DeepSeek-R1 VS EDARE 特点

原文部分（节选）

“For DeepSeek- R1, we extend this approach by incorporating both rule- based rewards for reasoning- oriented data and model- based rewards for general data, thereby enhancing the adaptability of the learning process across diverse domains.” Nature+1

“DeepSeek- R1- Zero demonstrates capabilities such as self- verification, reflection, and generating majority voting, highlights its strong foundational capabilities and its potential for further advancements in reasoning tasks.” arXiv

中译

对于 DeepSeek- R1，我们将已有的方法拓展为：对以推理或逻辑为导向的数据使用**规则型奖励**（rule- based rewards），对更广泛的一般性数据使用**基于模型的奖励**（model- based rewards），从而增强学习过程在多样领域上的适应性。

DeepSeek- R1- Zero 显现出自我验证（self- verification）、反思（reflection）以及生成多数投票机制（majority voting）的能力，突显其在推理任务上具有坚实的基础能力和进一步发展的潜力。 arXiv+1

方法重点

下面是 DeepSeek-R1 在训练流程、奖励设计、CoT（思维链）涌现方面的方法要点，中译并附带简要说明：

项目	内容	意义 / 我们可以学到的
训练流程（“Training Pipeline”）	*DeepSeek- R1- Zero* 从基础模型（base model）直接用强化学习(RL)开始，不通过 SFT（supervised fine- tuning）作为初始阶段。然后在后期版本中加入冷启动数据（cold start data）来提高可读性与格式（format）一致性。 arXiv+236 氪+2	这与我们 EDARE 中“尽早检测顿悟 + 用熵轨迹与注意力跳跃”非常一致。我们可以考虑是否在我们系统中也提供一个不依赖很多监督数据的路径（类似 “Zero RL” 流程）。

项目	内容	意义 / 我们可以学到的
奖励设计	- **规则奖励（rule- based rewards）**：对推理密集型任务（数学、逻辑、编码）使用规则奖励，比如 “答案正确性”，“格式正确（包含思考步骤、标注、代码通过测试）”。36 氪+1 - **模型奖励（model- based rewards）**：对非推理任务或通用任务，引入模型评价，比如人类偏好 / 模型评价者判断，以提升通用性。Nature+1 - 在某些版本中，还设计“格式一致性奖励”（language consistency reward），奖励中英混杂问题的改善。 36 氪+1	我们可以在反思触发中加入“格式一致性 /可读性”这个维度作为辅助判断；这可以用作我们“reason”中要解释的一部分。
CoT（思维链 / chain- of- thought）的自然涌现	在 DeepSeek- R1- Zero 的训练过程中，模型未被强迫写长的中间推理步骤，但随着 RL 的推进，它自然地在回答逻辑/数学问题时生成越来越长的 CoT；这些中间版本中会慢慢“展开步骤”、“验证过程”、“思考时间变长”。 arXiv+236 氪+2	这对应我们希望在 EDARE 中用 entropy/attention 跳跃来捕捉 “顿悟节点”与“新思维路径”的出现。我们可以试图量化 CoT 长度 + 新关键词 + 熵下降的关系。
成本与资源	DeepSeek 在公开资料中声称训练 R1 的成本相对较低（如 $294,000）在其 Supplementary Material 中披露。其用的 GPU 芯片包括 H800（针对中国市场）、一些早期阶段可能使用 A100（受限出口前）等资源。 Reuters+2Nature+2	对我们意味着：即使资源不是极为豪华，也能通过合理设计奖励 +训练流程 +蒸馏/压缩版本获得强推理能力。我们要继续注重资源效率。

对比图表：DeepSeek- R1 vs 我们 EDARE 框架

以下是我为你做的对比表格，展示 DeepSeek- R1 方法中关键要素 vs EDARE 当前设计，对我们意味着什么：

比较项	DeepSeek- R1 方法特点	我们 EDARE 当前设计	差距 / 可借鉴之处
初始化方式	DeepSeek- R1- Zero 从 base model 用 RL 直接训练，不依赖 SFT 启动；后期用冷启动数据改善格式与可读性。	我们目前也设定 M2/M3 反思与 trigger 流程，不严格依赖大量监督数据。	可以考虑一个“Zero RL”风格的路径，以测试我们在弱监督条件下触发能力的可靠性。
奖励设计维度	准确性奖励 + 格式奖励 + 模型奖励 +可读性 /语言一致性（在后期版本）	我们目前有熵下降、attention 跳跃、confidence + 语义 /可读性（格式）作为辅助。	可以进一步显性加入格式奖励（如 CoT 格式一致性、正确标记思考步骤等）与语言一致性。
思维链 CoT 涌现	模型训练过程中自然延长 CoT 步骤、验证、反思、自我校正 → 在推理性能中显著提升。	我们设计的 attention 跳跃 + 熵轨迹正是想捕捉这种自然展开与顿悟；尚需监测 CoT 长度、新关键词出现等。	可以加入 CoT 长度记录 + 新关键词功能，以量化“思考路径扩展”的阶段。
资源与效能	使用合理 GPU（如 H800 等），训练成本公开；在运行推理与蒸馏阶段推出轻量版本。	我们在设计中已经考虑到计算成本（减少迭代次数、加权初始值、简化库依赖等）与可视化/日志模块轻量化。	保持这个设计方向，同时尽量公开我们性能 vs 成本对比，以增强可说服力。