



链式思考（CoT）在奖励函数设计中的价值

链式思考提示要求模型生成一步步推理后给出答案，这通常能提高复杂任务的准确度。在奖励建模领域，有研究发现，将奖励建模视为推理任务可以显著增强模型的解释性和性能¹。例如，RM-R1 提出 **链式推理训练**后，奖励模型不仅给出最终分数，还生成了连贯的评价理由，从而提高了判断的一致性和可解释性¹。**支持观点：** CoT 提示使输出逻辑透明，可作为结构化奖励设计的重要工具，帮助工程师理解奖励来源¹。**反对观点：** 但也有人指出，传统的 CoT 提示侧重于线性推理，可能限制搜索空间宽度，如 PCGRLLM 等工作认为 CoT 方式会限制奖励函数探索，需要引入更广泛的提示方法²。综上，目前虽然 CoT 被视为一种增强可解释性的有力手段，但并未被公认为奖励设计的**唯一或核心机制**。

LLM驱动的奖励工程：CoT与替代方法

- **采用 CoT 的系统**：一些早期工作直接使用 CoT 引导奖励函数生成。例如，ChatPCG 框架在奖励生成流程中使用 CoT 自对齐（self-alignment）技术，以确保生成的奖励函数与题意相符³。该方法让 LLM 在生成奖励代码前“思考”如何评估游戏状态，从而提高了奖励的合理性³。
- **无 CoT 的替代方案**：也有许多系统并不依赖显式的推理链，而是直接让 LLM 生成完整的奖励函数代码并通过迭代反馈优化。例如，Eureka 算法利用 GPT-4 等大模型直接**生成奖励函数的代码**，并通过演化算法优化，未使用逐步推理提示⁴。同样，CARD 框架中的 Coder 模块仅给 LLM 提供环境描述和目标，然后让其输出奖励函数代码⁵；生成后通过自动执行和 Evaluator 反馈机制不断改进，而非显式要求模型“思考”。这些例子表明，**CoT 并非唯一基础**，现代系统更倾向于结合代码生成、自动反馈和搜索策略来设计奖励函数^{4 5}。

奖励生成系统中的动态温度与模型选择机制

- **动态温度调节（DTRO）**：在 LLM 推理中已经探索过根据置信度调节采样温度的方法，以平衡多样性和准确性。例如，Shin 等提出的 EGoT（Enhancing Graph of Thought）方法采用余弦退火随推理层数逐渐降低温度：前期保持高温产生多样输出，后期降低温度获得精确答案⁶。另有研究表明，链式思考提示通常提高模型输出的置信度⁷，与之配合的动态温度机制（如基于答案间置信度差异动态调节温度⁸）可以进一步在探索性与收敛速度间进行自适应平衡。虽然现有文献中未见专门以“DTRO”命名的模块，但这些启发式温度调节思路实际上就是提升奖励生成稳定性的上层策略。
- **动态模型选择（DMSRO）**：类似地，多模型选择机制被应用于提升生成系统的表现。在 LLM 集成领域，**DER** 方法把不同 LLM 专家视为马尔可夫决策过程，根据输入动态选择问答路径，以最小化资源消耗并最大化效果⁹。在奖励模型训练中，**LASeR** 方法则动态挑选最有鉴别力的规则或奖励模型（相当于多臂赌博机）来标注样本¹⁰。此外，RuleAdapter 等工作会根据评价差异为每个样本选择最相关的安全评判规则¹¹，也是动态选择思想的体现。虽然目前尚未有文献直接称之为 DMSRO，这些策略本质上都属于对奖励生成/评估管道的上层优化，与 CoT 互补：CoT 提供结构化推理增加输出置信⁷，而动态选模机制则根据任务需求在不同模型间切换，实现更高鲁棒性^{9 10}。

研究趋势与未来方向

综上近年来研究表明，使用大模型自动生成奖励的思路正在快速发展：Eureka、CARD 等工作展示了**无需人工编码**即可通过 LLM 生成并迭代改进奖励函数^{4 5}；RM-R1 等研究强调**引入推理链**能提高奖励评判的透明度和性能

①；而 EGoT/Graph-of-Thought 等框架则将**动态推理控制**（如温度调度）引入多阶段 reasoning 流程 ⑥。这些进展表明，未来奖励工程可能融合**更丰富的提示结构与自动优化机制**：既利用链式推理或自我发现的推理结构增强可解释性，也通过动态参数和模型集成策略提升系统鲁棒性。总体来看，研究趋势朝着建立**自动化、可解释**的奖励设计流水线演进，即用更强的LLM能力和反馈循环减少人力干预，同时保持设计的透明度和可控性 ① ⑥。

参考文献：本文观点主要基于近期相关文献：③ ① ⑥ ④ ⑤ ② ⑨ ⑩ ⑪ ⑦ 等。

① RM-R1: Reward Modeling as Reasoning

<https://arxiv.org/html/2505.02387v1>

② ③ arxiv.org

<https://arxiv.org/pdf/2502.10906>

④ Eureka: Human-Level Reward Design via Coding Large Language Models | OpenReview

<https://openreview.net/forum?id=IEduRUO55F>

⑤ A Large Language Model-Driven Reward Design Framework via Dynamic Feedback for Reinforcement Learning

<https://arxiv.org/html/2410.14660v1>

⑥ Enhancing Graph Of Thought: Enhancing Prompts with LLM Rationales and Dynamic Temperature Control | OpenReview

<https://openreview.net/forum?id=l32IrJtpOP>

⑦ ⑧ Revisiting Self-Consistency from Dynamic Distributional Alignment Perspective on Answer Aggregation

<https://arxiv.org/html/2502.19830v1>

⑨ Dynamic Ensemble Reasoning for LLM Experts

<https://arxiv.org/html/2412.07448v1>

⑩ LASeR: Learning to Adaptively Select Reward Models with Multi-Armed Bandits | OpenReview

<https://openreview.net/forum?id=fDcn3S8oAt>

⑪ RuleAdapter: Dynamic Rules for training Safety Reward Models in RLHF | OpenReview

<https://openreview.net/forum?id=tcMWVgYf8f>