CO-GRAFT：通过角色级策略嫁接归因搜索智能体的强化学习增益

Ramraj Chandradevan, Sayontan Ghosh, Vinoth Selvendran{first.last}@gmail.com

arXiv:2611.31185v1 [cs.IR] 23 Nov 2026

摘要

强化学习（RL）使大语言模型搜索智能体的精确匹配（EM）得分较强化学习前的初始化最高提升 32 分，但总体准确率提升并不能揭示训练过程中哪些底层功能能力得到了改善，或彼此如何交互。我们提出 CO-GRAFT，这是一种与奖励无关的干预式归因框架：经强化学习训练的智能体与其强化学习前的初始化，在实时检索的闭环执行中协同生成单条轨迹。每种策略分别控制四种功能角色中的一个互不重叠的子集：中间推理、搜索或回答决策、搜索查询生成以及最终答案提取。在不修改权重或重新训练的情况下评估所有 $2^4 = 16$种角色分配，由此定义一个合作博弈；该博弈的夏普利值将强化学习总增益分解到各角色，而 Möbius 交互项则区分角色单独带来的改善与多角色耦合。我们将 CO-GRAFT 应用于基于 Qwen2.5 和 Llama-3.2 模型的 Search-R1 智能体（涵盖两种模型规模、基础版和指令微调初始化、GRPO 和近端策略优化（PPO）优化器，以及自然问题（NQ）和 HotpotQA 数据集），结果表明，查询生成带来的增益从未超过总增益的 15%；相反，增益取决于强化学习前的答案有效性、搜索保留、模型规模以及固定的任务级搜索预算。这使归因重心从推理转向答案提取，尤其是在规模较大或答案响应检查点格式良好的情况下；而规模较小的基础模型则依赖推理与决策之间的耦合。

1 引言

大语言模型（LLM）搜索智能体通过在多轮交互中交替进行推理与外部检索，回答多跳且知识密集型问题（Jin等人，2025b；Song等人，2025；Chen等人，2025；Zheng等人，2025b；Sun等人，2025；Gao等人，2025；Guo等人，2025）。每一轮中，智能体会依次执行四种不同的功能角色（Jin等人，2025b；Song等人，2025；Chen等人，2025）：生成中间推理（R），做出二元决策（D），选择调用搜索或终止并作答；若选择搜索，则构造搜索查询（Q）；检索结束后，提取最终答案（A）。通过强化

强化学习（RL）通过精确匹配（EM）准确率，相较于非强化学习推理和检索基线提高6到25个百分点（Jin等人、2025b，a），并且在我们的评估中，相较于提示式强化学习前搜索智能体最高提高32个百分点。

然而，整体EM改进并不能揭示强化学习后训练期间哪些功能能力发生了变化，也不能揭示多种能力如何相互作用以提高准确率（Jin 等，2025a）。无论智能体是通过结果奖励（Jin 等，2025b；Chen 等，2025；Guo 等，2025）、过程奖励（Lightman 等，2024；Wang 等，2024b；Zheng 等，2025b），还是通过格式（Jin 等，2025b；Guo 等，2025）和检索效用奖励（Song 等，2025；Sun 等，2025；Gao 等，2025）等辅助奖励塑形进行训练，最终EM的提升都可能源于四种不同的角色行为（Jin 等，2025b,a）：制定更有效的搜索查询，更准确地对检索到的段落进行推理，更恰当地决定何时搜索、何时回答，或将最终答案格式化为符合评估标准的形式。确定究竟是哪些角色驱动了EM增益至关重要，尤其是对于结果奖励强化学习（Jin 等，2025b；Guo 等，2025）而言，因为其监督仅应用于最终答案；但这一点同样适用于任何经过训练的搜索智能体，可用于诊断失败模式并针对未来的训练改进进行调整。

现有的评估与信用分配方法无法在闭环执行过程中分离各功能角色的因果贡献。观察性行为分析衡量的是表层统计特征，例如响应长度（Jin等人，2025b；Guo等人，2025）、搜索频率（Jin

等人，2025b，a）或查询的词汇重叠度（Jin等人，2025a），这些指标与训练进展相关，却无法验证给定角色的变化是否确实提升了端到端 EM 准确率。单角色消融（Jin等人，2025a）一次只扰动一个组件，因此无法捕捉多角色之间的交互；要产生增益，强化学习策略必须同时控制两个或更多角色。离线轨迹干预（Zheng等人，2025b）评估的是静态前缀，而非闭环执行（结合实时检索的交互式反馈）；与此不同，过程奖励（Lightman等人，2024年；Wang等人，2024b）和步骤级信用分配方法（Gao等人，2025）会在优化过程中改变奖励信号，而不是分解已训练检查点所习得的能力。

为克服这些局限，我们提出**CO-GRAFT**（**协同策略嫁接**），这是一种与奖励无关的干预归因框架。在该框架中，经强化学习训练的智能体与其强化学习前的初始版本通过实时检索，共同生成一条轨迹。每个策略控制同一条轨迹中四个角色里互不重叠的子集（图 1b），为其负责的角色生成词元片段，并在角色边界处将生成的前缀传递给另一策略。评估所有 $2^4 = 16$ 角色组合，即可定义夏普利值（Shapley，1953），无需修改模型权重、访问训练奖励或重新训练。通过这 16 项测量，莫比乌斯交互项（Harsanyi，1963年）可将单角色的独立提升与成对及更高阶（多角色）依赖区分开来，而夏普利值则汇总每个角色在所有角色组合中的边际贡献，形成单一的加性归因分数。

将 CO-GRAFT 应用于 Search-R1 智能体（Jin等人，2025b，a），涵盖 Qwen2.5（Qwen等人，2025）和 Llama-3.2（Grattafiori等人，2024）模型（涉及两种模型规模、基础版和指令微调初始化，以及在自然问题（Shao等人，2024）和近端策略优化（Schulman等人，2017）优化器下的 NQ（Kwiatkowski等人，2019年）和 HotpotQA（Yang等人，2018）数据集），揭示了强化学习如何在不同角色之间重新分配能力。首先，搜索查询生成（Q）在任一数据集上所占的强化学习增益都从未超过 15%，这表明改进的检索查询并不是准确率提升的主要驱动因素。其次，强化学习前答案有效性、搜索保留和模型规模决定了归因如何在推理与答案提取之间分配：更大的模型以及指令微调的检查点在强化学习之前就能生成格式良好的答案，会将归因从推理（R）转向最终答案提取（A）；而较小的基础版模型则表现出推理与决策之间的强正耦合（R×D）。第三，决策（D）增益反映的是学习固定的任务级搜索预算，而不是根据问题级证据充分性来决定何时停止。

我们将贡献总结如下：

- 我们提出了 **CO-GRAFT**，一种干预归因框架，将搜索智能体轨迹划分为四个功能角色（R、D、Q 和 A），并在不修改权重或重新训练的情况下，在实时检索环境中评估经强化学习训练的智能体与其强化学习前初始化之间的所有 $2^4 = 16$角色分配组合。

- 我们利用夏普利值和莫比乌斯交互项，对强化学习带来的总精确匹配增益进行精确分解，从而区分独立的角色改进与成对及高阶交互作用。

- 我们使用该框架，在 Qwen2.5 和 Llama-3.2 模型上对 Search-R1 智能体进行评估，涵盖自然问题和 HotpotQA，结果表明，强化学习增益由强化学习前答案有效性、搜索保留、模型规模以及固定的任务级搜索预算驱动，而查询生成的贡献始终被限制在总增益的 $\leq 15\%$ 以内。

2 CO-GRAFT

我们引入 CO-GRAFT，这是一种将经强化学习训练的搜索智能体的性能增益分解到其轨迹功能角色上的归因框架。为确定强化学习改进的来源，我们首先将每条轨迹划分为四个完备且互不重叠的角色。随后，我们构建嫁接式轨迹展开，在受控的角色分配下，由强化学习策略及其强化学习前初始化共同生成轨迹，最后利用夏普利分解和莫比乌斯交互项量化独立贡献与协同贡献。

角色。CO-GRAFT 将搜索智能体的轨迹分解为四种不同角色，涵盖策略生成的每个 token。遵循标准的 Search-R1 轨迹格式（Jin等人，2025b），智能体交替生成推理块（<思考>r</思考>）与

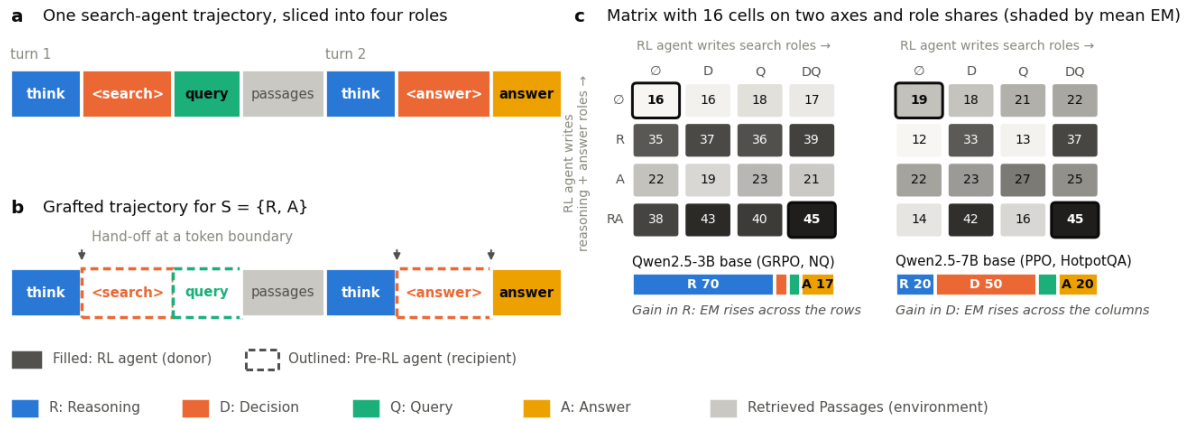


图 1：用于归因强化学习性能提升的 CO-GRAFT框架概览。(a) 将搜索智能体轨迹划分为四种角色：推理 R、决策 D、查询 Q 和答案 A；灰色段落由环境返回。(b) $S = \{\text{R,A}\}$的嫁接轨迹：强化学习智能体（实心）生成 R 和 A，而其强化学习前模型（轮廓线）在实时检索期间生成 D 和 Q。(c) 包含全部 16 种角色组合的 4×4 网格，根据强化学习智能体拥有的角色（行 $\cup$列）编制索引，其余角色则由强化学习前模型生成：列逐步加入搜索角色（D、Q），行逐步加入推理和答案角色（R、A），从 $\emptyset$（左上角，只有强化学习前模型）逐步到 $\{\text{R, D,Q, A}\}$（右下角，只有强化学习智能体）。每个单元格报告嫁接策略在相同的 1,000 个评估问题上的平均精确匹配 (%)（颜色越深表示精确匹配越高），并列出各角色在总增益中的 Shapley百分比份额 $v(N) - v(\emptyset)$，分别对应 Qwen2.5-3B 基础版（GRPO、自然问题）和 Qwen2.5-7B 基础版（近端策略优化、HotpotQA）。

搜索调用（<搜索> q </搜索>）——每次调用后，环境返回的段落都包含在 <信息>..... </信息> 标签中——随后以最终响应（<答案> a </答案>；图 1a）结束。我们将四种功能角色形式化为 $N = \{\text{R、D、Q、A}\}$：R 包含下一个动作标签之前生成的所有推理 token；D 负责离散动作选择（<搜索>、 <答案>或未完成轮次的终止）；Q 包含搜索查询字符串；A 则包含最终答案文本。由于 <信息>块由检索器从外部注入，而非由任一策略模型采样生成，因此它们不属于任何参与方。关键的是，这种划分将轨迹中模型生成的每个 token 都归入且仅归入一个角色，从而确保任务准确率的任何变化，都能直接追溯到输出 token 发生变化的特定角色。

嫁接式轨迹展开。 嫁接式轨迹展开是指由强化学习智能体及其强化学习前模型共同生成的一条轨迹，二者分别负责角色的不重叠子集（图1b）。令 π^+ 表示经强化学习训练的智能体（即供体），令 π^- 表示其完全一致的强化学习前初始化模型（即接收方），二者共享相同的分词器和提示模板。对于嫁接策略 $S \subseteq N$，每个角色片段 ρ 均由 π^+ 在 $\rho \in S$ 时生成，否则由 π^- 生成。在生成过程中，负责写入当前片段的策略会基于完整的前置前缀进行条件化，其中包括配对模型生成的词元，以及检索器针对每个搜索查询返回的实时段落。为防止策略交接过程中产生伪影，我们的执行框架强制执行三项不变性约束：(1) 由于模型在一次过程中生成推理及其后续动作标签，执行框架会丢弃 R 的负责方生成的标签，并让 D 的负责方从相同的前缀解码动作标签；(2) 所有交接严格发生在完整词元边界处，因此拆分词元（例如“␣<”在“ <search>”中）不会扭曲后续分词；(3) 只有在生成完整的开始和结束标签对时，执行框架才会解析动作，并将格式错误或未闭合的标签视为无效轮次，将错误提示传递给下一步。其余执行机制均遵循标准的搜索智能体推理循环，在 $S = \emptyset$ 处恢复纯接收方行为，在 $S = N$处恢复纯供体行为。

Shapley 归因。评估所有 $2^4 = 16$角色组合（图1c）定义了一个合作博弈，其夏普利值可对强化学习增益进行精确且与顺序无关的归因。令$v(S)$ 表示嫁接策略 S 在固定评估集上的平均精确匹配（EM）分数，其中$v(\emptyset)$ 和 $v(N)$ 分别表示接收方和供体基线。角色

$i \in N$（Shapley，1953）是

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} w_S \left[v(S \cup \{i\}) - v(S) \right], \quad (1)$$

其中 $w_S = |S|!\,(|N| - |S| - 1)!/|N|!$ 根据边际贡献所占的比例加权4! 种角色排列中，子集 S 先于角色 i出现。根据效率公理，$\sum_{i \in N} \phi_i = v(N) - v(\emptyset)$ 恒等成立，从而确保四个角色归因之和等于强化学习的精确总增益。我们将角色份额表示为该总增益的归一化百分比（图 1c 中的柱形）（$\phi_i / [v(N) - v(\emptyset)]$）。由于总增益接近零时百分比归一化会变得不稳定，因此只有当整体强化学习增益至少达到 4 个精确匹配点（即单点解码噪声下限的四倍）时，我们才报告份额；更小的差异则标记为“无增益”。

莫比乌斯交互。Shapley 值将每个角色的总归因汇总为单个标量，而莫比乌斯交互项（Harsanyi，1963年）则将这份归因拆分为角色单独获得的部分，以及仅在与其他角色共同协作时才获得的部分。对于每个包含两个或更多角色的组合 $T \subseteq N$ （$|T| \geq 2$），交互项为

$$m(T) = \sum_{S \subseteq T} (-1)^{|T| - |S|}\, v(S). \quad (2)$$

直观而言，$m(T)$分离出只有当 π^+ 同时控制T 中的所有角色时才能实现的剩余增益：当各角色独立贡献时，该值为零；协同耦合时，该值为正；角色相互重叠、形成替代项时，该值为负。由于即使整体强化学习增益较小，原始交互幅度仍然具有信息价值，因此我们直接以 EM 分数报告所有评估模型的莫比乌斯项。

3 实验设置

在本节中，我们将详细介绍搜索智能体、数据集、指标、训练方案、检索设置、轨迹测试框架和辅助对照测试，并使用我们提出的 CO-GRAFT框架研究不同智能体角色对强化学习性能增益的归因。我们考察公开发布的 Search-R1(Jin等人，2025b，a) 检查点，并通过使用相同的 Search-R1 训练和评估设置训练其他模型来补充这些检查点。

3.1 智能体

接收模型。我们从两个模型系列中选择搜索智能体 × 两个起始点 × 两种规模。第一个系列包括 Qwen2.5-3B 和 Qwen2.5-7B，每种模型均有基础版和指令微调(Qwen等人，2025) 两种版本。第二个系列包括 Llama-3.2-1B 和 Llama-3.2-3B，每种模型均有基础版和指令微调(Grattafiori等人，2024年) 两种版本。每个接收模型都使用其供体的训练提示运行，因此 $v(\emptyset)$ 是经提示处理的强化学习前模型，先前研究将其报告为基线（Jin等人，2025b）。

供体模型。对于上述每个对应的接收模型，我们都会获得一个经过强化学习（GRPO）训练的供体智能体。Qwen2.5-3B、Qwen2.5-7B 和 Llama-3.2-3B（基础版和指令微调版）的供体模型均已公开。对于 Llama-3.2-1B（基础版和指令微调版），我们遵循相同的训练方案（§3.3）自行训练供体模型。

3.2 评估数据集与指标

我们使用与 Search-R1 设置相同的评估数据集。由于让每个问题都经过 16 种角色分配的计算成本很高，因此我们不使用完整测试集，而是在固定的随机子集上进行评估：NQ（Kwiatkowski 等，2019年）和 HotpotQA（Yang 等，2018）各选取 1,000 个问题。我们的主要指标是最后一个答案块上的精确匹配，同时还报告有效答案率（包含格式正确的答案块的问题所占比例）。

3.3 训练方案

我们采用 Search-R1 的 GRPO 训练方案来训练搜索智能体，其中包括：最终答案的结果精确匹配奖励、格式正确的轨迹对应的 0.2 格式奖励、从损失中屏蔽检索到的词元、进行四个搜索轮次（每轮最多生成 500 个词元）、组大小为 5、学习率10^{-6}、批量大小为 512、KL 系数为 10^{-3}，以及 $1k$ 个步数（每步 240 个提示）。此外，我们在 NQ 和 HotpotQA 的合并训练集划分上进行训练，并分别在两个测试集上进行评估。

3.4 检索环境

遵循 Search-R1 的设置（Jin等人，2025b），我们使用 E5-base-v2（Wang等人，2024a）作为稠密双编码器，在包含 2100 万个段落的 2018 年维基百科转储数据集上执行精确内积搜索。每次展开期间都会实时查询检索器，并返回排名前三的段落作为检索到的证据。

3.5 执行框架

为执行多轮展开，训练和评估中的每条轨迹最多包含四个搜索轮次，随后是一个最终答案轮次；每轮最多生成 500 个令牌。任何未以完整动作标签结束的轮次都会收到无效动作提示，并在轮次预算允许时进入下一轮。

由于轨迹嫁接会改变展开的执行流程，我们通过在原始 Search-R1 代码库和 CO-GRAFT 执行框架下，使用相同的评估集评估 $S = \emptyset$（仅强化学习前智能体）和 $S = N$（仅强化学习智能体），验证在未发生角色切换时，CO-GRAFT 执行框架是否能保留单策略行为。在 32 次比较中（8 个模型对、两个数据集和两种策略行为），CO-GRAFT 在 28 种情况下的准确率与 Search-R1 相差不超过一个精确匹配点的波动范围，其余四种情况下的差异最多为 2.3 分。因此，该执行框架能够复现基线策略行为，且不会引入实现伪影。

3.6 控制实验

我们进行了三项控制实验，以区分角色归因与格式伪影、固定搜索预算及查询文本条件化的影响。

答案内容与格式。为区分标签遵循度与严格精确匹配下的事实性提升，我们使用宽松指标（子串匹配、令牌 F1 和有效答案率）重新计算归因。在自然问题数据集上进行宽松评分时，A 对于 Qwen2.5 模型，归因会大幅减弱（例如，Qwen2.5-7B 基础版从 38% 降至 -5%），但对 Llama-3.2-1B-Instruct 仍然存在（12 对 16 个精确匹配分），这表明A在 Qwen2.5 中，增益反映的是格式遵循度的提升；而在 Llama-3.2 中，则反映了内容质量的改善。不过，表1 报告的是严格精确匹配结果。

自适应停止 vs. 固定搜索深度。为检验 D 学习的是按问题进行自适应停止，还是统一的搜索次数，我们将 D 替换为固定预算 $B \in \{0, \ldots, 4\}$，同时保留强化学习智能体的其余角色。在所有评估数据对中，最优固定预算与完整强化学习智能体的表现相差不超过一个精确匹配点（例如，对于 Qwen2.5-3B 近端策略优化，在自然问题上为 $B = 1$，在 HotpotQA 上为 $B = 2$），这表明模型学习到的是任务级搜索深度，而非自适应停止。

段落检索与查询条件化。为区分检索质量与查询文本条件化，我们在 Qwen2.5-3B 和 Qwen2.5-7B（近端策略优化）上的全部 16 种组合中，固定强化学习智能体检索到的段落。冻结段落会使 Q 的归因降低 60–100%（0.6–2.5 个精确匹配分数），这证实 Q 的增益来自检索，而非文本条件化。然而，向强化学习前模型提供强化学习检索到的段落，最多只能增加 3 个精确匹配分数；而让强化学习智能体控制 R，在实时检索和冻结检索两种情况下都能带来 11–22 个精确匹配分数的增益。

4 结果

本节将在表 1 中呈现主要结果，并从多个因素出发理解这些发现，包括问题复杂度、模型规模和指令微调。

4.1 问题复杂度的影响：单跳 vs. 多跳

问题复杂度对角色分布的影响仅处于次要地位，主要表现为在推理与答案之间转移归因，而总体模式仍由搜索行为驱动。我们比较了七个模型在单跳自然问题与多跳 HotpotQA 上的角色归因（其中两个模型因其强化学习增益低于报告阈值而被排除），发现查询角色在不同任务类型之间保持不变，且占比始终不超过 15%。相比之下，推理与答案呈现出明显的权衡：在多跳 HotpotQA 上，推理归因始终更高（七个模型中有六个如此，最大差距达 33 个百分点）；而在单跳自然问题上，答案归因更高（七个模型中有六个如此，最大差距达 29 个百分点）。这反映出，多跳查询会促进归因从最终答案生成转向中间推理步骤。与此同时，当搜索行为保持稳定时（搜索调用达到非零水平），决策占比低于 25%；但当模型增加或减少搜索调用时，决策占比会大幅上升（当搜索次数降至零时，最高可达 38–86%）。总体而言，问题复杂度造成的分布偏移小于搜索行为变化所引发的差异。

4.2 模型规模的影响：较小模型与较大模型

模型规模会改变哪些角色获得归因，将归因从推理和决策转向答案和查询。在两种规模的增益都至少达到四个百分点的三个模型系列中

接收方	type	data	强化学习前	RL	Δ	srch	有效	R	D	Q	A
Qwen2.5-3B	base	NQ	15.8	45.2	+29.4	0.9→1.1	53→98	**70**	7	6	17
		HotpotQA	9.2	38.9	+29.7	0.9→1.6	51→97	**81**	1	5	13
Qwen2.5-7B	base	NQ	25.5	48.2	+22.7	1.4→1.1	79→100	**40**	9	13	38
		HotpotQA	18.3	44.1	+25.8	1.6→1.6	66→100	**45**	18	5	32
Qwen2.5-3B	指令微调 自然问题		21.8	41.3	+19.5	1.0→1.3	97→100	36	21	4	**39**
		HotpotQA	22.0	34.9	+12.9	1.2→1.9	94→99	**69**	12	4	15
Qwen2.5-7B	指令微调 自然问题		30.9	41.9	+11.0	1.0→1.2	97→99	34	6	14	**46**
		HotpotQA	30.7	39.0	+8.3	1.3→1.6	95→99	35	6	14	**45**
Llama-3.2-1B 基础版		NQ	2.1	17.2	+15.1	0.2→0.0	40→100	27	26	0	**47**
		HotpotQA	1.7	17.8	+16.1	0.2→0.0	18→100	**40**	38	0	22
Llama-3.2-3B 基础模型		NQ	12.6	44.2	+31.6	0.7→1.0	66→100	33	18	15	**34**
		HotpotQA	9.6	32.0	+22.4	0.7→1.0	63→100	23	25	12	**40**
Llama-3.2-1B 指令微调 自然问题			12.7	28.8	+16.1	0.8→0.8	68→98	-23	41	16	**66**
		HotpotQA	6.8	19.6	+12.8	0.8→0.8	57→96	1	36	16	**46**
Llama-3.2-3B 指令模型 自然问题			29.0	32.2	+3.2	1.4→0.0	86→100	— 无增益 —			
		HotpotQA	21.3	22.6	+1.3	1.6→0.0	83→100	— 无增益 —			

表 1：使用我们提出的 CO-GRAFT框架进行的强化学习增益归因研究的主要结果。强化学习前、强化学习和 Δ 分别表示强化学习前智能体、强化学习智能体的精确匹配（%）及二者之差；而搜索 和有效答案 分别表示每个问题的搜索次数，以及强化学习前智能体 → 强化学习智能体所示的有效答案率（%）。当增益（Δ）超过 4 个百分点时，报告各角色（R：推理、D：决策、Q：查询、A：答案）的占比（%）。

（Qwen2.5 基础版、Qwen2.5 指令微调和 Llama-3.2 基础版），在六组比较中的五组里，较大模型分配给答案与查询的归因总计多出 17 至 40 个百分点。这种转变在 Qwen2.5 基础版中更为明显：3B 模型高度依赖推理（自然问题上为 70% 对 40%；HotpotQA 上为 81% 对 45%），而 7B 模型则将更多归因转向答案，高出 19 至 21 个百分点；Qwen2.5 指令微调模型在 HotpotQA 上也呈现出类似的权衡（推理为 69% 对 35%；答案为 15% 对 45%）。同样，在 Llama-3.2 基础版中，1B 模型放弃了搜索（查询 $\leq$ 1%），而 3B 模型则持续进行搜索调用，并将 12–15% 的归因分配给查询。由于较大模型在强化学习前的有效答案率更高（例如，Qwen2.5 在自然问题上为 53% 对 79%），但 Qwen2.5-Instruct 除外，其有效答案率相同（97%），因此增益归因总体上表明，模型规模增大推动归因转向答案和查询。

4.3 指令微调的影响：基础版与指令微调版

指令微调不会使归因分配发生统一变化；相反，其影响取决于后训练如何改变模型的有效答案率和搜索轨迹。在四组基础模型–指令模型对中，我们发现，有效答案率较高的指令微调模型获得的强化学习增益较小，且归因更多转向答案；而放弃搜索的模型则会使强化学习增益停滞，并打破这一规律。具体而言，在 Qwen2.5 系列（3B 和 7B）中，指令微调变体的有效答案率起点比对应基础版高 18 至 44 个百分点，因此强化学习需要学习的格式内容更少，整体增益也低 9.9 至 17.5 个百分点，同时归因从推理转向答案。在自然问题上，推理的归因变化为（{v1}至 {v2} 个百分点），答案的归因变化为（{v3}至 {v4} 个百分点）。

相较之下，Llama-3.2-1B 的趋势恰好相反：两个变体的有效答案率起点都较低，但指令微调模型由于保留了搜索能力（0.9 至 1.0，而基础版为 0.2 至 0.0），增益高出 6.1 至 11.2 个百分点，使查询归因提高了 10 至 12 个百分点。最后，Llama-3.2-3B 指令模型体现了这两种行为的叠加效应——有效答案率起点较高（在自然问题上为 86%，基础版为 66%），同时完全放弃搜索（从 1.4–1.6 降至 0.0），因此增益几乎可以忽略（+1.3至 +3.2 个百分点，而基础版为 +22.4 至 +31.6 个百分点）。综合来看，这些比较表明，强化学习前答案有效性和搜索保留情况，比单独考察指令微调更能可靠地预测角色归因。

5 分析

我们深入分析了四个具有代表性的 Search-R1 检查点，以揭示不同角色如何相互作用，以及强化学习优化器是否会改变归因方式。从 Qwen2.5-3B 基础版开始——它在表 1 中的推理占比最高——我们选取了三组受控消融实验，每组仅改变一个维度：指令微调（Qwen2.5-3B 指令版）、同一模型系列内的模型规模（Qwen2.5-7B 基础版），以及相近规模下的模型系列（Llama-3.2-3B 基础模型）。由于这四个模型均有公开的 GRPO 和近端策略优化检查点，这一统

		莫比乌斯项															沙普利值			
		一个角色				两个角色							三个和四个角色							
智能体	gain	R	D	Q	A	R×D	R×Q	R×A	D×Q	D×A	Q×A	RDQ	RDA	RQA	DQA	RDQA	R	D	Q	A
NQ																				
Qwen2.5-3B 基础版	29.4		+0.6		+5.7	+2.0	−1.1	−2.2	−1.7	−2.9	−1.4	+1.8	+5.2	+1.6	+2.1	−1.6	**+20.6**	+1.9	+1.8	+5.0
Qwen2.5-3B 指令微调	19.5		+6.2			−2.5	−4.4	−8.9	−1.7	−7.3	−4.1	+2.9	+7.7	+4.1	+3.2	−4.1	+7.0	+4.0	+0.9	**+7.6**
Qwen2.5-7B 基础版	22.7	+11.7	+4.7		+8.6	−4.6	−0.6	+0.9	−1.6	−0.7	+0.1	+2.1	+0.4		+3.0	−3.8	+9.1	+2.1	+2.9	+8.6
Qwen2.5-3B 基础版	31.6	+0.1			+7.0	+8.8	+0.4	+2.1	−2.4	+0.7	−0.3	+7.0	+4.0	+0.5	+1.0	+2.8	+10.3	+5.8	+4.8	**+10.8**
HotpotQA																				
Qwen2.5-3B 基础版	29.7	+21.4	−0.5	+1.5	+4.2	+2.8	+2.7	+0.8	−0.6	−0.8	−0.1	−0.6	+1.1		+0.4	−0.8	**+23.9**	+0.3	+1.6	+3.9
Qwen2.5-3B 指令版	12.9		+0.9		+4.0	+1.0	−4.5	−4.5	−2.1	−0.3	−3.1	+3.3	+0.8	+5.0	+1.3	−1.8	+8.9	+1.6	+0.5	+2.0
Qwen2.5-7B 基础版	25.8	+9.9	+5.8		+6.2	+1.3	−1.1	+7.8	−0.6	+1.9	+0.1	+0.6		0.0	−0.7	+0.8	**+11.6**	+4.6	+1.3	+8.4
Llama-3.2-3B 基础模型	22.4	−0.2			+5.1	+5.4	0.0	+2.2	+0.2	+3.0	+1.1	+3.0			+0.1	+3.3	+5.3	+5.6	+2.6	**+8.9**

表 2：四个 GRPO智能体在自然问题和 HotpotQA 上的莫比乌斯项与夏普利值（EM 分数）。角色 R、D、Q 和 A 与表 1 中的含义相同。行总和对应增益的 15 个莫比乌斯项。单元格底色以蓝色突出显示正项，以红色突出显示负项。最大的夏普利值以粗体显示。

一的选择既支持我们对下文角色交互的分析，也支持对优化器敏感性的分析。

5.1 角色交互：成对项与高阶项

虽然夏普利值量化了每个角色获得的归因，但莫比乌斯交互项（公式 2；表 2）揭示了这些增益究竟是独立产生的，还是通过多角色协同产生的。当两个角色带来的协同增益超过其各自单独作用之和时，成对交互为正；当二者作为替代项产生重叠时，则为负。在我们的四个重点检查点和两个数据集上，成对项已经改变了表 1 中份额的解读方式：在 Llama-3.2-3B 基础模型中，R×D 在 NQ（+8.8）和 HotpotQA（+5.4）上的成对项均为最大，而单独的推理（+0.1 和 −0.2）与决策（−2.5 和 −0.5）自身几乎没有或完全没有贡献，这表明两个角色必须协同作用，才能提升精确匹配率。相反，在 Qwen2.5-3B 指令微调模型的 NQ 上，与答案之间为负的成对项（R×A−8.9、D×A −7.3）以及较强的单独贡献（推理为 +11.0、决策为 +6.2 、答案为 +13.8 ，总增益为 19.5）表明存在冗余重叠，而非联合协同；与此同时，Qwen2.5-7B 基础版表现出任务敏感性（单跳与多跳）：R×A 在 NQ 上从+0.9 变为在 HotpotQA 上的 +7.8。

高阶莫比乌斯项能进一步阐明多角色动态，但前提是将其与低阶效应结合起来解读，而非孤立看待。由于三方或四方项反映的是扣除低阶项后剩余的部分，因此高阶值为正，可能抵消负向的成对交互，而不能证明所有纳入角色之间存在同时协同；在 Qwen2.5-3B 基础版中，对自然问题上的 R×D×A（+5.2 抵消了 R×A −2.2 和 D×A −2.9）以及 Qwen2.5-3B 指令微调版（+7.7 抵消了 −8.9 和 −7.3）的结果便体现了这一点。相比之下，Llama-3.2-3B 基础版在自然问题上呈现正向的 R×D×Q 项（+7.0 ），同时成对项为小幅正值或非负值（R×D +8.8、R×Q+0.4、D×Q −2.4），这表明强化学习查询在推理与决策共同发挥作用后能带来额外价值。总的来说，沙普利值指出归因落在何处，而莫比乌斯项则表明这部分贡献体现的是独立能力（Qwen2.5-3B 基础版）、协同适应（Llama-3.2-3B 基础版），还是角色重叠（Qwen2.5-3B 指令微调版）。这项分析可为强化学习迭代训练提供指导，帮助判断奖励塑形应针对单个角色，还是联合角色组合。

5.2 RL优化器敏感性：GRPO 与近端策略优化

只有当强化学习优化器将搜索行为推至轮次上限时，它才会改变主导角色；否则，初始检查点决定哪些角色获得性能增益。比较四种架构的公开 GRPO 和近端策略优化检查点（表3），我们发现，在三种架构中，两种优化器下的主导角色保持一致：在 Qwen2.5-3B 基础版上，推理占主导地位（自然问题上 GRPO 为 70%，近端策略优化为 83%）；Qwen2.5-3B 指令微调版在自然问题上更偏向答案（39% 对 46%），在 HotpotQA 上则更偏向推理（69% 对 84%）；Llama-3.2-3B 基础版在自然问题上，推理和答案的占比相差不超过六个百分点（GRPO 下分别为 33% 和 34%，近端策略优化下分别为 38% 和 32%），而在 HotpotQA 上更偏向答案（40% 对 37%）。值得注意的是，即便近端策略优化提高了搜索频率，相较于 GRPO

模型	data	opt.	Δ	搜索	R	D	Q	A
Qwen2.5-3B base	NQ	GRPO	+29.4	1.1	**70**	7	6	17
		PPO	+27.7	2.1	**83**	-1	2	16
	Hotpot	GRPO	+29.7	1.6	**81**	1	5	13
		近端策略优化	+26.8	2.2	**67**	7	8	18
Qwen2.5-3B 指令微调	NQ	GRPO	+19.5	1.3	36	21	4	**39**
		PPO	+12.8	1.9	31	17	6	**46**
	Hotpot	GRPO	+12.9	1.9	**69**	12	4	15
		PPO	+11.3	2.1	**84**	7	-9	18
Qwen2.5-7B base	NQ	GRPO	+22.7	1.1	**40**	9	13	38
		PPO	+20.8	4.0	-2	**77**	7	11 39
	Hotpot	GRPO	+25.8	1.6	**45**	18	5	32
		PPO	+26.6	3.9	20	**50**	10	20
Llama-3.2-3B base	自然问题	GRPO	+31.6	1.0	33	18	15	**34**
		PPO	+28.1	1.0	**38**	17	13	32
	Hotpot	GRPO	+22.4	1.0	23	25	12	**40**
		近端策略优化	+21.9	1.0	25	22	16	**37**

表3：GRPO 与 PPO 公开检查点之间的角色增益归因。Δ，*srch*、R、D、Q 和 A 与表1中的含义相同。占比最大的角色以粗体标示。

在 Qwen2.5-3B 的两个变体上也是如此（自然问题上的每问题搜索次数分别为 2.1 对 1.1 次，以及 1.9 对 1.3 次），这表明搜索深度的适度变化不会重塑角色归因。唯一的例外是 Qwen2.5-7B 基础版：在自然问题上，近端策略优化使搜索次数达到四次的轮次上限，将 77% 的归因分配给决策，并使推理的独立占比为负（−27%）；而 GRPO 平均搜索 1.1 次，并将归因均分给推理（40%）和答案（38%）。总体而言，决定角色分布的是初始模型，而非强化学习优化器；只有当优化器将搜索深度推至轮次边界时，才会打破这一分布。

6 相关工作

以往关于搜索智能体强化学习的研究，主要评估端到端任务性能，而非分解分析哪些行为变化带来了增益。现有的强化学习搜索智能体（Jin 等，2025b,a；Song 等，2025；Chen 等，2025；Zheng 等，2025b；Sun 等，2025；Gao 等，2025）优化基于结果的奖励，并报告汇总的精确匹配（EM）分数。将优化限制在查询生成（Jiang et al., 2025）或为搜索步骤和答案步骤分别分配奖励（Wang et al., 2025）的方法，预先指定了哪些轨迹角色需要改进，而不是衡量它们的实际贡献；对训练配置的消融研究（Xu et al., 2026）改变的是算法超参数，而非功能角色。此外，策略内信用分配方法，包括步骤级奖励（Zheng et al., 2025a；Wang 等人，2026）、轮次删除（Zhu 等人，2026）、反事实分支（Li 等人，2026a）、文档重放（Mukhopadhyay 等人，2026）和逐角色优势估计（Wu 等人，2026），只在单一策略内部计算信号，不会将训练后的智能体与其强化学习前的初始化版本进行比较。CO-GRAFT 直接衡量训练后的策略与其初始化版本在各功能角色上的性能差异。

第二类研究在推理期间组合多个模型，或定位后训练带来的行为变化。Venhoff 等人(2026）和Choi 等人(2026）在闭卷生成过程中，按单个词元切换基础版与后训练检查点；而 CO-GRAFT 则在多轮轨迹中分配完整的功能角色，并利用实时检索器反馈，精确归因强化学习增益。为定位强化学习改进，Li 等人(2026b）将增益归因于权重空间中的参数子集，而Lyu(2026）证明，数学强化学习通过可验证奖励带来的域外增益，主要源自答案格式；这一模式与我们的 Qwen2.5 研究结果一致。最后，ShapleyFlow（Yang 等人，2026）将夏普利值应用于固定智能体工作流中的静态模块，而 CO-GRAFT 则在交互式检索过程中，将两种策略间的训练增益归因于不同的功能轨迹角色。

7 结论

我们提出了 CO-GRAFT，这是一种框架，可在闭环执行下将经强化学习训练的搜索智能体的精确匹配增益分解到其四种轨迹角色中。研究表明，强化学习前的答案有效性、搜索保留和模型规模决定了增益究竟源于推理与决策的耦合、固定的任务级搜索预算，还是最终答案提取；而在自然问题和 HotpotQA 上，搜索查询生成（Q）带来的增益始终不超过总增益的 15%。基于结果的精确匹配无法揭示训练过程中哪些功能行为得到改善，因此，未来针对检索能力的强化学习奖励设计应明确监督查询构造（Q）和证据充分的停止决策（D）。

局限性

CO-GRAFT 归因的是一对策略之间的性能差异，而不是评估单个检查点；其解读受到合作

博弈论分解的性质以及我们的实验设置所影响。归因份额衡量相对于所选基线策略（强化学习前的初始化）的增益；夏普利值则对所有角色排列中的边际贡献取平均。当多角色交互显著时（例如 Qwen2.5-7B PPO 配对），仅凭加性沙普利份额无法捕捉联合依赖关系，因此必须结合莫比乌斯交互项来解读。此外，功能边界依据结构化动作标签进行划分，将内部规划、段落综合和结构化格式编排一并归入中间推理角色（R）；对 R 进行更细粒度的分解则留待未来工作。

我们的实证评估也受到模型规模、推理服务随机性和诊断范围的限制。计算资源限制使我们本地训练的模型规模不超过70亿参数，因此无法纳入14B或320亿模型配对；而批量贪心解码会引入约一个精确匹配点的服务方差。此外，固定预算和检索交换诊断仅涵盖自然问题和 Hot- potQA 上的两个 Qwen2.5 PPO 配对；在评估集上选择最优预算，则会在这一噪声底限内提供一个上界估计。最后，采用查询级过程奖励、更长优化计划或实时网络搜索的训练方案，可能会以不同方式在各角色间重新分配增益，而 CO-GRAFT 可直接用于衡量这些变化。

参考文献

Mingyang Chen、Linzhuang Sun、Tianpeng Li、Haoze Sun、Yijie Zhou、Chenzheng Zhu、Haofen Wang、Jeff Z. Pan、Wen Zhang、Huajun Chen、Fan Yang、Zenan Zhou 和 Weipeng Chen。2025。研究：通过强化学习教会大语言模型借助搜索进行推理。预印本，arXiv:2503.19470。

Youngbin Choi、Sehyun Hwang、Suhyeon Jun、Seoyeon Lee、Sunghyun Choi、MoonJeong Park 和 Dong- woo Kim。2026。新模型，旧习惯：多轮大语言模型智能体中模型切换时的行为迁移。预印本，arXiv:2610.19712。

Jiaxuan Gao、Wei Fu、Minyang Xie、Shusheng Xu、Chuyi He、Zhiyu Mei、Banghua Zhu 和 Yi Wu。2025。超越十轮：借助大规模异步强化学习，释放长程智能体搜索的潜力。预印本，arXiv:2508.07976。

Aaron Grattafiori、Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al- Dahle、Aiesha Letman、Akhil Mathur、Alan Schel- ten、Alex Vaughan、AmyYang、AngelaFan、Anirudh Goyal、Anthony Hartshorn、Aobo Yang、Archi Mitra、Archie Sravankumar、Artem Korenev、Arthur Hinsvark，以及其他542人。2024年Llama 3 系列模型。预印本，arXiv:2407.21783。

Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Peiyi Wang、Qihao Zhu、Runxin Xu、Ruoyu Zhang、Shirong Ma、Xiao Bi、Xiaokang Zhang、Xingkai Yu、Yu Wu、Z. F. Wu、Zhibin Gou、Zhihong Shao、Zhu- oshu Li、Ziyi Gao、Aixin Liu 等另外 175 人。2025。Deepseek-r1 通过强化学习激励大语言模型进行推理。《自然》，645(8081):633–638。

John C. Harsanyi。1963年。n人合作博弈的简化议价模型。《国际经济评论》，4：44–70。

Pengcheng Jiang、Xueqiang Xu、Jiacheng Lin、Jinfeng Xiao、Zifeng Wang、Jimeng Sun 和 Jia wei Han。2025。s3：通过强化学习训练搜索智能体并不需要那么多数据。载于*2025*年自然语言处理实证方法会议论文集，第21599–21617页，中国苏州。计算语言学协会。

Bowen Jin、Jinsung Yoon、PriyankaKargupta、Sercan O Arik 和 Jiawei Han。2025a。关于大语言模型智能体中强化学习与推理-搜索交错的实证研究。*arXiv*预印本 *arXiv:2505.15117*。

Bowen Jin、Hansi Zeng、Zhenrui Yue、Jinsung Yoon、Sercan Arik、Dong Wang、Hamed Zamani 和 Jiawei Han。2025b。Search-r1：通过强化学习训练大语言模型进行推理并利用搜索引擎。*arXiv*预印本 *arXiv:2503.09516*。

Tom Kwiatkowski、Jennimaria Palomaki、Olivia Red- field、Michael Collins、Ankur Parikh、Chris Alberti、Danielle Epstein、IlliaPolosukhin、Jacob Devlin、Ken- ton Lee、Kristina Toutanova、Llion Jones、Matthew Kelcey、Ming-Wei Chang、Andrew M. Dai、Jakob Uszkoreit、Quoc Le 和 Slav Petrov。2019年。自然问题：问答研究的一个基准。计算语言学协会会刊，7：452–466。

Mingchen Li、Hansi Zeng、ZhuoQian、Jiatan Huang、Sunjae Kwon、Hamed Zamani 和 Hong Yu。2026a。Rice-po：将检索交互转化为推理智能体的信用信号。预印本，arXiv:2605.26352。

Xirui Li、Ming Li 和 Tianyi Zhou。2026b。强化学习能为视觉推理带来什么提升？一种弗兰肯斯坦式分析。预印本，arXiv:2602.12395。

Hunter Lightman、VineetKosaraju、Yuri Burda、Harri- son Edwards、Bowen Baker、Teddy Lee、Jan Leike、John Schulman、Ilya Sutskever 和 Karl Cobbe。2024年。让我们逐步验证。载于国际学习表征会议，第 2024 卷，第 39578–39601 页。

Zongye Lyu。2026。相同答案，新框：RLVR 的域外增益有多少来自答案格式？预印本，arXiv:2610.20639。

Debayan Mukhopadhyay、Utshab Kumar Ghosh 和 Shubham Chatterjee。2026。桥接证据：静态检索效用无法预测多步智能体搜索中的因果效用。预印本，arXiv:2607.15253。

Qwen、An Yang、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chengyuan Li、Dayiheng Liu、Fei Huang、Haoran Wei、Huan Lin、JianYang、Jianhong Tu、Jianwei Zhang、Jianxin Yang、Jiaxi Yang、Jingren Zhou、Junyang Lin 等另外 24 位作者。2025。Qwen2.5 技术报告。预印本，arXiv:2412.15115。

John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。2017。近端策略优化算法。预印本，arXiv:1707.06347。

Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、Y. K. Li、Y. Wu 和 Daya Guo。2024年。DeepSeekMath：在开放语言模型中推动数学推理能力的极限。预印本，arXiv:2402.03300。

Lloyd S Shapley. 1953. n 人博弈的一个值。载于 Harold W. Kuhn 和 Albert W. Tucker 编，《博弈论贡献 II》，第 307–317 页。普林斯顿大学出版社，普林斯顿。

Huatong Song、Jinhao Jiang、Yingqian Min、Jie Chen、Zhipeng Chen、Wayne Xin Zhao、Lei Fang 和 Ji- Rong Wen。2025。R1-searcher：通过强化学习激励大语言模型的搜索能力。预印本，arXiv:2503.05592。

Hao Sun、Zile Qiao、Jiayan Guo、Xuanbo Fan、Yingyan Hou、Yong Jiang、Pengjun Xie、Fei Huang 和 Yan Zhang。2025。ZeroSearch：无需搜索即可激励大语言模型的搜索能力。*arXiv* 预印本 *arXiv:2505.04588*。

Constantin Venhoff、Iván Arcuschin、Philip Torr、Arthur Conmy 和 Neel Nanda。2026。基础模型懂得如何推理，思考模型学会何时推理。载于第 *43* 届国际机器学习会议论文集，第 306 卷，收录于机器学习研究会议录，第 123688–123732 页。PMLR。

Junzhe Wang, Zhiheng Xi, Yajie Yang, Hao Luo, Shi- han Dou, Tao Gui, and Qi Zhang。2026。通过贡献加权组相对策略优化增强基于大语言模型的搜索智能体。载于计算语言学协会第*64*届年会论文集（第*1*卷：长篇论文），第31704–31718页，圣迭戈，加利福尼亚州，美国。计算语言学协会。

Liang Wang、Nan Yang、Xiaolong Huang、Binx- ing Jiao、Linjun Yang、Daxin Jiang、Rangan Ma- jumder 和 Furu Wei。2024a。通过弱监督对比式预训练获得文本嵌入。预印本，arXiv:2212.03533。

Peiyi Wang、Lei Li、Zhihong Shao、Runxin Xu、Damai Dai、Yifei Li、Deli Chen、Yu Wu 和 Zhifang Sui。2024b。Math-shepherd：无需人工标注，逐步验证并强化大语言模型。载于第*62*届计算语言学协会年会论文集（第*1*卷：长篇论文），第9426–9439页，泰国曼谷。计算语言学协会。Yiding Wang、Zhepei Wei、Xinyu Zhu 和 Yu Meng。2025。超越结果奖励：解耦搜索与回答以改进大语言模型智能体。预印本，arXiv:2510.04695。Xingyu Wu、Yuchen Yan、Zhengxi Lu、Siqi Chen、Xin Zhang、Aiting Liu、Chao Deng、Jie Liu、Jin Ma、Jian Shao、Jun Xiao 和 Yongliang Shen。2026。ItерSynth：通过角色解耦的迭代合成重新思考深度搜索智能体。载于*2026*年自然语言处理实证方法会议论文集，第49432–49458页，匈牙利布达佩斯。计算语言学协会。Yinuo Xu、Shuo Lu、Jianjie Cheng、Meng Wang、Qian- long Xie、Xingxing Wang、Ran He 和 Jian Liang。2026。如何训练深度研究智能体？Search-R1中的提示、奖励与策略优化。预印本，arXiv:2602.19526。

Yingxuan Yang, Bo Huang, Siyuan Qi, Chao Feng, Haoyi Hu, Yuxuan Zhu, Jinbo Hu, Haoran Zhao, Ziyi He, Xiao Liu, ZongYu Wang, Muning Wen, Lin Qiu, Xuezhi Cao, Xunliang Cai, Yong Yu, and Weinan Zhang。2026。模块化智能体工作流的基于归因的分析与优化。载于计算语言学协会会议论文集：*ACL 2026*，第7247–7264页，圣迭戈，加利福尼亚州，美国。计算语言学协会。

Zhilin Yang、Peng Qi、Saizheng Zhang、Yoshua Bengio、William Cohen、Ruslan Salakhutdinov 和 Christo- pher D. Manning。2018。HotpotQA：一个用于多样化、可解释多跳问答的数据集。载于*2018* 年自然语言处理经验方法会议论文集，第 2369–2380 页，比利时布鲁塞尔。计算语言学协会。Xuhui Zheng、Kang An、Ziliang Wang、Yuhang Wang 和 Yichao Wu。2025a。StepSearch：通过逐步近端策略优化激发大语言模型的搜索能力。载于*2025* 年自然语言处理经验方法会议论文集，第 21805–21830 页，中国苏州。计算语言学协会。Yuxiang Zheng、Dayuan Fu、Xiangkun Hu、Xiaojie Cai、Lyumanshan Ye、Pengrui Lu 和 Pengfei Liu。2025b。DeepResearcher：通过在真实世界环境中进行强化学习来扩展深度研究。载于*2025* 年自然语言处理经验方法会议论文集，第 414–431 页，中国苏州。计算语言学协会。

QiangZhu、Jiajun Wu 和Longyi Wang。2026。Lotapo：用于多轮搜索推理中自生成过程奖励的留一轮归因。预印本，arXiv:2607.13501。

A 从莫比乌斯项到夏普利值

莫比乌斯交互项和沙普利值是对同一组 16 项组合测量的两种互补分解。对于每个子 $S \subseteq N$ 集，CO-GRAFT 测量 $v(S)$，即经强化学习训练的智能体拥有 S 中的角色、而强化学习前初始化拥有其余角色时所获得的精确匹配分数 $N \setminus S$。莫比乌斯反演将总增益划分为 15 个分量，每个非空子集 $T \subseteq N$对应一个： $m(T)$ 在扣除由 T 的子集所解释的所有低阶效应后，捕捉到只有当强化学习智能体同时拥有 T 中的所有角色时才会产生的边际协同效应（式2）。因此，任意组合 S 的增益等于其中包含的莫比乌斯项之和：

$$v(S) - v(\emptyset) = \sum_{T \subseteq S, T \neq \emptyset} m(T), \qquad (3)$$

对于完整组合 $S = N$，全部 15 个项之和即为总强化学习增益 $v(N) - v(\emptyset)$。

夏普利值通过将每个交互分量 $m(T)$ 平均分配给参与其中的角色，为每个角色赋予其应得的份额：

$$\phi_i = \sum_{T \ni i} \frac{m(T)}{|T|}. \qquad (4)$$

对于中间推理（R）， ϕ_{R} 等于独立项 $m(\{\mathrm{R}\})$，加上包含 R 的三个两两项各自的一半、包含 R 的三个三方项各自的三分之一，以及四方项的四分之一。因此，莫比乌斯项可以识别哪些角色（单独或联合）带来性能增益，而沙普利值则将每个角色的整体贡献概括为一个标量，同时不保留高阶耦合结构。

表4 展示了式4在自然问题上两个截然不同的检查点中的表现。对于 Qwen2.5-3B 基础版，高阶交互项可以忽略不计：单独推理达到 $+18.8$ 个 EM 分数，其夏普利值为20。6，这表明其归因与独立推理改进紧密对应。对于 Llama-3.2-3B 基础模型，情况正好相反：单独推理仅提升 $+0.1$ 分，但其夏普利值达到10。3，其中5。7分来自两两交互（主要是 R×D在 $+8.8$ 中的交互），另有4。5 分来自三方和四方项。由于只有当强化学习策略同时控制决策时才会产生这种归因，因此不能脱离 ϕ_{R} 来独立解释ϕ_{D}。

仅角色	单独	对	组三项	四项	Shapley
Qwen2.5-3B 基础版，自然问题（增益 29.4）					
R	+18.8	−0.7	+2.9	−0.4	+20.6
D	+0.6	−1.3	+3.0	−0.4	+1.9
Q	+2.5	−2.1	+1.8	−0.4	+1.8
A	+5.7	−3.3	+3.0	−0.4	+5.0
Llama-3.2-3B 基础模型，自然问题（增益 31.6）					
R	+0.1	+5.7	+3.8	+0.7	+10.3
D	−2.5	+3.6	+4.0	+0.7	+5.8
Q	+2.4	−1.2	+2.8	+0.7	+4.8
A	+7.0	+1.3	+1.8	+0.7	+10.8

表4：根据表2 的莫比乌斯项重构的夏普利值（EM 分数）。单独：单角色项；接下来三列分别是包含该角色的双角色项之和的一半、三角色项的三分之一，以及四角色项的四分之一。四部分之和等于夏普利值，存在舍入误差。

B 交互矩阵

图 2 展示了表 2 中各智能体和数据集的独立莫比乌斯项与两两莫比乌斯项。对角线元素显示单角色增益，非对角线元素显示两两交互，右侧条形图显示夏普利值，而 H 汇总了二维网格中未显示的三角色项和四角色项。由此可见三种结构模式：(1) 对于 Qwen2.5-3B 基础版，独立推理对角线项（在自然问题上为 $+19$ ，在 HotpotQA 上为 $+21$ ）远大于其他所有元素，而每个两两交互项均在 ± 3 点以内，证实角色贡献总体上具有较强的加和性。(2) 对于自然问题上的 Qwen2.5-3B-指令微调，四个对角线项均为正值（从 $+4$ 到 $+14$），而全部六个成对项均为负值（从 -2 到 -9），表明存在显著的跨角色冗余。(3) 对于 Llama-3.2-3B 基础模型，推理和决策的独立对角线项为零或负值（自然问题上分别为 0 和 -3；HotpotQA 上分别为 0 和 -1），而 R×D 在两个数据集上都是最大的两两交互项（分别为 $+9$ 和 $+5$），这表明推理与搜索终止之间存在很强的耦合。最后， H 在自然问题上的 Qwen2.5-3B-指令微调（$+13.8$）和 Llama-3.2-3B 基础模型（$+15.3$）中都相当显著，因此需要结合表 2 的高阶列共同考察交互矩阵。

C 示例轨迹

我们通过定性案例研究展示三种不同的归因模式：由中间推理（R）、最终答案提取（A）以及搜索或回答决策（D）驱动的增益。对于每个案例，我们选取一个评估实例，其中强化学习前初始化失败，而经强化学习训练的智能体成功，

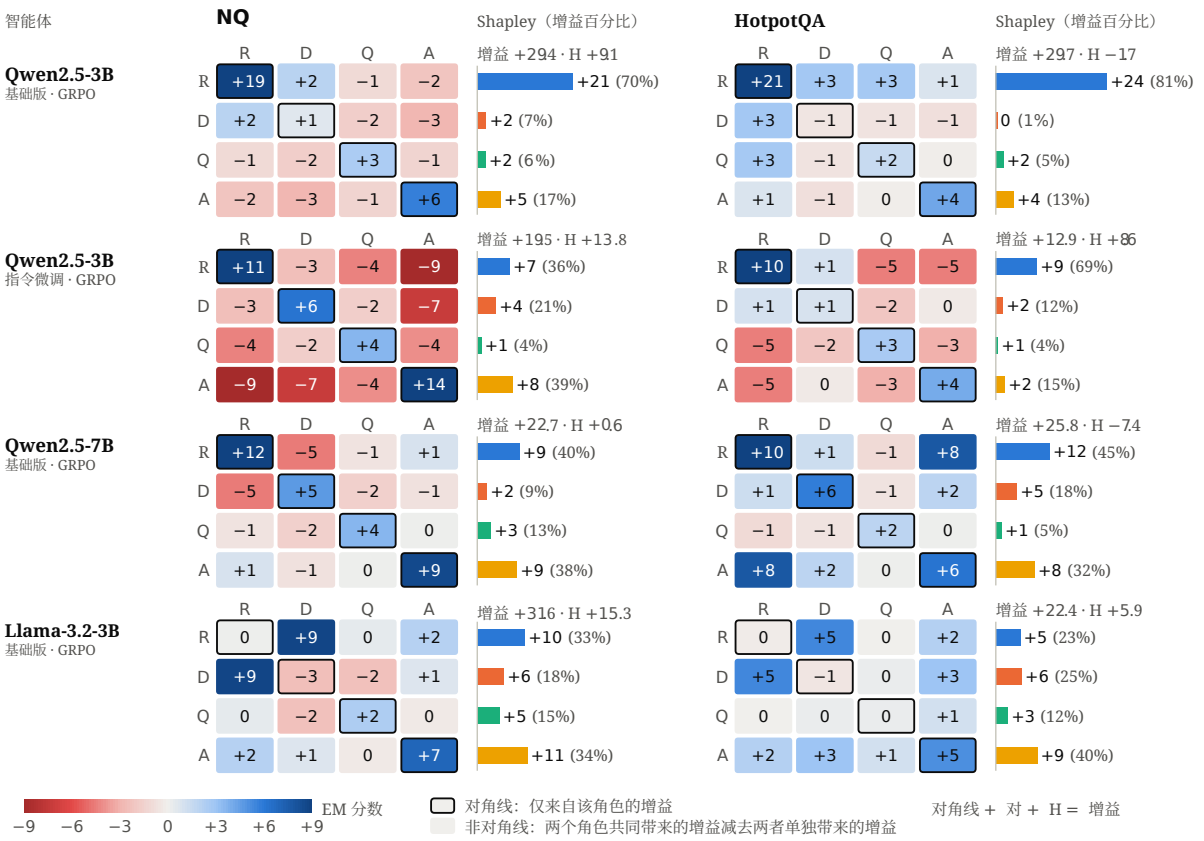


图 2：表 2 中四个经 GRPO 训练的智能体在自然问题（左）和 HotpotQA（右）上的交互矩阵，以精确匹配分数衡量。带边框的对角线条目表示当强化学习智能体仅控制单个角色时的单独角色增益。非对角线条目表示成对的莫比乌斯项（两个角色共同带来的增益减去单独角色增益之和）。右侧的条形图报告每个角色的夏普利值（以精确匹配分数表示，并显示其占总增益的百分比）。 H 表示所有三个角色和四个角色莫比乌斯项的未加权总和，而对角线条目、上三角成对项和 H之和等于总强化学习增益。热图中的颜色深浅显示在条形图中，颜色饱和度上限为 ± 9 分。

且在 16 种角色分配下的表现与对目标角色的控制相对应。图 3–5 展示完整轨迹（为提高可读性，省略检索到的段落文本并合并换行），并以颜色标示各生成片段所对应的当前角色策略。



图3：推理主导的增益：Qwen2.5-3B基础版（GRPO）在自然问题上的表现。推理角色获得了该智能体在自然问题上70%的增益（表 1）。两个模型接收相同的三个段落，其中包含标准答案。在强化学习智能体撰写 R 的八种角色分配中，该问题的精确匹配得分均为1；在另外八种分配中均为0。



图4：答案引导的增益：Qwen2.5-3B 基础版（GRPO）在自然问题上的表现。答案角色获得了该智能体在自然问题上17%的增益（表 1）。两个模型接收相同的三个段落，其中包含标准答案。在强化学习智能体写下 A 的全部八种角色分配中，该问题的精确匹配均为1；在另外八种分配中均为0，并且强化学习前模型写下的答案中没有任何一个包含标准答案。



图5：决策驱动的增益：Qwen2.5-7B 基础版（近端策略优化）在 HotpotQA 上的表现。决策角色获得了该智能体在 HotpotQA 上增益的 50%（§5.2）。在强化学习智能体编写 D 的八种角色分配中，有四种情况下该问题的精确匹配为 1；而在强化学习前模型编写该内容的全部八种角色分配中，精确匹配均为 0。