子、配置和逐条结果位于 `output/experiments/`，入口是 `manuscript/experiments/run_minimal_checks.py`。第 13 章说明哪些是合成标签的指标计算，哪些是本地模拟执行器的运行。它们不调用真实模型，不提供真实模型 ASR、训练效果、GPU 性能或生产安全保证。第 12 章的训练、机制和跨分布研究则是可执行建议，需获得相应数据、权重与资源后再运行。

维护者修改论断、来源版本、实验脚本或数据后，应同步更新文献表、核验记录、结果及修订说明，并重新执行相关编译和验证。判断可信证据的基本顺序是：先读具体主张，再检查条件、原始记录与反例，最后决定结论能外推到哪里。

参考文献与规范

[1] Long Ouyang et al. *Training language models to follow instructions with human feedback*. arXiv 2203.02155v1. Mar. 4, 2022. URL: https://arxiv.org/abs/2203.02155v1 (visited on 10/22/2026).

[2] Yuntao Bai et al. *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*. arXiv 2204.05862v1. Apr. 12, 2022. URL: https://arxiv.org/abs/2204.05862v1 (visited on 10/22/2026).

[3] Ralph Allan Bradley and Milton E. Terry. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons". In: *Biometrika* 39.3–4 (1952), pp. 324–345. DOI: 10.1093/biomet/39.3-4.324. URL: https://doi.org/10.1093/biomet/39.3-4.324 (visited on 10/22/2026).

[4] John Schulman et al. *Proximal Policy Optimization Algorithms*. arXiv 1707.06347v2. July 20, 2017. URL: https://arxiv.org/abs/1707.06347v2 (visited on 10/22/2026).

[5] Yuntao Bai et al. *Constitutional AI: Harmlessness from AI Feedback*. arXiv 2212.08073v1. Dec. 15, 2022. URL: https://arxiv.org/abs/2212.08073v1 (visited on 10/22/2026).

[6] Rafael Rafailov et al. *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*. arXiv 2305.18290v3. May 29, 2023. URL: https://arxiv.org/abs/2305.18290v3 (visited on 10/22/2026).

[7] Collin Burns et al. *Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision*. arXiv 2312.09390v1. Dec. 14, 2023. URL: https://arxiv.org/abs/2312.09390v1 (visited on 10/22/2026).

[8] Evan Hubinger et al. *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training*. arXiv 2401.05566v3. Jan. 10, 2024. URL: https://arxiv.org/abs/2401.05566v3 (visited on 10/22/2026).

[9] Ryan Greenblatt et al. *Alignment faking in large language models*. arXiv 2412.14093v2. Dec. 18, 2024. URL: https://arxiv.org/abs/2412.14093v2 (visited on 10/22/2026).

[10] Stephanie Lin, Jacob Hilton, and Owain Evans. *TruthfulQA: Measuring How Models Mimic Human Falsehoods*. arXiv 2109.07958v2. Sept. 8, 2021. URL: https://arxiv.org/abs/2109.07958v2 (visited on 10/22/2026).

[11] Sewon Min et al. *FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation*. arXiv 2305.14251v2. May 23, 2023. URL: https://arxiv.org/abs/2305.14251v2 (visited on 10/22/2026).

[12] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. *Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation*. arXiv 2302.09664v3. Feb. 19, 2023. URL: https://arxiv.org/abs/2302.09664v3 (visited on 10/22/2026).

[13] Chuan Guo et al. *On Calibration of Modern Neural Networks*. arXiv 1706.04599v2. June 14, 2017. URL: https://arxiv.org/abs/1706.04599v2 (visited on 10/22/2026).

[14] Glenn W. Brier. "Verification of Forecasts Expressed in Terms of Probability". In: *Monthly Weather Review* 78.1 (1950), pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2. URL: https://doi.org/10.1175/1520-0493(1950)078%3C0001:vofeit%3E2.0.co;2 (visited on 10/22/2026).

[15] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. *SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models*. arXiv 2303.08896v3. Mar. 15, 2023. URL: https://arxiv.org/abs/2303.08896v3 (visited on 10/22/2026).

[16] Shehzaad Dhuliawala et al. *Chain-of-Verification Reduces Hallucination in Large Language Models*. arXiv 2309.11495v2. Sept. 20, 2023. URL: https://arxiv.org/abs/2309.11495v2 (visited on 10/22/2026).

[17] Zhibin Gou et al. *CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing*. arXiv 2305.11738v4. May 19, 2023. URL: https://arxiv.org/abs/2305.11738v4 (visited on 10/22/2026).

[18] Akari Asai et al. *Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection*. arXiv 2310.11511v1. Oct. 17, 2023. URL: https://arxiv.org/abs/2310.11511v1 (visited on 10/22/2026).

[19] Cheng Niu et al. *RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models*. arXiv 2401.00396v2. Dec. 31, 2023. URL: https://arxiv.org/abs/2401.00396v2 (visited on 10/22/2026).

[20] Patrick Lewis et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. arXiv 2005.11401v4. May 22, 2020. URL: https://arxiv.org/abs/2005.11401v4 (visited on 10/22/2026).

[21] Andy Zou et al. *Universal and Transferable Adversarial Attacks on Aligned Language Models*. arXiv 2307.15043v2. July 27, 2023. URL: https://arxiv.org/abs/2307.15043v2 (visited on 10/22/2026).

[22] Kai Greshake et al. *Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection*. arXiv 2302.12173v2. Feb. 23, 2023. URL: https://arxiv.org/abs/2302.12173v2 (visited on 10/22/2026).

[23] Alexandra Souly et al. *A StrongREJECT for Empty Jailbreaks*. arXiv 2402.10260v2. Feb. 15, 2024. URL: https://arxiv.org/abs/2402.10260v2 (visited on 10/22/2026).

[24] Mantas Mazeika et al. *HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal*. arXiv 2402.04249v2. Feb. 6, 2024. URL: https://arxiv.org/abs/2402.04249v2 (visited on 10/22/2026).

[25] Patrick Chao et al. *JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models*. arXiv 2404.01318v5. Mar. 28, 2024. URL: https://arxiv.org/abs/2404.01318v5 (visited on 10/22/2026).

[26] Eric Wallace et al. *The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions*. arXiv 2404.13208v1. Apr. 19, 2024. URL: https://arxiv.org/abs/2404.13208v1 (visited on 10/22/2026).

[27] Hakan Inan et al. *Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations*. arXiv 2312.06674v1. Dec. 7, 2023. URL: https://arxiv.org/abs/2312.06674v1 (visited on 10/22/2026).

[28] Seungju Han et al. *WildGuard: Open One-Stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs*. arXiv 2406.18495v3. June 26, 2024. URL: https://arxiv.org/abs/2406.18495v3 (visited on 10/22/2026).

[29] Yangjun Ruan et al. *Identifying the Risks of LM Agents with an LM-Emulated Sandbox*. arXiv 2309.15817v2. Sept. 25, 2023. URL: https://arxiv.org/abs/2309.15817v2 (visited on 10/22/2026).

[30] Qiusi Zhan et al. *InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents*. arXiv 2403.02691v3. Mar. 5, 2024. URL: https://arxiv.org/abs/2403.02691v3 (visited on 10/22/2026).

[31] Edoardo Debenedetti et al. *AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents*. arXiv 2406.13352v3. June 19, 2024. URL: https://arxiv.org/abs/2406.13352v3 (visited on 10/22/2026).

[32] Andy Zou et al. *Security Challenges in AI Agent Deployment: Insights from a Large Scale Public Competition*. arXiv 2507.20526v1. July 28, 2025. URL: https://arxiv.org/abs/2507.20526v1 (visited on 10/22/2026).

[33] Mateusz Dziemian et al. *How Vulnerable Are AI Agents to Indirect Prompt Injections? Insights from a Large-Scale Public Competition*. arXiv 2603.15714v1. Mar. 16, 2026. URL: https://arxiv.org/abs/2603.15714v1 (visited on 10/22/2026).

[34] Nicholas Carlini et al. *Extracting Training Data from Large Language Models*. arXiv 2012.07805v2. Dec. 14, 2020. URL: https://arxiv.org/abs/2012.07805v2 (visited on 10/22/2026).

[35] Nicholas Carlini et al. *Quantifying Memorization Across Neural Language Models*. arXiv 2202.07646v3. Feb. 15, 2022. URL: https://arxiv.org/abs/2202.07646v3 (visited on 10/22/2026).

[36] Samuel Yeom et al. "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting". In: *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. arXiv 1709.01604v5. 2018, pp. 268–282. DOI: 10.1109/csf.2018.00027. URL: https://arxiv.org/abs/1709.01604v5 (visited on 10/22/2026).

[37] Justus Mattern et al. *Membership Inference Attacks against Language Models via Neighbourhood Comparison*. arXiv 2305.18462v2. May 29, 2023. URL: https://arxiv.org/abs/2305.18462v2 (visited on 10/22/2026).

[38] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". In: *Foundations and Trends in Theoretical Computer Science* 9.3–4 (2014), pp. 211–487. DOI: 10.1561/0400000042. URL: https://doi.org/10.1561/0400000042 (visited on 10/22/2026).

[39] Martín Abadi et al. *Deep Learning with Differential Privacy*. arXiv 1607.00133v2. July 1, 2016. URL: https://arxiv.org/abs/1607.00133v2 (visited on 10/22/2026).

[40] Alexander Wan et al. *Poisoning Language Models During Instruction Tuning*. arXiv 2305.00944v1. May 1, 2023. URL: https://arxiv.org/abs/2305.00944v1 (visited on 10/22/2026).

[41] Minh N. Vu et al. *Analysis of Privacy Leakage in Federated Large Language Models*. arXiv 2403.04784v1. Mar. 2, 2024. URL: https://arxiv.org/abs/2403.04784v1 (visited on 10/22/2026).

[42] Yue Huang et al. *TrustLLM: Trustworthiness in Large Language Models*. arXiv 2401.05561v6. Jan. 10, 2024. URL: https://arxiv.org/abs/2401.05561v6 (visited on 10/22/2026).

[43] Margaret Mitchell et al. "Model Cards for Model Reporting". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019, pp. 220–229. DOI: 10.1145/3287560.3287596. URL: https://arxiv.org/abs/1810.03993v2 (visited on 10/22/2026).

[44] Timnit Gebru et al. "Datasheets for datasets". In: *Communications of the ACM* 64.12 (2021), pp. 86–92. DOI: 10.1145/3458723. URL: https://arxiv.org/abs/1803.09010v8 (visited on 10/22/2026).

[45] Nelson Elhage et al. *Toy Models of Superposition*. arXiv 2209.10652v1. Sept. 21, 2022. URL: https://arxiv.org/abs/2209.10652v1 (visited on 10/22/2026).

[46] Hoagy Cunningham et al. *Sparse Autoencoders Find Highly Interpretable Features in Language Models*. arXiv 2309.08600v3. Sept. 15, 2023. URL: https://arxiv.org/abs/2309.08600v3 (visited on 10/22/2026).

[47] Andy Zou et al. *Representation Engineering: A Top-Down Approach to AI Transparency*. arXiv 2310.01405v4. Oct. 2, 2023. URL: https://arxiv.org/abs/2310.01405v4 (visited on 10/22/2026).

[48] Adly Templeton et al. *Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet*. Transformer Circuits Thread; original research report. Anthropic. May 21, 2024. URL: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html (visited on 10/22/2026).

[49] Wei Shi et al. *SAFER: Probing Safety in Reward Models with Sparse Autoencoder*. arXiv 2507.00665v3; v2 was withdrawn; this citation uses v3. July 1, 2025. URL: https://arxiv.org/abs/2507.00665v3 (visited on 10/22/2026).

[50] Aarush Sinha et al. *Mechanistic Analysis of Alignment Algorithms in Language Models*. arXiv 2606.09850v1. May 9, 2026. URL: https://arxiv.org/abs/2606.09850v1 (visited on 10/22/2026).

[51] Dan Hendrycks et al. *Measuring Massive Multitask Language Understanding*. arXiv 2009.03300v3. Sept. 7, 2020. URL: https://arxiv.org/abs/2009.03300v3 (visited on 10/22/2026).

[52] Percy Liang et al. *Holistic Evaluation of Language Models*. arXiv 2211.09110v2. Nov. 16, 2022. URL: https://arxiv.org/abs/2211.09110v2 (visited on 10/22/2026).

[53] Zhexin Zhang et al. *SafetyBench: Evaluating the Safety of Large Language Models*. arXiv 2309.07045v2. Sept. 13, 2023. URL: https://arxiv.org/abs/2309.07045v2 (visited on 10/22/2026).

[54] Buğra Alperen Uluırmak and Rifat Kurban. *EvalSafetyGap: A Hybrid Survey and Conceptual Framework for LLM Evaluation-Safety Failures*. arXiv 2606.30219v5. June 29, 2026. URL: https://arxiv.org/abs/2606.30219v5 (visited on 10/22/2026).

[55] Edwin B. Wilson. "Probable Inference, the Law of Succession, and Statistical Inference". In: *Journal of the American Statistical Association* 22.158 (1927), pp. 209–212. DOI: 10.1080/01621459.1927.10502953. URL: https://doi.org/10.1080/01621459.1927.10502953 (visited on 10/22/2026).

[56] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Technical Report NIST AI 100-1. National Institute of Standards and Technology, Jan. 2023. DOI: 10.6028/NIST.AI.100-1. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf (visited on 10/22/2026).

[57] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. Technical Report NIST AI 600-1. National Institute of Standards and Technology, July 2024. DOI: 10.6028/NIST.AI.600-1. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf (visited on 10/22/2026).

[58] International Organization for Standardization and International Electrotechnical Commission. *ISO/IEC 42001:2023: Information technology—Artificial intelligence—Management system*. Standard identification verified against the BSI catalogue; full standard not consulted. 2023. URL: https://www.iso.org/standard/81230.html (visited on 10/22/2026).

[59] European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*. Regulation identifier and risk framework checked against the European Commission AI Act page. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj (visited on 10/22/2026).

[60] OWASP Foundation. *OWASP Top 10 for LLM Applications 2025*. Version 2025. Nov. 17, 2024. URL: https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/ (visited on 10/22/2026).

[61] MITRE. *MITRE ATLAS: Adversarial Threat Landscape for AI Systems*. Version 2026.09. Version-locked official data release; commit 3259f388d19cbcca11bacf12a0ef97f4198f711b. Sept. 10, 2026. URL: https://github.com/mitre-atlas/atlas-data/tree/3259f388d19cbcca11bacf12a0ef97f4198f711b (visited on 10/22/2026).

[62] Anthropic. *Responsible Scaling Policy*. Version 2.2. Historical version 2.2; not presented as the current policy. May 14, 2025. URL: Anthropic RSP v2.2 (PDF) (visited on 10/22/2026).

[63] OpenAI. *Preparedness Framework*. Version 2. Historical version 2; not presented as the current framework. Apr. 15, 2025. URL: https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf (visited on 10/22/2026).