如何防御模型蒸馏攻击

——以 DeepSeek 模型蒸馏事件为例的深度技术分析

作者：winzheng Research Lab 发布日期：2026年2月13日 密级：公开发布 | Classification: Public

---

目录

一、执行摘要 (Executive Summary)

模型蒸馏（Model Distillation）是一种将大型教师模型（Teacher Model）的知识压缩传输至较小学生模型（Student Model）的技术。这项技术最初由 Geoffrey Hinton 等人在 2015 年提出，本意是优化部署效率。然而，近年来，模型蒸馏被恶意利用为一种攻击手段——通过系统性地查询商业 API 接口来窃取模型能力，这已成为全球 AI 行业的重大安全威胁。

2025年初，DeepSeek 的模型蒸馏事件引发了全球关注。有多方证据表明，DeepSeek 在训练其模型时大量使用了来自 OpenAI 等商业模型的蒸馏数据。这一事件不仅引发了对 AI 知识产权的广泛讨论，也迫使各大 AI 厂商重新审视其防御体系。

本报告将深入分析模型蒸馏攻击的技术原理、DeepSeek 事件的技术细节、以及企业级别的综合防御策略，为 AI 基础设施公司和模型服务提供商提供可操作的技术指南。

二、模型蒸馏技术原理

2.1 知识蒸馏的核心机制

知识蒸馏的核心思想是利用教师模型的"软标签"（Soft Labels）而非仅仅依赖原始数据的"硬标签"（Hard Labels）来训练学生模型。软标签包含了教师模型对各类别的概率分布信息，这种"暗知识"（Dark Knowledge）揭示了类别间的相似性关系。

在大型语言模型（LLM）场景下，蒸馏变得更加复杂且危险。攻击者可以通过 API 获取教师模型对大量 prompt 的完整响应，这些响应本身就携带了丰富的知识、推理模式和语言风格信息。

2.2 LLM 蒸馏攻击的技术流程

以下是典型的 LLM 蒸馏攻击工作流：

阶段	操作	技术细节
① 数据收集	大规模查询目标 API	使用多样化 prompt 库，覆盖代码、数学、推理、创意写作等全领域
② 数据清洗	过滤与去重	移除低质量响应、去重、质量评分排序
③ 模型训练	SFT 微调学生模型	使用收集的 Q&A 对进行有监督微调（Supervised Fine-Tuning）
④ 对齐优化	RLHF/DPO 对齐	使用教师模型生成的偏好数据进行 RLHF 或 DPO 对齐
⑤ 评估验证	能力对标测试	在标准基准上与教师模型对比表现

2.3 蒸馏损失函数的数学表达

经典的知识蒸馏损失函数结合了软目标损失和硬目标损失：

$$L = \alpha \times T^2 \times KL(p_{teacher}(T) \,||\, p_{student}(T)) + (1-\alpha) \times CE(y, p_{student})$$

其中 **T** 为温度参数（Temperature），控制输出概率分布的"软化"程度。较高的温度会使概率分布更平滑，从而暴露更多的"暗知识"。在 LLM 蒸馏攻击中，攻击者并不需要访问 logits，仅凭生成的文本就可以完成有效蒸馏。

三、DeepSeek 蒸馏事件深度分析

3.1 事件背景

DeepSeek 作为中国领先的 AI 创业公司，在 2024 年底至 2025 年初发布了一系列在多项基准测试中表现优异的模型，包括 DeepSeek-V3 和 DeepSeek-R1。然而，多项技术分析显示，这些模型的训练过程中大量使用了从 OpenAI 模型蒸馏的数据。

3.2 技术证据分析

以下是多个独立来源的技术证据：

证据类型	具体发现	技术含义
输出特征匹配	DeepSeek 模型在特定场景下出现与 OpenAI 模型相似的固定表达模式	表明训练数据中包含了目标模型的输出特征
拒绝模式复制	某些拒绝响应的语言风格与 OpenAI 模型高度一致	Safety alignment 的行为模式被"继承"到学生模型
能力分布异常	小模型在特定任务上超越同规模模型的期望表现	蒸馏可以在特定任务上达到超线性的能力转移
API 使用模式	异常的大规模 API 调用模式被检测	符合系统性蒸馏攻击的数据收集特征

3.3 DeepSeek-R1 的蒸馏策略特点

DeepSeek-R1 在蒸馏策略上展现了特别的技术特点。该模型采用了混合训练路径：先使用大规模蒸馏数据进行基础能力建设，再通过强化学习（RL）进行推理能力增强。其 Chain-of-Thought 推理路径的生成模式与 OpenAI o1 的输出风格存在显著相似性，这被认为是蒸馏的直接证据之一。

值得注意的是，DeepSeek 的技术报告中也提到了使用"合成数据"进行训练，而这类"合成数据"的来源和生成方式正是争议的核心。在 AI 行业中，使用另一个 AI 模型生成的数据来训练自己的模型是否构成知识产权侵权，目前仍是一个法律灰色地带。

四、综合防御体系架构

基于对模型蒸馏攻击的深入研究，我们提出一套多层次、系统化的防御架构，覆盖从 API 层到模型层的全方位防护。

4.1 API 层防御：智能速率限制与异常检测

4.1.1 自适应速率限制

传统的固定速率限制已不足以应对复杂的蒸馏攻击。我们建议采用基于行为分析的自适应速率限制系统，实时评估每个 API 用户的调用模式，并动态调整配额。

1. **多维度行为特征提取**：包括查询频率、prompt 多样性、主题覆盖广度、请求时间分布等
2. **实时风险评分**：通过机器学习模型对每个请求序列进行蒸馏风险评分
3. **动态配额调整**：高风险用户自动降低速率限制，低风险用户保持正常服务

4.1.2 查询模式异常检测

蒸馏攻击通常具有可识别的模式特征，与正常用户行为存在显著差异。我们建议部署多层异常检测系统，监控以下关键指标：查询多样性异常高（正常用户往往集中在特定领域）、系统性的能力探测行为、以及大量请求 token 逻辑概率输出的请求。

4.2 输出层防御：智能水印与信息控制

4.2.1 隐式模型水印 (Watermarking)

模型水印是一种在模型输出中嵌入不可见标识的技术，用于追踪和证明模型输出的来源。关键技术包括：

1. **Token 级别水印**：在采样过程中微调 token 选择概率，嵌入统计可检测的模式
2. **语义级别水印**：在保持语义不变的前提下，在表达方式上嵌入可述性水印

3. **结构水印**：在输出的逻辑结构和推理路径中嵌入特征标识

4.2.2 输出信息控制

限制模型输出中的信息量是防御蒸馏的另一重要策略。具体措施包括：不提供完整的 logits/logprobs 输出，仅提供 top-k token 概率；对输出添加可控的噪声，降低蒸馏数据的信噪比；以及对输出进行随机化后处理，破坏蒸馏数据的一致性。

4.3 模型层防御：架构级保护

4.3.1 可学习性降低技术

这是一种前沿的防御技术，旨在使模型输出难以被学生模型有效学习，同时不影响正常用户的使用体验。核心思路是在保持单次响应质量的前提下，在多次响应之间引入受控的不一致性，使得收集到的蒸馏数据难以用于有效训练。

4.3.2 对抗性训练防御

通过在模型训练阶段引入对抗性目标，使模型在保持正常能力的同时，其输出对蒸馏攻击具有天然的抗性。这可以通过以下方式实现：

1. **多目标训练**：同时优化任务性能和反蒸馏目标
2. **输出多样性增强**：提高相同输入下输出的随机性
3. **特征混淆机制**：在输出空间中嵌入干扰性特征

4.4 法律与合规层防御

技术手段之外，法律和合规体系也是防御体系的重要组成部分：

1. **服务条款强化**：明确禁止将 API 输出用于模型训练，并建立技术审计机制
2. **知识产权保护**：探索将 AI 模型输出纳入知识产权保护框架的可能性
3. **国际合作机制**：推动建立跨国 AI 知识产权保护框架和共同标准

五、防御效果评估与对比

不同防御策略的有效性存在显著差异。以下是对各种防御方案的综合评估：

防御策略	防御效果	对用户体验影响	实施复杂度	综合评分
自适应速率限制	★★★☆☆	★★★★☆	★★★☆☆	B+
查询异常检测	★★★★☆	★★★★★	★★★★☆	A–
模型水印	★★★★☆	★★★☆☆	★★★★★	B+
输出信息控制	★★★★★	★★☆☆☆	★★★☆☆	B
可学习性降低	★★★★★	★★★☆☆	★★★★★	A
综合防御体系	★★★★★	★★★★☆	★★★★★	A+

从上表可以看出，单一防御策略均存在局限性。最优解决方案是将多种防御策略组合部署，形成多层次的纵深防御体系。其中，API 层防御提供第一道防线，输出层防御提供追踪和取证能力，模型层防御则从根本上降低蒸馏效果。

六、企业级实施指南

6.1 分阶段部署路线图

我们建议企业按照以下三个阶段逐步部署防御体系：

阶段	关键任务	预期效果
第一阶段（1–3个月）	部署自适应速率限制、建立基础监控体系、更新服务条款	拦截 60% 以上的低级蒸馏攻击
第二阶段（3–6个月）	实施模型水印、部署异常检测 ML 模型、开发取证工具链	拦截 85% 以上的蒸馏攻击，具备取证追踪能力
第三阶段（6–12个月）	研发可学习性降低技术、对抗性训练集成、建立行业联盟	全方位防御能力，显著提升蒸馏攻击成本

6.2 关键性能指标 (KPI)

企业应建立以下核心 KPI 来衡量防御体系的有效性：

- **蒸馏攻击检测率**：目标 > 95%
- **误报率**：目标 < 0.1%
- **正常用户响应延迟增加**：目标 < 50ms
- **水印存活率**（经过微调后仍可检测）：目标 > 80%

七、未来展望与行业建议

7.1 技术发展趋势

模型蒸馏攻击与防御将持续演化。以下是我们对未来发展方向的判断：

攻击端趋势：分布式蒸馏攻击将变得更加隐蔽，通过分散查询来规避检测；跨模型蒸馏将成为新趋势，从多个教师模型融合知识；自动化蒸馏工具链会降低攻击门槛。

防御端趋势：基于大模型的智能检测系统将大幅提升检测精度；可验证计算和零知识证明可能为模型授权提供新范式；联邦学习技术可能提供新的合作训练模式，减少蒸馏动机。

7.2 对行业的建议

基于我们的分析，向 AI 行业提出以下建议：

- **模型提供商**：将反蒸馏防御纳入核心产品路线图，将其视为与模型安全同等重要的基础设施
- **企业用户**：在选择 AI 服务时将反蒸馏能力纳入评估标准，确保自己的数据和使用模式不会被用于未授权的蒸馏
- **研究机构**：加大对可学习性降低、模型水印、异常检测等关键领域的研究投入
- **监管机构**：加快 AI 知识产权保护立法，建立明确的模型蒸馏合规框架

八、结论

模型蒸馏攻击已成为 AI 行业面临的最严峻的安全挑战之一。DeepSeek 事件深刻揭示了当前防御体系的不足，也为行业敲响了警钟。然而，挑战也意味着机遇——投资于先进的防御技术不仅能保护模型资产，还能建立竞争壁垒。

我们强烈建议 AI 基础设施公司将反蒸馏防御作为与模型安全对齐同等重要的战略优先级，并在技术、法律、行业合作三个层面同时发力，构建全方位的防护体系。
