ChatGPT

On-Policy Distillation 的前世今生

概念背景与发展历史

知识蒸馏 (Knowledge Distillation) 最初由 Hinton 等人在 2015 年提出，是一种将**大模型**的知识提炼给**小模型**的技术 [1]。教师模型（teacher）的输出软目标用于监督学生模型（student）的训练，使小模型在较小规模下逼近大模型性能。这一思想最早应用于监督学习领域，如图像分类和语言模型压缩。其后，“蒸馏”理念被引入强化学习 (RL)：2015 年末，DeepMind 的 Andrei Rusu 等人提出了**策略蒸馏 (Policy Distillation)** [2]。他们证明可以将深度强化学习智能体的策略提取出来，训练一个**更小但性能同样专家级**的新网络。此外，该方法还能**将多个任务的策略合并到单一网络中** [2]。几乎同时，Emilio Parisotto 等人提出了**Actor-Mimic** 算法 [3]（发表于 2016 年 ICLR），利用多个专家教师（预训练的深度 Q 网络）指导一个学生网络**同时学习多项任务** [3]。这些奠基性工作标志着蒸馏技术在强化学习和多任务学习中的首次重大应用，使得原本需要庞大网络和长时间训练的 RL 策略可以被更小的模型高效复现。

在此之后，研究者开始探索**蒸馏与「on-policy/off-policy」训练范式的结合**。传统蒸馏通常是**离线的 (off-policy)**，即学生在教师预先生成的固定数据上模仿教师输出（类似于**行为克隆**或监督微调）。这种**off-policy蒸馏**训练稳定、高效，但存在**训练分布与实际使用分布不匹配**的问题：学生只见过教师走访的情境，缺乏应对自身错误的经验 [4]。相反，强化学习则是**在线的 (on-policy)**，智能体通过与环境交互从**自身经历**中学习，能够直接针对自己的错误进行调整，但 RL 通常**反馈稀疏**（每个序列仅一个奖赏信号）且训练成本高 [4]。为兼具两者优点，近年来产生了**On-Policy Distillation**的概念：**学生模型使用自身策略与环境交互**，然后**由强大的教师模型对每一步行为给予密集反馈**，从而实现高效训练 [5]。这个理念可类比为**“大师逐招点评”**：学生自行下棋（产生trajectory），教师对每一步棋给予打分或指正，而非仅在终局给胜负评价 [5]。On-Policy Distillation 的提出旨在**弥合 RL 与监督蒸馏的鸿沟**——既保留学生从自身决策中学习、纠正失误的**相关性**（on-policy优点），又获得每一步都有指导信号的**高密度监督**（蒸馏优点） [5]。这一概念的雏形可以追溯到 2011 年的**DAgger 算法**（Dataset Aggregation）**【注】：人类专家会纠正学习代理在自己轨迹中犯下的错误，实现**on-policy模仿学习**来避免错误累积。同理，On-Policy Distillation 将“专家”由人类替换为强教师模型，用教师策略指导学生走出自己轨迹中的失误。

经过多年的探索，**On-Policy Distillation**在 2023 年前后开始走向成熟。Google Brain 的 Agarwal 等人（2024 年 ICLR）提出**广义知识蒸馏 (Generalized KD, GKD)** [1] [6]，正式将“学生自生成数据+教师反馈”引入序列模型蒸馏领域，以解决**自回归生成中训练分布偏差**的问题。他们的方法允许使用多种损失函数（如KL散度变体），并且**可以无缝结合RL微调（RLHF）** [6]。同期，业界实践（如阿里巴巴达摩院的 Qwen3 技术报告）也验证了**On-Policy Distillation 的巨大效益**：相比纯RL微调，小模型通过教师逐步指导，**以约1/10的算力达到更高性能** [7] [8]。可以说，On-Policy Distillation 的概念在最近几年经历了从萌芽到落地的重要发展历程，既有顶会论文奠基其理论，也有工业界大模型训练实践佐证其价值。

关键论文与奠基性工作

下面梳理On-Policy Distillation及相关方法的发展里程碑和代表性论文：

- **2015 年 – 知识蒸馏 (Knowledge Distillation)** [1]：「Distilling the Knowledge in a Neural Network」(Hinton 等)。提出教师-学生网络蒸馏框架，用软目标训练小模型复现大模型性能，开启了模型压缩的新方向。
- **2015 年 – 策略蒸馏 (Policy Distillation)** [2]：Rusu et al., 提交 ICLR 2016. 首次将蒸馏用于强化学习策略压缩。**贡献**：提炼出深度Q网络(DQN)等RL智能体的策略，训练一个参数量大幅减少但仍**专家级**的

学生策略网络；并演示可将多个教师策略**合并**成单一学生网络。多任务蒸馏的学生在Atari游戏上甚至**超过了**各单任务教师和一个联合训练的大模型 [2]。

- **2016 年 – Actor-Mimic** [3]：Parisotto et al., ICLR 2016. 提出多任务和迁移学习的新方法。**贡献**：利用**深度RL + 模型压缩**技巧，让一个学生网络在多款游戏上同时学习决策策略，**模仿多个专家教师**（预先训练的DQN)。结果显示学生网络学到的表示具有**迁移泛化**能力，在无额外指导下加速了新游戏的学习 [3]。这是**多教师蒸馏**和**跨任务知识转移**的早期成果。
- **2011 年 – DAgger 算法**（行为克隆改进）：Ross et al., NIPS 2011. **贡献**：不是蒸馏工作，但奠定了**on-policy 模仿学习**思想基础。通过在机器人、游戏等任务中**让专家纠正学生的决策**，DAgger 解决了纯行为克隆（off-policy模仿）中的误差积累问题。这一思路与日后 On-Policy Distillation 不谋而合：都强调用**专家在学生访问的状态下提供示范**来迭代提升策略。
- **2017 年 – AlphaGo Zero 自我蒸馏**：Silver et al., Nature 2017. **贡献**：提出不使用人类数据、纯自我对弈提升围棋AI性能的AlphaGo Zero算法。每一步通过蒙特卡洛树搜索（MCTS）改进当前策略，然后**训练网络去模仿MCTS所得的强化策略**。该循环可视为一种**on-policy策略蒸馏**：网络从自己对弈（on-policy）的局面中学习，更强的搜索策略充当教师提供监督。这实现了每次自我博弈后对策略的蒸馏提升，最终产出超越人类的围棋策略网络。随后这一**专家迭代**思想也在其它博弈中推广。
- **2018 年 – Kickstarting Deep RL** [9] [10]：Schmitt et al., NeurIPS 2018. **贡献**：扩展策略蒸馏到**连续在线RL**场景。提出“Kickstarting”方法，允许**预训教师边指导、学生边与环境交互学习**。没有架构限制，且学生最终能超过教师性能 [9]。在 DeepMind Lab-30 多任务基准上，Kickstarting 显著提高了样本效率：学生仅用**约1/10的训练步数**就达到了从零训练模型的性能，并最终**超出教师42%** [11] [10]。此外，一个学生还能同时借鉴**多个专家教师**（各专长不同任务），取得远超单教师的效果 [11]。Kickstarting 将**策略蒸馏损失**加入RL目标中，被证明是提升数据效率和加速收敛的有效手段。
- **2023 年 – 广义知识蒸馏 (Generalized KD)** [1] [6]：Agarwal et al., ICLR 2024. **贡献**：正式提出**On-Policy Distillation**框架用于自回归大模型压缩。学生模型在训练中**生成自身输出序列**，教师模型对这些**学生轨迹**提供反馈信号，从而消除传统蒸馏中训练分布与推理分布不符的缺陷 [1]。此外，作者引入**多样化损失**（如反向KL散度、JS散度等）来提高学生训练的灵活性，尤其当小模型容量不足以精确匹配大模型分布时，可选择更稳健的目标函数 [6]。该方法还能**无缝结合RL微调（RLHF）**过程，在压缩聊天模型时同时对齐人类偏好 [6]。实验在摘要、翻译、数学推理等任务上表明，相比传统蒸馏，GKD 显著提升了学生模型性能。
- **2024 年 – DistiLLM**：Ko et al., ICML 2024. **贡献**：针对大模型蒸馏提出改进的**偏置KL损失（Skewed-KL）**和**自适应离线数据采样**策略 [12]。在保证稳定训练的同时，DistiLLM 部分融入了学生生成的数据，从而提高蒸馏效率。据报道，该方法在多项基准上超越以往KD技术。它代表了**off-policy 蒸馏方向的新进展**，通过调整损失函数和采样分布，弥补纯on-policy方法可能出现的训练不稳定等问题。
- **2024 年 – Adversarial Moment-Matching Distillation** [13] [14]：Zhang 等, NeurIPS 2024. **贡献**：提出同时优化**on-policy**和**off-policy**目标的生成对抗蒸馏方法。该方法不直接最小化学生和教师分布差异，而是通过匹配**行动值函数的高阶矩**来传递知识 [15]。训练过程可视作一个**两玩家博弈**：一方调整学生策略，另一方（两个小网络）近似教师的on/off策略价值 [16] [17]。他们发现，将**学生实时生成的输出 (on-policy)** 与 **教师预生成或监督数据 (off-policy)** 的训练目标**混合优化**，能显著提升大模型指令遵循等任务的蒸馏效果 [13] [14]。在包含GPT-4反馈的评测中，这种**混合蒸馏**在多个数据集上取得了**最佳结果** [13] [14]。此工作揭示了同时利用两种信号的协同效应，并为蒸馏目标的设计提供了新思路。
- **2025 年 – Qwen3 On-Policy 蒸馏实践** [7] [8]：阿里巴巴达摩院 Qwen3 技术报告 (arXiv 2025). **贡献**：在大模型实际训练中验证了On-Policy Distillation的强大效果。Qwen3-8B 模型经过离线蒸馏微调后，分别使用**纯RL微调**和**on-policy 蒸馏微调**进行比较 [18]。结果表明：**蒸馏策略不仅性能更优**，而且所需算力只是RL方法的大约**十分之一** [7] [8]。例如，在数学推理基准AIME上，传统RL微调耗费17,920 GPU小时达到67.6分，而on-policy蒸馏仅用1,800 GPU小时就达到74.4分 [8]。该报告还指出，利用教师logits指导，学生模型拓展了探索空间，提高了复杂推理能力，而纯RL并未带来类似收益 [7]。Qwen3 的研究是业界**大模型持续学习**的代表案例，证明了On-Policy Distillation在实践中**提效增质**的潜力。

【注】 由于篇幅所限，此处未详述的一些相关工作还包括：OpenAI 在强化学习对话模型中使用RLHF + 蒸馏小模型的探索（如 2020 年总结任务中对大模型进行策略蒸馏）、DeepMind AlphaStar (2019) 在星际争霸AI中用

蒸馏结合人口机制压缩策略，以及近期Meta等在指令微调数据集公开和蒸馏领域的工作等等。这些共同丰富了On-Policy Distillation的发展图景。

强化学习与模仿学习中的应用实例

On-Policy Distillation及其相关蒸馏方法已在强化学习和模仿学习领域产生诸多应用：

- **强化学习中的策略压缩与迁移**：策略蒸馏最直接的应用是在训练耗时、模型庞大的RL智能体上进行**模型压缩**。例如，上述 DeepMind 工作将训练良好的 Atari 游戏策略提取到小网络中，在不损失性能的情况下大幅减少模型计算量 [2] 。这对部署在资源受限的平台（如移动设备、机器人）非常有价值。同样技术也用于**多任务学习**，把多个独立训练的策略合并进一个通用策略网络 [2] 。Actor-Mimic 即证明了单一模型可以在多游戏环境中表现出接近各专家的水平，并将不同任务的知识融会贯通，有助于**迁移学习** [3] 。

- **专家演示与模仿学习**：蒸馏本质上是一种模仿，他可以用来从专家（人类或强AI）行为中学习策略。在离线情况下，这就是**行为克隆**：学生模型在固定的专家数据集上训练（典型如无人车通过人类驾驶记录学习）。而on-policy蒸馏提供了**交互式模仿**的新范式。例如**DAgger 算法**的机器人导航应用中，学生一边试错，专家一边纠偏，这实际就是**“边学边蒸馏”**的过程。如今如果把专家替换成一个性能卓越的大模型，我们就可采用on-policy distillation来让学生模型学会专家擅长的技能，同时还能学会**从自身错误中恢复** [19] 。这种思想已用于无人驾驶、机器人操控等需要高可靠性的领域，能显著减少“长尾错误”的发生。

- **人类游戏AI与自我博弈**：AlphaGo Zero 展示的自我博弈+蒸馏策略取得巨大成功后，这一模式被推广到其它游戏AI中。例如国际象棋、德州扑克等环境下，研究者使用搜索或规划方法得到局部最优决策，再将其蒸馏进策略网络，以此加速策略改进。2022 年的**Expert Iteration**方法正是对这类过程的抽象：在每个迭代中用强“专家”（如搜索算法）指导“学徒”网络学习，然后学徒提升后再进行更深入的搜索，如此循环。这些都是**On-Policy策略蒸馏**理念在对抗游戏领域的体现，使AI能够快速接近甚至超越人类水准。

- **大型语言模型 (LLM) 微调**：近年最受关注的应用之一是在**大语言模型**的微调中运用蒸馏。由于直接用RLHF微调大模型代价极高，许多工作尝试**蒸馏出一个小型的、高效的对话或专用模型**。典型做法是先用强大的教师模型（如GPT-4或PaLM等）生成高质量指令响应数据，然后对较小的学生模型做**监督微调(SFT)**——这其实就是标准的**off-policy 知识蒸馏**流程 [20] 。例如 **Stanford Alpaca (2023)** 项目利用 OpenAI 模型生成52K指令响应，让一个7B参数的学生模型学会了**指令跟随**能力 [20] ；又如 **OpenAI Codex** 则通过蒸馏GPT的代码生成能力，将参数压缩数十倍以部署在VSCode插件中。这些案例显示蒸馏能将昂贵模型的专长“传承”下来。学界也有类似探索：Taori 等人（2023）证明小模型经过大模型指令数据微调后能在指令遵循上媲美原模型 [21] 。Guha 等人（2025）的 **OpenThoughts** 项目通过蒸馏提升了小模型在数学和科学问答上的推理能力 [20] ；Vedula 等人（2025）则成功**蒸馏医学笔记**信息抽取模型，用大模型教师指导小模型提取临床要点 [21] ；Ding 等人（2023）通过蒸馏构建了多轮对话小模型，大幅增强了其上下文对话连贯性 [22] 。这些应用遍及各个垂直领域，证实了蒸馏在LLM**专业化**上的有效性。值得注意的是，此类传统蒸馏由于是离线进行，也暴露出一些问题，比如学生模型可能**学到教师的语气风格却未真正掌握事实知识**（出现所谓“鹦鹉学舌”现象） [23] 。因此业界开始引入on-policy机制来改进LLM微调——让学生在**与实际用户或环境交互**时得到教师纠偏。例如最新的研究把RLHF产生的反馈信号**密集地施加在学生生成的每个token上**，而非仅根据对话整体质量回传一个标量奖励。这样学生能知道**具体哪一步出了错**，从而更有效地提升对话质量和事实准确性。实践中，Anthropic、OpenAI 等公司据传也采用了类似思想：先用人类反馈训练出强大的对话模型，再通过**蒸馏这个RLHF教师**来训练小模型，用于移动端或其他资源受限场景。虽然细节未公开，但理念与On-Policy Distillation一致，即**用强模型在线指导弱模型学习**，兼顾效果与成本。

- **持续学习与个性化定制**：On-Policy Distillation 还展现出在**模型持续学习**和**个性化细调**方面的独特优势。一般的小模型若直接在新领域数据上微调，容易发生**灾难性遗忘**，丢失原有能力。然而研究表明，可以用模型早期checkpoint作为教师，通过on-policy distillation把**已学会的技能重新注入学生**。比如前述 Qwen3 实验中，小模型在加入新领域知识后，其原先的指令遵循能力出现退化，于是研究者用之前版本的模型当教师，开启一个on-policy蒸馏阶段，结果成功**恢复了模型的问答和助理能力**，同时保留了新知识 [24] [25]。这种**交替微调-蒸馏**的流水线让模型能够不断吸收新知识又不忘旧本领，非常适合企业内部“长期学习”的需求。类似地，对于特定企业或用户定制的小模型，也可用通用大模型作为教师，在目标领域数据上on-policy蒸馏，从而**个性化地训练模型**。与单纯用目标数据微调相比，加入教师反馈能确保模型**既适应新域又维持总体能力**。因此，我们看到越来越多开源项目开始支持这类流程，例如 HuggingFace 社区的TRL库、清华的ChatGLM等，都探讨在指令微调中融入交互式蒸馏以提升效果。

总的来说，无论在**强化学习**还是**模仿学习/大模型微调**中，On-Policy Distillation正推动着**训练范式的革新**：以前要么依赖离线专家数据、要么耗费大量在线试错，现在可以借助强教师“在线指路”快速学得高性能策略。这种方法的应用场景从游戏AI、机器人控制扩展到了对话系统、代码助手等诸多领域，显示出广阔的前景。

技术细节：On-Policy vs Off-Policy Distillation 与知识蒸馏的关系

为了更清晰地理解On-Policy Distillation的技术特点，我们将其与传统的Off-Policy蒸馏和一般知识蒸馏进行对比：

- **知识蒸馏 (KD)**：广义而言，包括一切教师指导学生学习的过程，通常指经典的离线蒸馏。教师模型输出概率分布或 logits，学生以此为软标签最小化交叉熵或KL散度损失 [1]。KD 一般在**静态数据集**上进行，**不涉及环境交互**。优点是训练稳定、实现简单，在分类、NLP等任务中被大量应用，用于模型压缩和知识迁移。缺点是在自回归生成场景下会遇到**训练/推断分布不匹配**的问题——学生在训练时永远跟随教师输出的轨迹，可能造成“学会了考试答案却不会独立解题”。

- **Off-Policy Distillation（离策略蒸馏）**：指学生模型学习**教师预先生成的轨迹**。这其实是知识蒸馏在序列建模（特别是RL或LLM微调）中的典型用法，也叫**监督微调 (SFT)**。例如用强模型生成问答对，让学生模仿答案；或让专家智能体走迷宫，学生模仿专家动作。Off-policy蒸馏提供了**密集的逐步监督**（每一步都有教师信号），训练过程等价于**模仿专家**。其突出优点是**效率高**：教师信号丰富明确，学生收敛速度快 [23]。特别在LLM场景，一次性生成海量高质量指令数据，然后SFT学生，比起强化学习那种逐轮交互，成本低廉。然而缺点在于**“教师分布陷阱”**：学生只在教师见过的情境下表现良好，如果学生一旦偏离教师常规（例如一开始就输出了教师从未犯的错误），接下来就进入了**教师未覆盖的状态空间**，学生无从借鉴正确做法，误差将**步步放大** [19]。这就是所谓**复合误差 (compounding error)**问题。在短文本生成还好，在长对话或长规划中，离线蒸馏往往后劲不足，学生容易“自我崩溃”。另一个问题是**表面模仿**：学生可能学会了教师的语气和格式，却**没有真正理解任务**。Gudibande 等人（EMNLP 2023）发现，用GPT-4生成数据来训练学生模型时，学生模型在事实问答上倾向于给出听起来很自信但实际错误的回答，说明它只是模仿了GPT-4的风格而未掌握GPT-4背后的知识或推理过程 [26] [23]。这些都是Off-policy蒸馏固有的局限。

- **On-Policy Distillation（在线蒸馏）**：学生模型**以自身当前策略**与环境交互，产生轨迹，然后参考教师模型对这些**学生轨迹**的反应来更新自己。这里教师的作用可以是**批评者**（critic）或**专家**（expert）：常见做法是让教师模型对学生每一步动作打分，或直接提供该状态下教师理想的下一步输出分布，学生据此调整参数。与off-policy相比，on-policy蒸馏最大的特点是**训练数据来自学生自己的决策分布**，从而**消除了分布失配**的问题 [5]。学生不断在自己容易出错的地方受教，逐渐学会**纠错和自我恢复**。此外，由于有教师的细粒度指导，相比纯RL，on-policy蒸馏的反馈信号**更加密集丰富**，不再只有输赢这一个稀疏信号 [4]。这使得训练**高效且稳定**：例如在复杂数学推理任务上，用on-policy distillation只需很少的训练step就能取得与RL长期探索相当的成绩 [27]。当然，on-policy方法也有挑战：首先需要一个**强大的教师**时时待命，否则学生的trajectory无法得到有价值反馈。在现实中，教师通常是一个**性能远超学生**的大模型或一套昂贵的评估器（如带有知识检索或搜索模块的系统）。因此on-policy蒸馏经常搭

配“先离线蒸馏+再在线蒸馏”的两阶段：先用现有数据离线训练学生到一定水准，然后进入on-policy环节进一步提升。另一个技术细节是**损失函数的设计**：简单使用逐token交叉熵有时不足以指导学生逼近教师策略。文献提出了多种替代，如**KL散度**的不同形式。特别地，经验表明在on-policy蒸馏中采用**反向KL散度**（即以教师分布为基准的KL）效果突出[28]。反向KL是**“模式寻求”**(mode-seeking)的，会促使学生尽量覆盖教师策略中的高概率行为而不去探索教师未考虑的怪异输出，从而**避免了学生通过增加随机性来欺骗损失下降的“投机取巧”**（unhackable）[28]。例如在Qwen3微调中就使用了教师对学生分布的反KL，使学生在每个状态都朝教师行为靠拢[27]。此外，一些工作还引入**价值函数蒸馏**或**奖励蒸馏**等信号，融合策略和价值的学习，提高学生对长远回报的把握（如前述 NeurIPS 2024 方法通过匹配Q值分布来传递策略优劣[29][30]）。总而言之，On-Policy Distillation综合了**RL的相关性**与**蒸馏的高信息量**，在技术实现上需要考虑教师来源、采样策略、损失设计等细节，以实现性能与稳定性的最优平衡。

小结：知识蒸馏是总范畴，off-policy蒸馏是其传统形式，强调**模仿教师**；on-policy蒸馏是新形式，强调**在实践中向教师请教**。Off-policy方法数据利用率高，但学生可能局限于教师影子；on-policy方法学生学得针对性强，但需处理好训练稳定性和开销问题。实际应用中，两者常配合使用：**先离线学习打基础，再在线蒸馏补短板**。近期的研究甚至主张**联合优化**这两种目标，取得比单独更佳的效果[13][14]。这表明Off-policy和On-policy并非对立，而是可互补增益，为知识蒸馏开拓出更广阔的技术空间。

代表性算法与模型综述

为了便于读者纵览，我们以列表形式汇总若干与On-Policy Distillation密切相关的算法或模型，以及它们的特点：

- **Policy Distillation (2015, DeepMind)**[2]：首个将知识蒸馏用于强化学习策略的工作。通过离线蒸馏压缩DQN策略，实现**小模型专家级控制**；还能将多个游戏策略融入一个网络，实现**多任务统一**[2]。此算法验证了蒸馏在RL中的可行性，为后续工作铺平道路。
- **Actor-Mimic (2016, CMU)**[3]：多任务和迁移强化学习算法。由多个教师（各自擅长不同游戏）指导学生网络，同时学会多任务策略，并在没有教师的新任务上**快速适应**[3]。该模型引入**策略模仿+表示学习**思想，让学生通过蒸馏获得通用状态表示。
- **DAgger (2011, MIT)**：模仿学习算法，不直接称为蒸馏但理念相通。通过**交互采样+专家纠偏**迭代构造训练集，让学生策略逐步逼近专家策略。其核心思想（学生走自己的路，专家随行指导）后来在On-Policy Distillation中以“学生on-policy采样+教师反馈”形式重现，被认为是保证学生策略稳定性的关键。
- **Kickstarting Deep RL (2018, DeepMind)**[9][10]：“策略蒸馏+在线RL”结合算法。在学生的RL优化目标中加入一项蒸馏损失，鼓励学生策略贴近预训练教师。这**加速了学生学习**，在大型3D导航任务上学生以**1/10数据量**匹敌从零训练，并可超越教师[11][10]。Kickstarting 证明了**教师指导下的RL**能够同时提高**样本效率**和最终性能。
- **AlphaGo Zero 自博弈策略蒸馏 (2017, DeepMind)**：虽然不是以“蒸馏”命名，但本质是**迭代蒸馏**过程。利用当前策略网络配合MCTS产生更优策略分布，然后训练更新网络匹配之。该循环使策略持续改进，省去人工专家数据，堪称**极致形式的on-policy策略蒸馏**。AlphaGo Zero 的成功将这一思想推向顶峰，也带来了泛化的**Expert Iteration**框架，被用于其他博弈 AI。
- **RLHF + Distillation in LLMs (2020+, OpenAI等)**：在大语言模型对齐中使用“**先RL微调，后策略蒸馏**”的两段式训练。例如在人类反馈总结任务中，OpenAI先用RLHF训练一个大模型，然后将其策略**蒸馏**给一个小模型，以便部署【未公开细节】。这一思路后来广泛应用于开源社区：如有团队用人类偏好训练出奖励模型，再用该奖励模型指导学生对话模型进行on-policy优化（称为RLAIF等）。这些实践表明**RL信号和蒸馏目标可以协同**：RL探索提供方向，蒸馏确保高效学习和小模型可用性。
- **Generalized Knowledge Distillation (2023, Google)**[1][6]：提出**学生自举采样+教师密集纠偏**框架，并通过实验确立了其在各类生成任务上的有效性。作为On-Policy Distillation的代表算法，它提供了灵活的损失定制（如选择KL方向、融合MSE等）来适配不同学生模型容量，并证明**可与RLHF无缝结合**提升对齐效果[6]。

- **DistiLLM 系列 (2024, KAIST 等)** [12]：一系列关注**蒸馏过程优化**的研究。DistiLLM 引入**偏置KL损失**，强调让学生更多关注复制教师高概率输出，同时设计**经验回放策略**平衡on/off数据利用 [12]。后续的 DistiLLM-2 则加入**对比学习**思想，进一步提升蒸馏效果 [31]。这些算法改进细节提升了蒸馏的效果，展示了在目标函数和训练动态方面的创新。
- **Adversarial Moment-Matching (2024, CUHK)** [13] [14]：通过构造一个学生策略与教师策略价值差距的**对抗游戏**，达到蒸馏知识的目的。尤其突出的是他们将**on-policy与off-policy目标结合**起来优化，结果比单纯蒸馏教师数据或单纯on-policy各自都要好 [13] [14]。这代表了最新的趋势：不再局限于单一范式，而是融合多种训练信号以求最优。
- **Qwen3 + On-Policy Distillation 实践 (2025, Alibaba)** [7] [8]：在通用大模型上验证了on-policy蒸馏的**计算效率优势**和**性能增益**。通过少量GPU资源，就达到了甚至超过需要大规模算力RL微调的效果 [8]。这证明了该技术的实用价值，已经进入工业大型模型训练流程。Qwen3 的训练还表明，蒸馏不仅没有限制学生探索，反而提升了学生在复杂数学推理等方面的**解题空间** [7]。作为开源的前沿模型，Qwen3 的成功有望带动更多企业采用类似方案。

(以上列表涵盖了不同年份和机构的一系列工作，从概念提出、方法改进到实际应用验证，体现了On-Policy Distillation方法论的发展脉络。可以看到，学术界顶会论文提供了理论和算法创新，工业界顶尖实验室则输出了大规模验证和开源实践，共同推进了该领域的进步。)

最新研究进展（2023年至今）

进入 2023 年以来，On-Policy Distillation 相关研究呈现**加速发展**和**方向拓展**的态势。几大值得关注的趋势包括：

- **大模型对齐中的蒸馏革命**：随着大模型（尤其聊天模型）的流行，如何在保证性能的情况下降低推理成本、满足部署约束成为焦点问题。2023-2024 年，涌现出多篇探索**RLHF结果蒸馏**的研究，试图用蒸馏技术取代昂贵的强化学习环节。例如前述 Google 的 GKD 方法 [1] 证明，小模型可以通过on-policy蒸馏，同时得到**来自教师模型和人类偏好的双重指导**，从而在遵循人类指令上达到媲美RLHF大模型的效果。OpenAI等也在内部尝试类似思路，用已RLHF调优的大模型（如ChatGPT）生成交互数据，来监督训练较小模型。这一方向在实践中极具价值，因为它意味着**更小、更廉价的模型也能具备接近RLHF模型的对齐水准**，为开源社区复现强对话模型提供了可能。近期有研究进一步指出，将**人类反馈信号融入蒸馏损失**可以显著提升学生模型的礼貌性、安全性等指标，从而实现“无RLHF的RLHF效果”。可以预见，未来蒸馏技术将在模型对齐与伦理方面扮演更大角色。

- **混合蒸馏策略**：正如上节提到的新方法，学者们开始探索**同时结合on-policy和off-policy训练目标**的方法，以利用双方优点 [13] [14]。2024 年NeurIPS的成果表明，**混合优化**比纯粹的单一路径更优 [13] [14]。因此我们可能会看到一种标准范式的形成：训练过程既包括教师静态数据的模仿阶段，也包括学生在线试探的纠偏阶段，两者交替或联合进行。例如，可以设想未来的大模型微调管线：首先用教师生成的大量高质量数据进行几轮off-policy蒸馏，使学生快速接近教师性能；随后让学生与一小部分真实用户或仿真环境交互，再用教师对这些真实交互进行逐步评审（on-policy distill）来加强学生在实际分布下的鲁棒性；如此循环数次。混合策略有望进一步**提升学生模型的可靠性**，减少“一遇新情况就崩”的现象。目前一些开源项目（如HuggingFaceH4 On-Policy Distillation demo等）已经在尝试易用的混合蒸馏工具，方便开发者将该思想应用于任意模型族 [32]。可以预计，**标准化的混合蒸馏框架**将很快出现，为大规模模型的高效微调提供开箱即用的支持。

- **多模态与新领域扩展**：2023 年以来，蒸馏技术也在从语言领域扩展到**多模态**和**决策智能体**等新方向。例如有工作尝试将大型语言模型的**推理能力蒸馏给视觉语言模型**，提升后者在需要语言推理的问题上的表现 [33]。又如Meta的研究者开始讨论**如何将策略蒸馏用于机器人策略学习**，把模拟环境中学到的大型策略模型压缩到能在真实机器人上实时运行的小模型上。随着生成AI向多模态发展，On-Policy Distillation在这些场景同样具有意义：比如未来可能用一套强大的多模态基础模型指导较小的边缘模型

学会图像描述、视频理解等任务。初步的成果（如VOLD 2024工作，将LLM的逻辑推理蒸馏给视觉问答模型 [33]）已经展现了可喜的前景。可以预见，**跨模态蒸馏、跨领域蒸馏**会成为下一步热门课题。

- **开源工具和社区实践**：在最新进展中，不得不提及开源社区的推动作用。2023 年，大模型蒸馏迎来了一波开源热潮：Stanford Alpaca 引领了指令蒸馏数据发布风潮，随后Vicuna等开源模型均采用了蒸馏技巧；HuggingFace 等搭建了专门的库（如TRL）来支持RL与蒸馏训练。一些知名实验室也公开了他们的蒸馏代码和模型，例如阿里的 Qwen 系列、清华的ChatGLM2 等，都提供了蒸馏阶段的技术细节。社区还出现了专门探讨LLM蒸馏的论坛和综述文章 [34] [35]。可以说，**开源项目**正帮助将学术前沿的方法转换为易于使用的实践方案，加速了On-Policy Distillation的传播。例如，近期一个开源项目DeepSeek-AI展示了使用纯RL（on-policy强化学习）提升模型推理能力的新思路，与蒸馏方法形成竞合关系 [36]；而另一边，蒸馏派系则通过不断优化蒸馏目标（如DistiLLM系列）来缩小与RL训练效果的差距。由于社区活跃，我们看到学术论文的idea能被很快实现验证，这将持续推动技术演进。

- **性能与效率的权衡**：最新的研究还关注如何让蒸馏**又快又好**。例如2025年的一些技术报告提出，用**较低精度**的大模型当教师配合蒸馏，可以进一步降低计算成本而几乎不影响学生性能；或者在蒸馏过程中结合**剪枝/量化**技术，让学生模型高效的同时更小更快。一些工作开始报告**蒸馏的极限**：某些复杂知识小模型无论如何也学不全（“蒸馏瓶颈”），因此需要引入新的教师信号（如链式思维、中间步骤监督）才能弥补。总之，如何充分挖掘教师知识、突破学生容量限制仍是研究重点。特别是在LLM领域，2023年后出现了**100B+参数教师 vs 7B学生**的不对称局面，很多蒸馏策略在这种极端师生差距下需要重新审视。最近有论文提出分层蒸馏、循序蒸馏（先大降中，中降小）等方案来提高学习效率 [37] [38]。可以预见，围绕**蒸馏效率**的创新还会不断涌现，以满足实际应用中对**低成本高准确**模型的需求。

综上所述，过去两年On-Policy Distillation领域的发展可谓日新月异：从理论改进（广义蒸馏、混合目标、对抗训练）到应用拓展（LLM对齐、多模态蒸馏）再到工具开源，大量证据表明这种融合了强化学习与监督学习优点的方法**潜力巨大**。特别是在大模型时代，On-Policy Distillation为我们提供了一条**低成本获取高性能模型**的现实途径。展望未来，我们期待看到该方法在更多场景下落地，例如更复杂的连续控制、个性化医疗AI，以及通用多智能体系统等。同时，也需关注其中尚待解决的挑战，包括如何选择最佳的教师反馈信号、如何在保障稳定性的同时提升学生探索、以及如何理论上证明蒸馏过程的收敛与正确性。这些问题的解决将进一步巩固On-Policy Distillation在机器学习版图中的地位。可以肯定的是，随着研究的深入，**On-Policy Distillation将继续在实践中证明自己，成为连接智能体训练范式的重要桥梁** [5] [7]。

参考文献：本文内容引用和参考了包括ICLR、NeurIPS、ICML等顶会论文及知名研究机构开源报告在内的资料 [2] [3] [9] [1] [7] 等，完整出处请参见引用标注。

---

[1] [6] On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes | OpenReview
https://openreview.net/forum?id=3zKtaqxLhW

[2] [1511.06295] Policy Distillation
https://arxiv.org/abs/1511.06295

[3] [1511.06342] Actor-Mimic: Deep Multitask and Transfer Reinforcement Learning
https://arxiv.org/abs/1511.06342

[4] [5] [23] [27] [28] On-Policy Distillation LLMs Redefine Post-Training Efficiency
https://www.startuphub.ai/ai-news/ai-research/2025/on-policy-distillation-llms-redefine-post-training-efficiency/

[7] [8] [18] arxiv.org
https://arxiv.org/pdf/2505.09388

[9] [10] [11] [1803.03835] Kickstarting Deep Reinforcement Learning
https://arxiv.org/abs/1803.03835

12 [PDF] DistiLLM: Towards Streamlined Distillation for Large Language ...
https://arxiv.org/pdf/2402.03898

13 14 15 16 17 29 30 proceedings.neurips.cc
https://proceedings.neurips.cc/paper_files/paper/2024/file/cbae8efcc23a0cb6d15a20f245514020-Paper-Conference.pdf

19 20 21 22 24 25 26 On-Policy Distillation - Thinking Machines Lab
https://thinkingmachines.ai/blog/on-policy-distillation/

31 [PDF] DistiLLM-2: A Contrastive Approach Boosts the Distillation of LLMs
https://openreview.net/pdf?id=rc65N9xIrY

32 Unlocking On-Policy Distillation for Any Model Family - Hugging Face
https://huggingface.co/spaces/HuggingFaceH4/on-policy-distillation

33 VOLD: Reasoning Transfer from LLMs to Vision-Language Models ...
https://arxiv.org/html/2510.23497v1

34 LLM Model Distillation: A Research Survey - Medium
https://medium.com/@abhi-84/llm-model-distillation-a-research-survey-3c2a2eeb61a7

35 DistiLLM-2: A Contrastive Approach Boosts the Distillation of LLMs
https://arxiv.org/html/2503.07067v1

36 DeepSeek Releases R1 and Opens up AI Reasoning
https://patmcguinness.substack.com/p/deepseek-releases-r1-and-opens-up

37 [PDF] Aligning Feature Dynamics for Effective Knowledge Distillation
https://aclanthology.org/2025.acl-long.1125.pdf

38 (PDF) Adapt-and-Distill: Developing Small, Fast and Effective ...
https://www.researchgate.net/publication/353491000_Adapt-and-Distill_Developing_Small_Fast_and_Effective_Pretrained_Language_Models_for_Domains