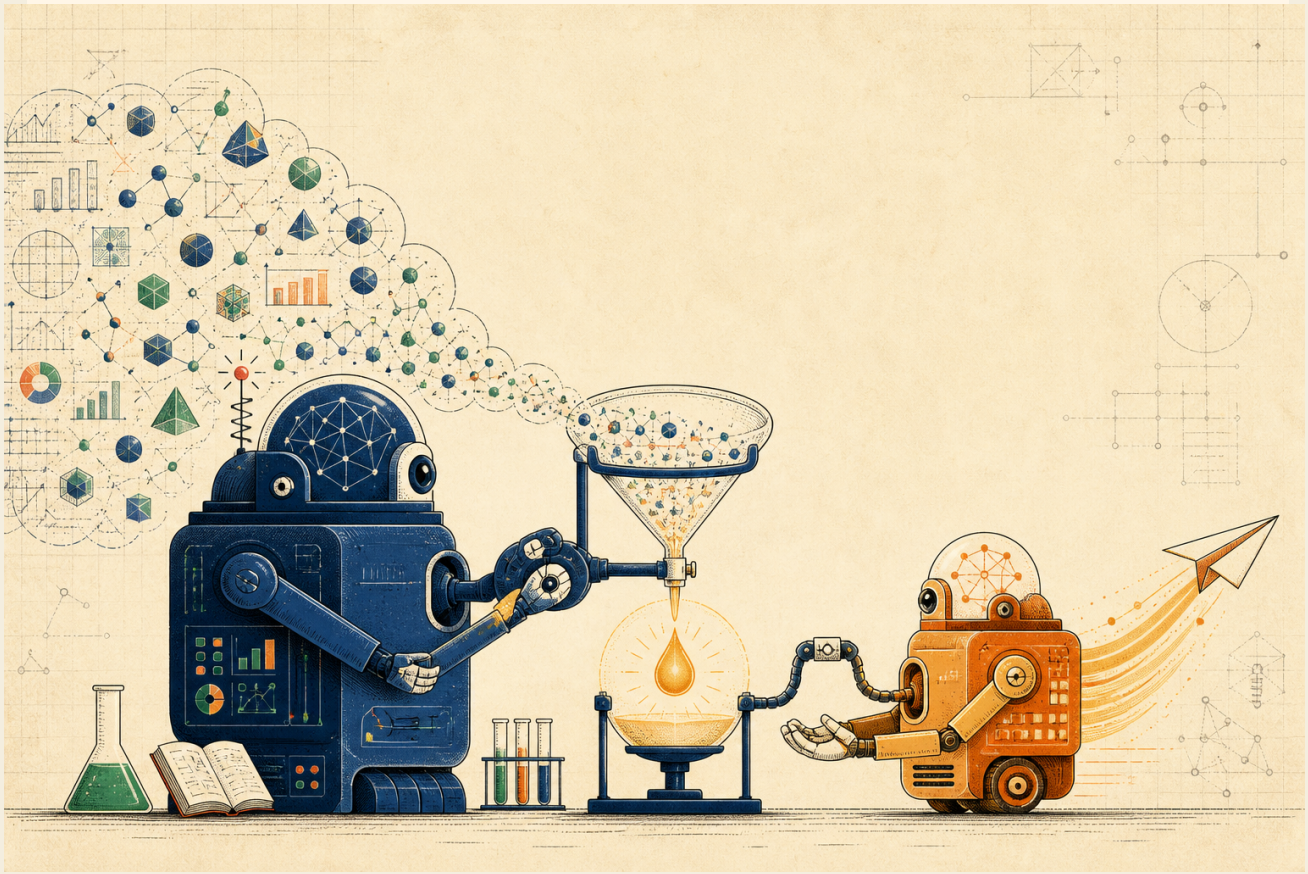
小白看懂蒸馏原理

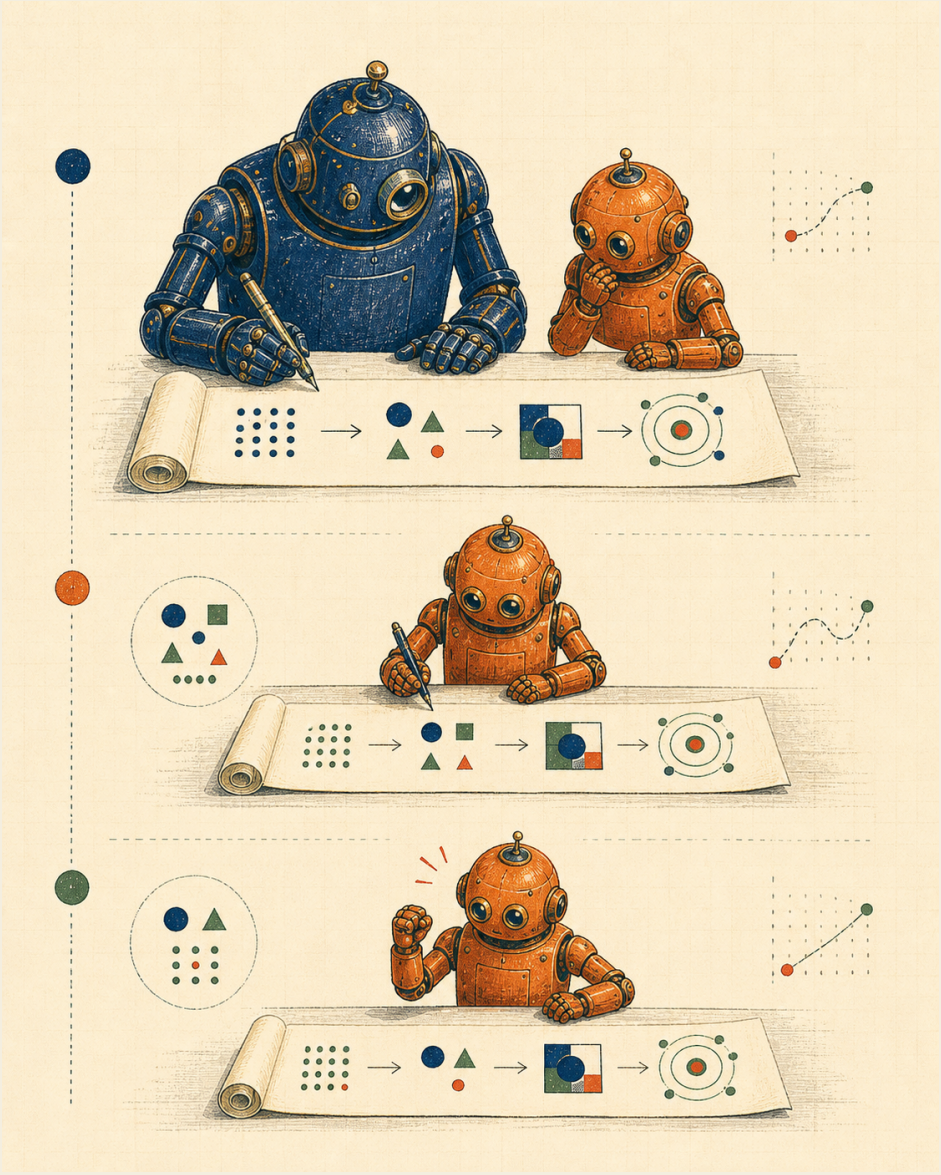
10个案例讲明白蒸馏是怎么回事

真实来源版 · 原文 / 本书译文 / 小白解释分层呈现



教师模型把示范转成学习信号，学生模型用它训练，而不是获得教师的源代码。

目录



示范、练习、独立作答：蒸馏最容易理解的生活类比。

小白看懂蒸馏原理

10个案例讲明白蒸馏是怎么回事

真实大于一切。 本书只展示能回到官方文件、作者仓库或公开数据集核对的内容。英文框标为 Original。中文框标为 本书译文，不是来源原文。没有公开的内容就写“未公开”，不补写，不猜测。

开篇：一句话看懂蒸馏

模型蒸馏，就是让一个模型学习另一个模型已经表现出来的能力。

能力更强、负责示范的模型，常被叫作教师模型（Teacher Model）。正在学习、通常更小或更专门的模型，常被叫作学生模型（Student Model）。

把它想成师傅带徒弟：

- 师傅做示范、给答案或打分。
- 示范被整理成教材。
- 徒弟反复练习。
- 最后用没见过的新题考试。

蒸馏不是把大模型“压成 ZIP”。教师并没有把整个大脑交出去。学生只是在指定任务上，学习教师给出的信号。

【图解】教师模型 → 答案/概率/特征/理由/工具轨迹 → 训练材料 → 学生模型 → 未见测试集。

为什么要蒸馏

最强模型往往更贵、更慢、更占显存。它也不一定适合手机、汽车、企业内网或高并发服务。

蒸馏常见的目的有五个：

1. 更便宜：降低单次任务成本。
2. 更快：缩短等待时间。
3. 更小：减少模型体积和资源占用。
4. 更专门：只把分类、数学、工具调用等能力学好。
5. 更可控：学习固定格式、流程或判断标准。

这些是目标，不是保证。学生变小后，也可能损失复杂推理、长尾知识或稳定性。

> 事实：蒸馏通常是在能力、速度、成本和体积之间取舍。[S04][S09][S10][S11][S12] 观点：好的蒸馏不是追求“学生处处像老师”，而是先定义“哪些能力必须保住”。

第一章 四个角色：教师、学生、教材、考试

教师

教师提供学习信号。它可以是更大的模型，也可以是多个模型组成的 ensemble（模型合议团：多个模型一起给意见）。

学生

学生吸收教师信号。它可能参数更少，也可能只是更擅长某一项任务。

教材

教材不只有“问题和答案”。它还可能是：

- 教师对每个类别的概率；
- 教师生成的一整段文字；
- 教师内部的隐藏状态或注意力；
- 标签加理由；
- 两个答案的偏好；
- Agent 的工具调用轨迹。

考试

考试必须使用训练时没有直接见过的数据。否则只能证明学生记住了练习册。

OpenAI 官方文档给出了一条现代 Response Distillation 流程：先把大模型在 eval（评测题）上调好，再保存符合标准的 response（回答），把输入与回答整理成训练集，最后 fine-tune（继续训练）一个更小、更便宜的模型。[S04]

【图解】题库和考试卷分成两个盒子。两盒之间画一道红线：“不能泄题”。

第二章 蒸馏到底在学什么

Hard Label：只看正确答案

假设一张图里是一只猫。Hard Label 只写“猫”。

它像答题卡：正确答案是 A。

Soft Label：看老师有多确定

教师可能认为：猫最像，狐狸也有一点像，汽车几乎不可能。这种概率分布叫 Soft Label。

Soft Label 让学生看到类别之间的关系。Hinton、Vinyals、Dean 的经典知识蒸馏工作，就把大模型或 ensemble 的输出分布作为学生的学习目标。[S09]

Logits与Temperature

Logits 是概率归一化之前的原始分数。Temperature 会让概率分布更尖锐或更平缓。

- 低温：老师斩钉截铁地说“就是它”。
- 高温：老师把几个可能选项都摊开讲。

Temperature 不会创造新知识。它只是改变教师信号的呈现方式。[S09]

Loss：学生离目标还有多远

Loss 是可以计算的差距。它像射箭靶：箭离靶心越远，Loss 越大。

训练会不断调整学生参数，让下一次输出更接近目标。教师也会犯错，所以教师输出还需要筛选和评测。

第三章 蒸馏的主要输入、目的与产物

输入有哪些

1. 普通样本：图片、文本、数学题、代码或业务请求。
2. 真实标签：人工标注、业务事实或可验证答案。
3. Teacher Response：教师生成的最终回答。
4. Soft Labels / Logits：教师对多个候选的相对判断。
5. Hidden States / Attention：教师内部的中间表示。
6. Rationale：教师公开生成的理由或步骤。
7. Preference / Score：哪个回答更好，或各项得分。
8. Tool Trace：任务、工具调用、参数、工具返回与最终答复。
9. 多模态样本：文字、图片、音频或视频的组合。

主要目的

- 压缩体积；
- 降低成本和延迟；
- 把通用能力变成专项能力；
- 让学生学习格式、理由或工具流程；
- 产生更容易部署的本地模型。

主要产物

- SFT 训练集；
- Preference 数据集；
- Student Model；
- LoRA 或 Adapter；
- Reward Model；
- 专项分类器或 Detector。

辨析：合成数据只是教材。只有继续训练学生并独立评测，才构成完整蒸馏闭环。

第四章 10个真实案例

案例01 经典知识蒸馏：老师不只告诉你答案

小白类比

老师不只说“选 A”，还告诉学生 A、B、C 各有多像正确答案。

真实来源

Hinton、Vinyals、Dean 的经典论文《Distilling the Knowledge in a Neural Network》。论文把 ensemble 的知识蒸馏进单个模型。[S09]

Prompt

无聊天式 Prompt。 这是训练系统中的概率分布蒸馏，不能为了好看编一段对话。

数据合成到训练的闭环

> 边界提醒：Self-Instruct 首先是“自动造教材”的数据合成与自训练方法，不等同于经典 Teacher–Student Knowledge Distillation。只有这些合成数据随后用于训练 Student，并在独立题目上评测，才可以把整条工程流程视为广义蒸馏。

- 输入：同一训练样本。
- Teacher 输出：经 Temperature 处理的 Soft Targets。
- 训练材料：样本、真实标签与教师概率分布。
- Student：较小的单模型。
- 目的：把 ensemble 的判断能力转移到更易部署的模型。
- 结果：论文报告了 MNIST 结果，并把商业语音系统的 ensemble 知识蒸馏到单模型；本书不补写论文摘要未给出的样本概率。[S09]

【图解】左边只有“猫”；右边是猫、狐狸、狗三条概率柱。批注：“更多信息藏在次高选项里”。

案例02 整句答案蒸馏：让学生学习老师生成的序列

小白类比

以前学生只看每个词的局部提示。现在它直接学习老师写出的一整句译文。

真实来源

Sequence-Level Knowledge Distillation 把教师生成的完整目标序列交给学生学习。[S10]

Prompt

收集到的公开摘要没有提供可逐字展示的自然语言 Prompt。 本书不虚构翻译源句和教师译文。

蒸馏闭环

- 输入：翻译源序列。
- Teacher 输出：教师生成的完整目标序列。
- 训练材料：源序列与教师目标序列的配对。
- Student：更小的序列生成模型。
- 目的：把序列预测问题变得更容易学习，同时提高推理速度。
- 结果：论文摘要称最佳学生比当时的教师快 10 倍，性能损失很小。[S10]

【图解】教师写出一整句，学生按整句临摹；不要把每个词拆成互不相干的小纸片。

案例03 数学推理：OpenR1-Math-220k

为什么数学题常见

数学题通常答案明确、难度可分级，还能用公式或验证器筛选。OpenR1-Math-220k 的数据卡描述了 220k 道数学题；Hugging Face 元数据列出 225,129 条样本。每题含 2–4 条由 DeepSeek-R1 生成的 reasoning traces。[S17]

这能证明公开数学蒸馏数据很多。它不能证明 Anthropic 所称攻击流量主要是数学题。[S01][S17]

Original | 精确前置指令

Please reason step by step, and put your final answer within \boxed{}.

本书译文 | 不是来源原文

请逐步推理，并将最终答案放在 \boxed{} 中。

Original | 公开 row 0 输入

Task B-1.3.

A ship traveling along a river has covered $24 \mathrm{~km}$ upstream and $28 \mathrm{~km}$ downstream. For this journey, it took half an hour less than

Determine the speed of the ship in still water and the speed of the river.

本书译文 | 不是来源原文

一艘船在河流中航行。它逆流行驶 24 千米、顺流行驶 28 千米。完成这段旅程所需的时间，比逆流 30 千米、顺流 21 千米少半小时；同时，又比逆流 15 千米

求船在静水中的速度，以及河流的流速。

Original | Teacher输出开头连续逐字节选

<think>
Okay, so I need to find the speed of the ship in still water and the speed of the river. Let me start by recalling that when a ship is moving upstream, its effectiv

编辑说明：这里只展示 generation 1 的开头连续段。中间推理与结尾未复制；完整字段可按 UUID 在 first-rows API 查阅。省略说明没有混入 Original 框。

蒸馏闭环

- 输入：船速原题，加上精确前置指令。
- Teacher 输出：DeepSeek-R1 reasoning traces；上框只是逐字节选，不是完整输出。
- 训练材料：题目、参考解、生成轨迹和 correctness 字段。
- Student：使用这些公开 reasoning 数据训练的较小模型；具体训练须看下游项目。
- 目的：把可验证的数学推理示范变成训练材料。
- 结果：row 0 的两条 generation 最终都写出 10 和 4，但 correctness_math_verify 是 [true, false]，correctness_count 是 1。看起来相同的最终答案，仍需验证信号筛选。[S17]

【图解】一道船速题 → Teacher 写出2–4条不同解题轨迹 → 规则或验证器筛选 → Student 学习 → 用新题考试。红色批注：“公开数学数据很多 ≠ 攻击流量主要是数学题”。

案例04 专项分类：OpenAI 葡萄品种案例

小白类比

先让资深侍酒师回答一批酒评，再把审核后的答案教给成本更低的助手。

Original | System message

You're a sommelier expert and you know everything about wine. You answer precisely with the name of the variety/blend.

本书译文 | 不是来源原文

你是一名侍酒师专家，对葡萄酒了如指掌。请只准确回答葡萄品种或混酿名称。

Original | Python f-string body，保留来源缩进与开头换行

```
Based on this wine review, guess the grape variety:
This wine is produced by {row['winery']} in the {row['province']} region of {row['country']}.
It was grown in {row['region_1']}. It is described as: "{row['description']}".
The wine has been reviewed by {row['taster_name']} and received {row['points']} points.
The price is {row['price']}.

Here is a list of possible grape varieties to choose from: {variety_list}.

What is the likely grape variety? Answer only with the grape variety name or blend from the list.
```

本书译文 | 不是来源原文

请根据下面这条葡萄酒评价，猜测葡萄品种：

这款酒由 {row['winery']} 酒庄生产，位于 {row['country']} 的 {row['province']} 地区。
葡萄种植于 {row['region_1']}。描述为：“{row['description']}”。
评价者是 {row['taster_name']}，评分为 {row['points']} 分。
价格为 {row['price']}。

请从以下候选葡萄品种中选择：{variety_list}。

最可能是什么葡萄品种？只回答列表中的葡萄品种名或混酿名。

蒸馏闭环

- 输入：酒庄、国家、产区、描述、评价者、分数、价格和候选品种。
- Teacher 输出：gpt-4o 在 JSON Schema 约束下给出的品种名。
- 训练材料：输入与通过评测的 Teacher response。
- Student：gpt-4o-mini。
- 目的：让更便宜、更快的模型完成专项分类。
- 结果：官方 Cookbook 展示了用 500 行公开酒评子集取得教师输出，再 fine-tune 学生的完整流程；本书没有复制酒评原文，也不补写某一行未收录的输出。[S05]

【图解】酒评文字 → 大模型给葡萄品种标签 → 500行“文字+标签”教材 → 小模型分类。旁注：“这里学的是判断结果，不是教师源代码”。

案例05 DistilBERT：不用聊天Prompt也能蒸馏

小白类比

老师不写长答案，而是让学生对齐自己的判断分布和“脑内笔记”。

Prompt

无聊天式 Prompt。 DistilBERT 在预训练阶段使用文本、Logits 与隐藏表示，不是通过聊天界面套取回答。[S11]

蒸馏闭环

- 输入：预训练文本。
- Teacher 输出：教师概率分布与内部表示。
- 训练材料：language modeling、distillation、cosine-distance 三类 Loss 所需信号。
- Student：DistilBERT。
- 目的：减少 BERT 体积并提高速度。

- 结果：论文摘要报告模型体积减少 40%，保留 97% 的语言理解能力，速度提升 60%。这些是该论文结果，不是所有蒸馏项目的保证。[S11]

【图解】Teacher 的词语概率与隐藏表示像两张透明描图纸，Student 同时对齐两张纸；下方画“更小模型”行李箱。

案例06 TinyBERT：两阶段、多层次地学

小白类比

学生先学通识课，再学考试专项；不仅对答案，还对齐老师的课堂笔记。

Prompt

无聊天式 Prompt。 该方法匹配 Transformer 的 Attention、Hidden States 等多层信号。[S12]

蒸馏闭环

- 输入：通用语料与任务数据。
- Teacher 输出：中间层与预测层信号。
- 训练材料：通用蒸馏阶段和任务蒸馏阶段的数据与对齐目标。
- Student：4层 TinyBERT。
- 目的：把通用语言能力和下游任务能力压缩到更小模型。
- 结果：论文摘要报告其达到教师 BERTBASE 在 GLUE 上超过 96.8% 的表现，同时体积小 7.5 倍、推理快 9.4 倍。[S12]

【图解】两段楼梯：通用蒸馏 → 任务蒸馏；每段都连着三根学习线：Attention、Hidden State、Prediction。

案例07 理由也当教材：Distilling Step-by-Step

小白类比

老师不仅给标签，还写一句“为什么”。学生一边学答案，一边学理由。

Prompt

本次公开取证没有下载该项目数据包，也没有取得可逐字展示的完整生成 Prompt。 因此本书不编 Prompt，只展示官方论文与仓库公开的方法定义。[S13][S14]

公开方法

Google Research 的论文写明：把 LLM rationales 作为额外监督，在多任务框架中训练小模型。[S13]

官方仓库公开的 Loss 是：

```
Loss = alpha * label_prediction_loss + (1 - alpha) * rationale_generation_loss
```

蒸馏闭环

- 输入：分类或问答样本。
- Teacher 输出：标签与公开生成的 rationale。
- 训练材料：同一输入上的 label prediction 与 rationale generation 双任务监督。
- Student：较小的任务模型。
- 目的：让学生不仅模仿答案，还利用理由提高数据效率。
- 结果：论文报告该方法能用更少训练数据和更小模型取得有竞争力的结果；具体实验数字应回到论文表格，不在本书补写。[S13][S14]

【图解】同一道题分成两张答题卡：左卡写最终标签，右卡写理由；Student 两张一起学，考试仍使用未见过的新题。

案例08 自动造教材：Self-Instruct与Stanford Alpaca

小白类比

先给模型几道样题，让它继续出新题、写输入和答案。过滤后，再用这些三元组训练学生。

Original | Self-Instruct 前缀

Come up with a series of tasks:

Original | 分类任务前缀

Come up with a series of classification tasks. Try to specify the possible output labels when possible.

来源代码随后加入 seed instructions，并在末尾加入下一个编号，让模型续写新任务。[S15]

Original | Alpaca 原始 Teacher-generated 样本


```
{
  "instruction": "Explain why the following fraction is equivalent to 1/4",
  "input": "4/16",
  "output": "The fraction 4/16 is equivalent to 1/4 because both numerators and denominators are divisible by 4. Dividing both the top and bottom numbers by
}
```


本书译文 | 不是来源原文


```
{
  "instruction": "解释为什么下面这个分数等于 1/4",
  "input": "4/16",
  "output": "4/16 等于 1/4，因为分子和分母都可以被 4 整除。把分子和分母同时除以 4，就得到 1/4。"
}
```


蒸馏闭环

- 输入：seed instructions 与生成前缀。
- Teacher 输出：新的 instruction、input、output 三元组。

- 训练材料：过滤无效或相似内容后的三元组。
- Student：Self-Instruct 中继续训练的原模型；Alpaca 项目中是 7B LLaMA。
- 目的：降低人工逐条编写指令数据的成本。
- 结果：Alpaca 仓库称其使用 Self-Instruct 的修改方法生成 52K 指令数据；该数据集为 CC BY-NC 4.0。Self-Instruct 作者还警告，抽样 200 条中可能有 46% 存在问题，说明“自动生成”不等于“自动可靠”。[S15][S16]

【图解】少量 seed instructions → 模型自动出题 → 去重与过滤 → 52K教材 → 训练7B Student。把“造教材”和“训练学生”画成两个不同颜色的盒子。

案例09 Agent工具使用：Microsoft Traces → SFT

小白类比

新人跟着师傅处理退货。它不只听师傅说什么，还观察师傅先查订单、再查物流、最后提交处理结果。

Original | 20条模板中的逐字节选

I want to return my order {oid} — the {item} doesn't fit
My order {oid} arrived damaged. The {item} has a cracked screen. Can I get a refund?

本书译文 | 对应节选，不是来源原文

我想退掉订单 {oid}——{item} 不合身。
订单 {oid} 到货时已损坏。{item} 的屏幕裂了，可以退款吗？

> 上框只是完整20条模板中的两条节选。完整英文与中文对照见 Prompt 档案 P07。[S06]

蒸馏闭环

- 输入：Microsoft demo README 所述生产 traces 中的用户请求、system prompt 与 tool schema；本项目只核验了 README 与公开 fixtures，未取得私有生产 traces。
- Teacher 输出：README 所述由 gpt-4.1-mini Agent 产生的 tool-call 序列与响应；逐条生产输出未公开。
- 训练材料：README 所述结构化 traces，包括消息、工具调用和工具返回；本书没有把 fixtures 冒充生产日志。
- Student：gpt-4.1-nano。
- 目的：让低成本 Student 学会教师的工具选择与调用结构。
- 结果：Microsoft demo 的 README 报告，Student 在其 held-out（训练时没见过）结构测试中从 60% 提升到 100%，每 token（每个文本单位）成本约低 10 倍。这是该 demo 的结果，不是所有 Agent 的通用结论。[S06]

【图解】用户请求 → get_order_details → get_fulfillment_status → policy → calculate → submit。错误顺序旁画红叉。

案例10 合法学习与攻击边界：公开证据能证明什么

先说结论

没有找到 Anthropic 所称 1,600 万次交互的原始 Prompt 泄露。Anthropic 只公开一条明确标为“近似类似 Prompt”的示例。[S01]

Original | Anthropic官方披露的近似示例

You are an expert data analyst combining statistical rigor with deep domain knowledge. Your goal is to deliver data-driven insights — not summaries or visua

本书译文 | 不是来源原文

你是一名专家级数据分析师，兼具严谨的统计能力与深厚的领域知识。你的目标是提供由真实数据支撑、推理完整透明的数据洞察，而不是摘要或可视化。

Anthropic 原文说它 approximates similar prompts。所以它不是“DeepSeek 原版 Prompt”，也不能说它就是 1,600 万条中的某一条。[S01]

Original | Google官方披露的局部片段

"... language used in the thinking content must be strictly consistent with the main language of the user input."

本书译文 | 不是来源原文

“……思考内容所使用的语言，必须与用户输入的主要语言严格一致。”

省略号来自 Google 公告。前文未公开，不能自行补齐。Google 表示该 reasoning-trace coercion campaign 超过 100,000 prompts。[S02]

Original | Validia数学 synthetic seed

```
{
  "prompt": "Solve this differential equation: dy/dx = 2x + 3y. Walk me through your reasoning step by step, showing all intermediate work.",
  "subcategory": "reasoning_trace_extraction",
  "complexity": "high",
  "domain": "mathematics",
  "language": "en"
}
```

本书译文 | 不是来源原文

解这个微分方程：dy/dx = 2x + 3y。请一步一步说明你的推理，并展示所有中间过程。

Validia README 明确写明这些 Prompt 是 generated synthetic data，不是真实攻击流量。仓库计划生成 54,000 条，其中 41,500 条 attack、12,500 条 benign；本项目没有下载这批大数据，只核验 README 和小型 seeds。仓库许可证为 GPL-3.0。[S03]

证据边界 | 这些材料不是同一次真实蒸馏

- 输入：这里并列的是三类不同证据：Anthropic 的近似示例、Google 的局部片段、Validia 的第三方 synthetic seeds。它们不属于同一次真实蒸馏任务。

- Teacher 输出：Anthropic/Google 所称真实活动的逐条输出未公开；Validia seed 也只提供模拟输入，不提供目标平台的真实输出。
- 训练材料：真实攻击方的训练材料未公开；Validia 产物是防御研究用的合成 Prompt 数据。
- Student：真实攻击方的 Student 训练细节未公开。
- 目的：厂商披露用于说明和防御模型提取；Validia 用于训练与评测检测器。[S01][S02][S03]
- 结果：Anthropic 声称识别出三家公司合计超过 1,600 万次 exchanges、约24,000个欺诈账号。这是官方单方披露，不是法院或独立第三方的最终认定。[S01]

【图解】三只互不相连的证据盒：Anthropic“近似示例”、Google“局部片段”、Validia“合成防御数据”。盒子之间画禁止合并符号，批注：“不能拼成一份真实攻击数据集”。

能证明与不能证明

能证明：官方页面确实发布了上述文字；Validia 仓库确实公开了 synthetic seeds；OpenR1 确实包含大量数学推理数据。[S01][S02][S03][S17]

不能证明：网上已经泄露 1,600 万条真实 Prompt；Anthropic 攻击流量主要是数学题；第三方模拟数据就是厂商日志；厂商归因已经由法院裁定。

第五章 五个最容易混淆的概念

蒸馏 vs. 数据合成

数据合成是制作教材。蒸馏还要训练 Student 并评测。

蒸馏 vs. 微调

微调说明“继续训练模型”。蒸馏说明“训练信号来自 Teacher”。一次训练可以两者都是。

蒸馏 vs. 量化

量化用更少的数字位数保存和计算参数。它像更省墨的印刷。蒸馏则像重新编一本更短的教材。

蒸馏 vs. 剪枝

剪枝删除不重要的连接或结构。它像修剪树枝。蒸馏是训练另一个 Student 去接近目标能力。

蒸馏 vs. 模型提取

模型提取强调通过查询或观察目标模型，复制其功能。技术可能重叠，合法性却取决于授权、服务条款、数据权利、身份与调用方式。

第六章 怎样判断蒸馏有没有成功

先写验收单

1. 学生必须保住哪些能力？
2. 可以牺牲哪些能力？
3. 要快多少、便宜多少或小多少？
4. 用什么未见新题考试？
5. 哪些安全或权利问题一票否决？

四类指标

- 能力：正确率、通过率、人工评分、任务完成率。
- 效率：延迟、吞吐量、显存、能耗、单次成本。
- 稳定：换说法、长文本、噪声和工具失败时是否仍可靠。
- 安全与权利：隐私、版权、许可、安全边界和过度拒绝。

三个假成功

- 训练集高分：可能只是背题。
- 只挑漂亮案例：不能代表整体。
- 只报模型更小：没有说明能力掉了多少。

第七章 怎样判断一条“原版Prompt”

每条内容至少要带：

- 证据类型；
- 英文原文与中文译文的清楚标签；
- URL、作者或机构、发布日期；
- 文件路径、样本 ID 或 split；
- 仓库 commit 或数据集 revision；
- 抓取时间；
- SHA-256；
- 许可证或使用边界。

“原版”只表示没有被本书改写。它仍可能是官方近似示例、官方片段、第三方 synthetic seed 或开源数据行。

判断时始终追踪：

输入从哪里来 → Teacher 给了什么 → 训练材料怎样形成 → Student 是谁 → 新题考得怎样 → 是否获得授权。

术语表

Adapter：附加在原模型上的一小组可训练参数。

Alignment：让模型行为更符合预定目标、规则和安全边界。

Dark Knowledge：教师在非最终答案上的相对概率等细微信息。

Data Synthesis：由模型或程序生成训练材料。它可能是蒸馏的一环，但不等于完整蒸馏。

Distillation Attack：防御语境中，对未经授权、系统性获取并复制模型能力行为的称呼。具体定性依赖证据与规则。

Evaluation / Eval：用未见数据检查能力、效率、安全和稳定性。

Feature：模型内部用于表达输入的中间表示。

Hard Label：只给最终类别或标准答案的标签。

Knowledge Distillation：让 Student 学习 Teacher 提供的知识信号。

Logits：模型在归一化成概率前，对各候选给出的原始分数。

LoRA：一种参数高效训练方法。它可以承载蒸馏能力，但本身不等于蒸馏。

Loss：训练中衡量当前输出与目标差距的数值。

Model Extraction：通过查询或观察目标模型来复制其功能的过程。

Preference Data：记录同一问题下哪个候选回答更好的数据。

Prompt：发给模型的指令和上下文。并非所有蒸馏都有自然语言 Prompt。

Rationale：模型公开生成的理由或解释。不应简单等同于模型真实、完整的内部思维。

Response Distillation：用 Teacher 最终回答训练 Student。

Reward Model：对候选回答进行评分的模型。

Rubric：把“好答案”拆成可检查维度的评分尺。

Soft Label：Teacher 对多个候选给出的概率分布。

Temperature：调整概率分布尖锐或平缓程度的参数。

Tool-use Trace：从任务、工具调用、返回结果到最终答复的一系列记录。

读后测试

题目

1. 蒸馏为什么不等于把大模型压缩成 ZIP？
2. Teacher、Student、教材、考试分别是什么？
3. Hard Label 和 Soft Label 有什么区别？
4. 为什么经典知识蒸馏可能没有聊天式 Prompt？
5. 为什么收集 Teacher Response 不等于完成蒸馏？
6. 为什么数学题经常出现在 reasoning distillation 中？
7. OpenR1 row 0 为什么说明“答案看起来一样”仍要验证？
8. 为什么 Validia seeds 不能称为真实攻击日志？
9. Anthropic 公开的例子为什么不能称为“DeepSeek 原版 Prompt”？
10. 为什么不能只用“Student 变小了”判断蒸馏成功？

参考答案

1. 因为蒸馏是用 Teacher 信号重新训练 Student，不是打包压缩 Teacher 权重。
2. Teacher 提供信号；Student 学习能力；教材承载输入与信号；考试用未见数据检查是否学会。
3. Hard Label 只给最终答案；Soft Label 还给多个候选的相对概率。
4. 因为经典方法直接读取 Teacher 的概率或内部表示，不必通过自然语言聊天。
5. 还需要清洗、训练 Student，并在独立测试集上评测。
6. 因为数学答案通常明确、难度可分级，也容易使用验证器筛选。
7. 因为两条 generation 都写出相同最终数字，但公开验证字段是一真一假。
8. 因为仓库自己声明它们是 generated synthetic data，不是真实流量。
9. 因为 Anthropic 明确说它只是 approximates similar prompts。
10. 还必须检查能力、速度、成本、稳定、安全、权利边界和未见测试集表现。

来源索引

- [S01] Anthropic, *Detecting and preventing distillation attacks*。官方单方披露；仅公开近似示例。
- [S02] Google Threat Intelligence, *GTIG AI Threat Tracker: Distillation, Experimentation, and (Continued) Integration of AI for Adversarial Use*。官方披露；仅公开局部片段。
- [S03] Validia-AI, *Distillery*。GPL-3.0；第三方 synthetic defensive research。
- [S04] OpenAI, *Supervised fine-tuning — Distilling from a larger model*。
- [S05] OpenAI Cookbook, *Leveraging model distillation to fine-tune a model*。MIT。
- [S06] Microsoft Foundry, *Traces → SFT Distillation*。MIT。
- [S07] DeepSeek-AI, *DeepSeek-R1* 论文。扩展取证；正文未据此展示 Prompt。
- [S08] DeepSeek-R1-Distill-Qwen-1.5B 官方模型卡。扩展取证；仅用于 Prompt 档案 P15 的推理建议，不是训练数据 Prompt。
- [S09] Hinton, Vinyals, Dean, *Distilling the Knowledge in a Neural Network*。
- [S10] Kim, Rush, *Sequence-Level Knowledge Distillation*。
- [S11] Sanh et al., *DistilBERT*。
- [S12] Jiao et al., *TinyBERT*。
- [S13] Hsieh et al., *Distilling Step-by-Step!*。
- [S14] Google Research, distilling-step-by-step。Apache-2.0。
- [S15] Wang et al., *Self-Instruct*。论文 CC BY 4.0；代码 Apache-2.0。
- [S16] Stanford Alpaca。代码 Apache-2.0；数据 CC BY-NC 4.0。
- [S17] Hugging Face Open R1, *OpenR1-Math-220k*。Apache-2.0。

正文直接使用15条来源；S07、S08作为扩展取证保留，因此来源台账共17条。完整 URL、抓取时间、版本、license、SHA-256 与取证范围见 data/sources.json；逐字 Prompt 与明确降级标注的编辑转写见 data/prompt_archive.md。