OpenAI o1的具体训练技术并未进行公开，但通过综合多种信息源可以推测的是OpenAI o1 运用的技术关键在于强化学习的搜索与学习机制。而对于 OpenAI o1 的技术路径，OpenAI o1 应当也是沿着 STaR 和 Quiet-STaR 类似的路线，即优化模型内部生成合理推理（即隐式的 CoT) 的过程。首先模型在生成内部思维链推理的过程中，o1模型训练过程与AlphaGo训练过程类似，o1模型通过蒙特卡洛树搜索树增强推理的合理性。蒙特卡洛树搜索（Monte Carlo Tree Search, MCTS）是一种用于决策过程中的搜索算法，特别适用于具有大状态空间的游戏或问题。MCTS结合了随机抽样与树状结构的探索，以在有限的时间内找到最优解或接近最优解。通过MCTS的模拟过程，可以生成一系列的推理步骤。每一步都代表了一个可能的动作，而MCTS会根据这些动作的预期结果来选择最有潜力的路径。这个过程允许MCTS将实例级别的反馈转化为更细粒度的步骤级别信号，从而提供更加具体的指导信息。和标准的 MCTS 用法有所不同，MCTS 节点上不再是最终生成答案中的某个 token 或某步，而是隐式推理过程中的每一步。 换句话说，MCTS的节点不再简单地代表某个具体的输出，而是代表模型在推理过程中的某个中间思维步骤。这意味着MCTS能够探索多个推理路径，而不仅仅是探索生成的某个单一答案，同时每个节点对应的是推理链中的一个决策点，而不是输出的一个token。通过在每个推理步骤上进行搜索和优化，模型能够更好地形成一个完整且连贯的推理链条，从而提高解决复杂问题的能力。总结来说，结合了MCTS的前瞻搜索能力和逐步自我评估机制，能够在复杂的推理任务中生成高质量的偏好数据，进而帮助提升LLM的推理性能。通过这种方式，模型不仅能够获得关于最终输出的反馈，还能得到有关推理过程中每个步骤质量的信息，从而促进更细致和准确的学习。

同时，为了提供更加细粒度的反馈和指导，需要引入过程性的奖励。这其中就涉及到过程监督的方法，过程监督是一种为模型推理中的每一步提供反馈的训练方法。相比于仅对最终结果进行评价的结果监督，过程监督能够在模型推理的每个中间步骤上提供精细的反馈，从而帮助模型更加准确地识别和修正错误。过程监督的核心思想是为模型的每一步推理提供监督，而不仅仅是根据最终输出进行评估。这种方法通过逐步检查模型生成的中间步骤，发现错误并及时纠正。在o1训练过程中具体步骤可能是首先模型在解决问题时，通过使用MCTS算法会有一个生成思维链的推理过程，每个内部推理步骤都会被视作一个独立的推理单元。然后对于每一个步骤，Critic模型都会对其进行判断,提供精准的反馈。如果某一步系统反馈较差，系统会立即终止后续步骤的生成，确保错误不继续传播到接下来的推理中。这种机制有助于模型识别错误发生的确切位置，方便及时纠正。相比之下，结果监督只对整个推理过程的最终答案给出“正确”或“错误”的反馈，不能告知模型错误出现在哪个具体步骤，导致难以精准改进。同时通过将推理过程进行分解，并且利用额外的更强更专项的 Critic Model，可以将推理过程的监督扩展到更复杂的问题上。通过给模型提供推理过程中每一步的奖励信号（而不是仅仅在最终结果正确时提供），解决了只依赖最终正确答案时奖励信号稀缺的问题，使得模型在训练中能够获得更频繁的反馈，从而更有效地学习如何优化推理过程。关于引入Critic Model生成反馈，之前 OpenAI 发布的 CriticGPT ， 通过 RLHF 方法训练模型能够为真实世界中的代码任务书写自然语言反馈，并成功泛化到 OOD 的分布上。这种反馈可以用来帮助人类进行更准确的评价，从而实现对于复杂输出的有效奖励反馈。先前 OpenAI 也深入验证了通过自我批判方法和 Critic Model 辅助人类评判在文本总结任务上的可行性。从可扩展监督的角度来说，这条路是必然的。对于 Critic 这类方法，关键的挑战在于如何将 Critic 的能力泛化到评判模型内部思维链推理步骤的任务上，，Critic Model 需要考虑输出各部分之间的依赖关系和逻辑推理问题，因此对于模型自身的推理能力要求更高。Critic Model同样也可能会存在 Generator-Discriminator-Critique (GDC) gaps ，即模型可能不会指出他们所发现的错误。基于评价比生成更简单的原则，Critic 的思路是可扩展监督一条有前景的技术路径，并且应该也被用在了辅助 o1 的训练当中。

关于o1训练过程中的迭代优化，o1可能使用了一种数据飞轮 + Bootstrap 的方式。数据飞轮（Data Flywheel）是指通过模型不断生成高质量数据来反过来提升模型性能，从而形成一个自我强化的良性循环。o1模型在推理过程中会通过反馈机制自动优化推理链，并通过强化学习引导模型进行更合理的推理。这些优化过程大大加速了模型性能的提升，形成了数据飞轮机制所需的自我强化回路。

o1模型通过自举（Bootstrap）机制来提升其推理能力。Bootstrap指的是通过现有的推理能力生成新的合理推理过程，这些推理过程进一步用于优化模型。这个过程非常符合数据飞轮的特性。具体过程为首先生成推理数据，模型在解决问题时不仅生成答案，还生成与之对应的推理过程（Rationales）。这些合理的推理过程被模型收集并加入到训练数据集中。然后通过反馈与训练，生成的新推理过程作为新的高质量数据反过来用于训练和微调模型，这个循环过程不断强化模型的推理能力。通过每次迭代，模型产生的新推理数据会进一步提升模型性能，最终形成了一个自我增强的反馈循环。在这个自举过程中，模型生成的数据不仅用于解决当前问题，还被用于训练下一轮模型的推理能力。这一过程与数据飞轮的理念高度一致，即通过模型自身产生的高质量数据形成正向反馈，推动模型性能持续提升。

最终通过强化学习的方法，o1 学会了优化其内部思维链，并不断改进其使用的策略。它学会识别并纠正错误，学会将复杂的步骤分解为更简单的步骤，并在当前方法无效时尝试不同的解决方案。这个过程大幅提高了模型的推理能力。

由o1-system-card可得知，在o1模型训练完成之后，openAI团队还会对o1模型进行安全性方面的训练与对齐，即采取mitigations（缓解措施） ，对模型在部署前采取的一系列改进和安全性优化，以减少潜在风险并确保模型的安全性和稳健性。这些缓解措施旨在防止模型生成不当内容、抵御越狱（jailbreak）攻击、减少幻觉、以及避免偏见等问题。