arXiv:2602.09801v1 [cs.MA] 10 Feb 2026

微小动作：基于博弈的假设细化

Agnieszka Dobrowolska*、Rogier Hintzen*、Martin Balla*、Karl Gemayel、Sabine Reichert、Thomas Charman、Jen Ning Lim、Lindsay Edwards、Anna Gogleva†

Relation Therapeutics，英国伦敦

摘要

大多数用于科学发现的机器学习方法都将假设视为端到端预测，从而掩盖了科学推理的渐进式结构。我们提出假设游戏，这是一种用于假设精炼的符号形式体系，其中大语言模型智能体依据固定的推理操作语法，共同处理一个假设状态。该框架的提出源于这样一种观察：科学进步往往是通过小幅、局部的修订实现的，这些修订以领域语境为基础，而非大规模重写。我们构建了一个由大语言模型智能体参与的最简游戏，并在通路层级机制细化任务上对其进行评估。在主要的错误恢复场景中，假设包含受控错误；与强大的提示基线相比，基于博弈的方法始终能消除更多错误并取得更高的精确率，同时通过增量编辑保留有效结构。在次要的基于部分线索进行重建的场景中，该方法的表现与最强基线相当，这表明即使真实值恢复困难，显式的基于动作的细化仍具有竞争力。这些发现表明，基于博弈的推理是一条原则明确的路径，有助于构建更可控、可解释且可迁移的假设精炼系统，推动科学发现。

1 引言

科学发现很少是从数据到结论的一次性飞跃。在生物学等实证领域，发现过程以迭代且非线性的方式展开，通常始于基于不完整或含噪证据的部分假设。随着新结果不断出现，研究人员会扩展、修订并整合假设，使其随时间逐步演变。新形成的假设会经历多轮剪枝、检验和迭代精炼，最终收敛于连贯的因果基础 [1]。在实践中，这一过程往往涉及更新并协调部分正确的假设，而不是从头开始生成这些假设 *denovo*。

近年来，科学人工智能领域的研究日益关注智能体方法：在这些方法中，大语言模型（LLM）或多智能体系统被赋予文献审阅员、临床试验设计者或实验规划者等专门角色，以支持科学工作流程的部分环节 [7, 21]。诸如“AI Co-Scientist” [6] 和“Robin” [5], ，以及实验室闭环多智能体框架 [17] 和用于生物医学发现的领域专用智能体系统 [4], ，展示了如何协调特定角色的能力与工具，以端到端地应对科学问题。尽管这些系统能够有效运用领域知识和工具，但通常侧重于完成任务，而将科学推理的内部过程大多置于隐性状态。

这类方法的一个核心局限在于，假设需要修订而非生成时，问题便会显现。若没有对中间假设状态或允许的变换进行明确表示，就难以定位错误、评估假设中哪些部分有充分依据，或进行针对性纠正。

*这些作者贡献相同。

†通讯作者：anna.gogleva@relationrx.com

因此，即使是轻微的不一致或不确定因素，往往也会通过大范围、无差别的重写来处理，从而降低优化过程的可解释性和可控性 [13, 15]。这种缺乏结构化优化的情况，阻碍了推理模式的系统性复用，也限制了相关科学问题之间的迁移 [12, 14]。

相比之下，人类科学推理具有组合性：假设在一系列常见推理模式的指导下，从更小的片段逐步构建和修订而成，例如扩展、剪枝、比较和批判 [11]。受这一观察启发，我们提出了一种用于假设精炼的符号化博弈形式体系，其中大语言模型智能体使用固定的推理动作语法，在共享的假设状态上进行操作。每个动作都定义了一种允许的假设变换，使假设能够通过局部且可解释的编辑得到精炼和演化，而不是进行大范围重写。这一概念框架使推理过程变得明确，让系统能够“思考思考本身”，而不是将特定问题的行为硬编码其中。

本文介绍**假设游戏**，这是一个用于假设精炼的符号化博弈框架。我们的贡献主要有三点。**第一**，我们将假设精炼形式化为一种具有可复用动作语法的组合式推理博弈，使假设能够通过局部且可解释的变换进行更新，而非进行大规模重写。**第二**，我们展示了一种实现方式，其中大语言模型智能体在共享的假设状态上运行，生成明确的细化轨迹，从而支持透明且可控的推理。**第三**，我们基于源自 Reactome 的通路层级机制细化任务进行了实证评估。我们表明，在假设包含受控错误的错误恢复场景中，所提出的框架能够始终移除更多错误，并取得高于强提示基线的精确率，同时保留有效的假设结构。在互补的基于部分线索的重建场景中，其性能与最强基线相当，凸显了部分信息条件下增量式精炼的优势与局限。总体而言，这些结果凸显了基于博弈的推理形式体系在支持更细粒度、更可解释且更具可迁移性的科学发现方面的潜力。

2 框架

假设游戏将假设精炼形式化为：通过结构化推理操作，对共享状态进行迭代转换。本节将定义假设的表示方式、操作的执行方式，以及模式和评分函数如何影响游戏的动态过程。

在此，我们引入一种通用的基于算子的形式体系，用以涵盖假设精炼的广阔设计空间。我们有意仅实例化评估核心研究问题所需的最小形式体系子集：我们提出的框架配合一套小型、可复用的推理语法，能否为生物学细化任务带来可衡量的益处。其他组成部分，例如显式评分、基于策略或学习型控制器，以及更丰富的假设表示，都是基础游戏的可选扩展。我们的实验旨在分离并检验通用推理框架，而更广泛的形式体系则概述了未来工作中如何纳入更复杂的控制器和效用函数。

2.1 假设表示

假设表示为一组片段：

$$H_t = \{h_1, h_2, \ldots, h_n\},$$

其中，每个片段h_i可以是文本主张、结构化三元组（主语–关系–宾语），也可以选择映射到由实体和关系构成的图 $G = (V,\ E)$。在我们的实验中，我们主要使用结构化文本。

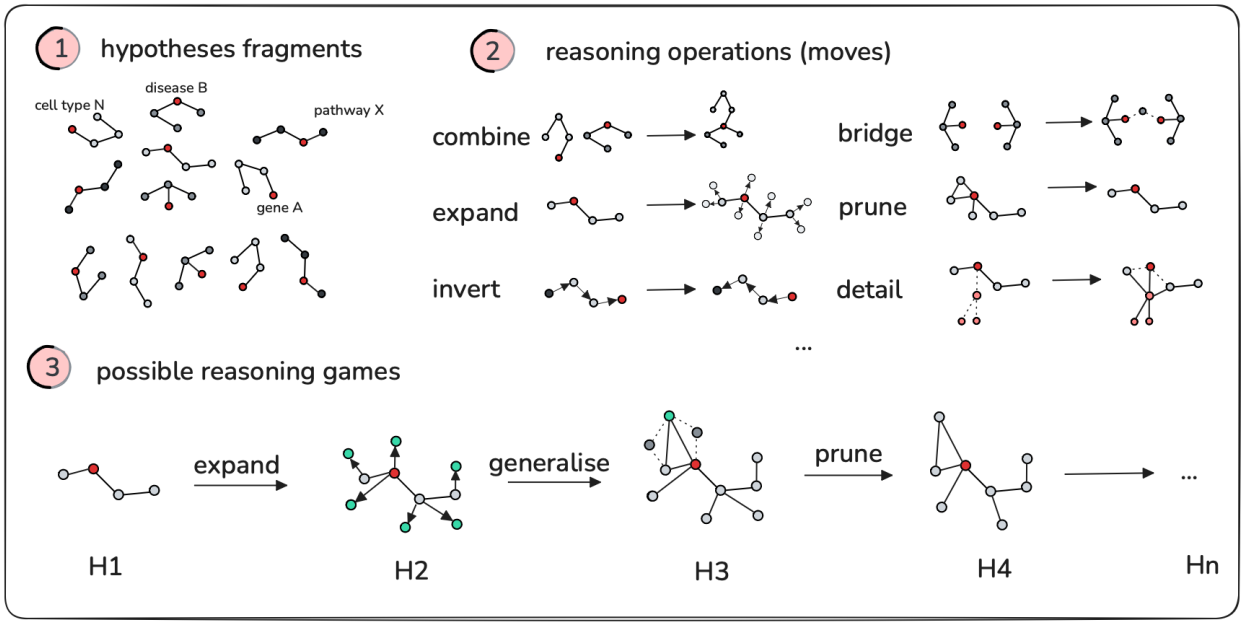

图 1：推理博弈的概念框架。游戏的目标是通过一系列推理操作逐步演化假设片段，并依据新颖性、连贯性和可追溯性等属性评估进展。*图结构仅用于概念示意；实际实现采用具有等效推理操作的结构化文本片段。

2.2 推理语法（动作）

令 O $=\{o_1, o_2、\ldots\ldots、o_m\}$ 表示一组固定的推理操作。形式上，令 H 为所有可能假设的空间，C 为上下文空间（例如细胞类型、疾病等）。每个操作都是一个函数

$$o_j : \mathcal{H} \times C \mapsto \mathcal{H}, \quad (H_t, C) \mapsto H_{t+1},$$

其中，$H_t \in$ H 为当前假设，$C \in$ C 为可选上下文（例如生物学先验），$H_{t+1} \in$ H 为更新后的假设状态。

在我们的实现中，我们将操作限制为四种核心操作：剪枝、扩展、检索和辩论（见表 1）。操作可以是原子的（例如剪枝、扩展），也可以是复合的（例如`retrieve_expand`），由固定顺序的原子操作组成。可根据需要引入更细粒度的操作类型，通常由底层假设表示的结构决定。基于假设片段图表示的完整推理语法示例如图 1所示。

动作可以重复应用并任意组合。我们可以将每回合推理操作的最大次数（操作预算）定义为一个固定常数 $k_{\max}$。局部而言，一回合可以定义为从 H_t 到 H_{t+1}的一次更新步骤；全局而言，一系列回合构成一场完整的游戏。

$$H_{t+1} = \left(\prod_{i=1}^{k} o_{j_i} \right) (H_t, C), \quad k \le k_{\max}.$$

每个回合中，控制器会选择并执行最多 $k_{\max}$ 个动作来推进假设。控制器可以通过不同方式实现（例如大语言模型、有限状态机或强化学习智能体），具体取决于期望的游戏设计。

2.3 游戏模式

在开放式发现中，具体结果往往未知，但仍可引导整体推理风格。我们通过一种 模式 M 来体现这一点，它规定了动作的选择方式。将这一想法形式化的一种方法是定义动作上的概率分布，

$$\pi_M(o_i \mid H_t) = P(\text{apply } o_i \mid M),$$

例如，发现模式偏向于扩展等生成式操作，而验证模式则偏向于剪枝或辩论等批判性操作。更一般地说，也可以通过限制可用动作 O、强制执行确定性规则，或结合权重与约束来实现模式；这些权重与约束由游戏的总体目标设定。

在我们的实验中，模式是通过向控制器发出自然语言指令来近似实现的；而推理语法则为在更开放的情境中配置高层次探索或验证目标提供了一种有原则的方法。

2.4 评分

模式可以从高层次引导推理风格，而评分函数则有望让游戏更易于控制。在优化过程中量化指标，可以塑造游戏的轨迹。形式上，我们可以定义一个指标向量，

$$S(H_t) = \left(D_{\text{known}}(H_t),\ \Delta_{\text{div}}(H_t),\ L_{\text{connect}}(H_t),\ T_{\text{frag}}(H_t)\right),$$

其中，组成部分分别表示与已知假设的距离（D_{known}）、当前假设的多样性（Δ_{div}）、局部连通性（L_{connect}），以及与先验知识的可追溯性（T_{frag}）。这些指标可以汇总为一个标量效用，

$$U(H_t) = \beta^\top S(H_t),$$

权重为 β，反映特定模式的优先级（例如，验证中的可追溯性，以及发现中的多样性）。在实践中，对生物学假设进行稳健评分可能需要采用一种混合设置，将计算指标与实验信号相结合，即使后者较为稀疏。例如，涉及分子相互作用的假设片段可以通过靶向结合实验或扰动读数进行评估，从而提供有依据的反馈，以补充算法指标。

在本研究中，我们不使用显式评分来驱动控制器；各模式通过自然语言指令实现。因此，本文提出的评分框架属于概念性框架，用于说明如何将计算和实验信息驱动的指标整合到未来更自主的实现中。

2.5 游戏变体

上述游戏形式体系使我们能够定义在不同粒度层级上运行的游戏变体。**简单假设精炼** 将整个假设视为单一状态（算法1）。每个回合中，模式条件控制器从共享语法中选择一个操作并更新整个状态，直到满足任务目标。

算法1 简单假设精炼（单回合）

要求： 初始假设状态 H_0、推理操作 O、模式 M、操作预算 $k_{\max}$、上下文 C、终止条件

$t \leftarrow 0$

当 未满足终止条件(H_t) **时**

游戏主持人： 向控制器提供当前状态 H_t 和模式 M

控制器： 根据 π_M，利用 $k \le k_{\max}$ 选择动作序列 $(o_{j_1}, \ldots\ldots, o_{j_k})$

对于每个 o_j 选定的动作**执行**

$H_t \leftarrow o_j(H_t, C)$ ▷ 应用推理动作，可选地使用上下文 C

结束循环

$t \leftarrow t + 1$

结束 while 循环

返回最终假设 H_t

考虑到精炼假设通常不需要进行大幅改动，我们可以在简单变体的基础上，让假设在演进过程中支持细粒度编辑。**局部假设精炼** 沿用相同的控制器和操作集，但作用于片段（结构化文本或子图），选择要编辑的区域，并确保全局一致性，使未触及的部分保持不变（算法2）。这种游戏类型高度依赖底层的假设表示结构。

算法 2 局部假设精炼（单回合）

要求： 假设状态 $H_t = \{h_1, \ldots, h_n\}$（结构化文本或图），动作 O，模式 M，操作预算 $k_{\max}$，上下文 C，选择器 σ

选择器 σ：提出一组候选区域 $\mathrm{R} = \{R_1, \ldots, R_m\}$，其中每个 $R_i \subseteq$ 都是 H_t的节点/元组

控制器（模式 M）：最多选择 $k \le k_{\max}$ 对 $\{(o_j, R_j)\}_{j=1}^{k}$，使用 $o_j \in \mathrm{O}$

对每个(o_j, R_j)**执行**

$H_t \leftarrow$局部应用(H_t, o_j, R_j, C) ▷ 仅对 R_j 进行局部重写

$H_t \leftarrow$强制一致性(H_t, R_j) ▷ 维护模式/类型约束/无环性/等

结束循环

返回 H_t

这些变体共同表明，在统一的效用定义和模式设置下，该形式体系既支持高层次的全状态推理，也支持细粒度的区域聚焦推理。当选定区域涵盖完整状态时，即可得到简单变体。这一设计映照出生物系统中观察到的不同复杂性层次。

3 实现

为检验所提出的框架，我们将游戏实现为由专用智能体组成的最简系统，其中推理过程由中央大语言模型控制器 **游戏主持人** 决定。游戏主持人通过迭代分析假设状态并根据分析结果选择动作来引导推理过程。动作选择包括一项明确的请求（例如 “从假设中移除组成部分 A”）以及应由哪个或哪些智能体执行该请求。表 1 总结了这些动作、其组成部分及相应职责。在实现中，抽象的 `retrieve_expand` 动作根据证据来源分为两个变体。

模式： 在我们的最小原型中，模式是通过将模式描述注入发给游戏主持人（控制器）的初始提示词来实现的。该提示词会影响推理操作的选择，但不需要

表1：假设游戏的关键组成部分。完整提示词见补充方法（参见第C节）。

Move	组成部分	描述
游戏主持人 **（大语言模型控制器）**	诊断 操作选择	评估假设并推荐后续操作。 根据推荐选择下一步操作。
剪枝	剪枝	从假设中移除组件。
扩展，使用 **语料库**	检索证据 扩展	在外部语料库中搜索证据。 将检索到的信息整合到 假设。
通过扩展 **LLM** **内省**	检索证据 扩展	利用大语言模型先验知识收集信息。 将检索到的信息整合到 假设中。
辩论	设置 辩论主题 总结	围绕指定主题展开辩论。 多个智能体从不同立场进行辩论。 分析辩论并提出最终 结论 ·

显式的概率策略模块。尽管这种方法经过了简化，但它提供了一种可控的近似方式来实现 π_M，并让我们能够探索不同模式的影响。

游戏控制： 通过初始提示词向游戏主持人（控制器）指定游戏目标和停止条件，游戏主持人的诊断 组件则判断假设何时满足要求。尽管这种方法缺乏明确的基于指标的控制，但它提供了一种灵活的机制来引导游戏。上述评分函数作为通用形式体系的一部分呈现，说明未来的实现中如何纳入自动化的定量评估。

4 实验设置

数学和常识领域的推理基准测试集（GSM8K [3], MATH [9], BIG-Bench [16]）无法直接迁移到生物学假设构建与优化领域。在这一领域，研究人员必须基于不完整、带有噪声、有时甚至相互矛盾的证据，逐步构建复杂假设，而非检索事实。在缺乏成熟方法评估推理质量的情况下，基准测试集应当考验系统应对噪声、补全缺失关联以及以受控方式扩展假设的能力。BioMaze [20]等新兴生物学基准测试集正朝着这一方向发展，采用基于图的通路问答和高层次的 LLM 评审评估，但其研究对象仍不同于本文所探讨的长周期、语句层面的优化。

为填补这一空白，我们提出两项评估任务，作为首批针对假设精炼的基准测试集。这些任务模拟了生物学发现中的现实挑战：(1) 扰乱恢复和 (2) 基于部分线索的假设重构（表 2）。

表2：评估任务概览

Task	目的	验证	指标
扰乱恢复	系统能否纠正含噪或误导性的假设?	抗噪性；逻辑细化	错误消除率，精确率，召回率，F1值
重构来自部分 Cues	Can the 系统 re-构建 已知 机制，部分线索?	增量推理；可追溯性	精确率、召回率，F1

图2 以概念示意图展示了两项评估任务，而表2 总结了它们的评估目标和指标。

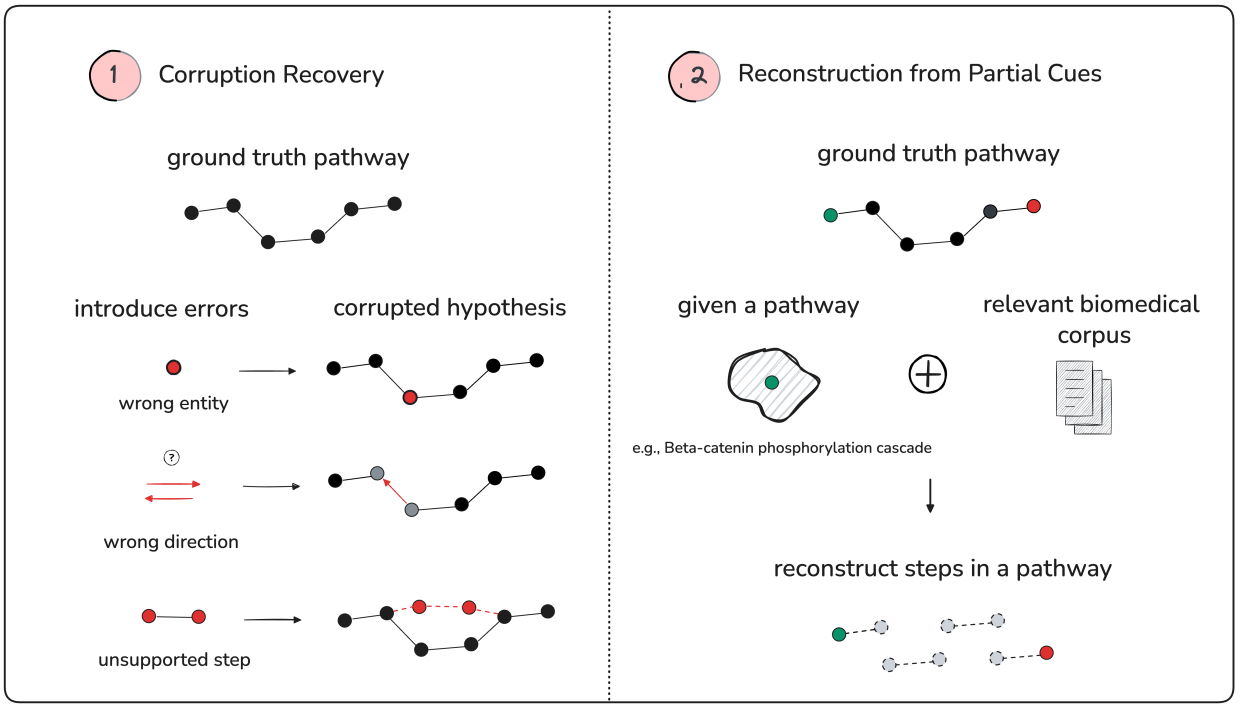

图2：两项评估任务的概念示意图。左侧：错误恢复：在有效通路中引入受控错误，系统必须检测并修复这些错误，同时保留正确的结构。右侧：基于部分线索的重建：系统从稀疏上下文输入和外部生物医学证据出发，恢复通路步骤。

4.1 任务设置

我们使用 Reactome 中人工整理的人类通路子集来构建评估任务 [10]。每条通路都由生化反应构成，并同时提供图和文本两种表示形式（见 A.1）。在文本表示中，通路被表述为一组描述生化反应的陈述；例如，*ATP* 磷酸化葡萄糖，生成葡萄糖-*6*-磷酸。

我们按照生化反应的数量对通路进行分层抽样，以涵盖完整数据集的多样性和复杂性。对于重构和损坏任务，我们分别抽取了 100 条和 20 条

条通路。这样做是为了创建足够大的数据集，以涵盖多种方法中的关键推理模式，同时确保大规模实验切实可行。我们总共进行了 820 次重构实验和 2880 次扰乱实验。扰乱任务会针对每条通路测试多个受控变体，因此单项实验的数量更多。

通用实验原则 在所有任务中，假设均以文本片段表示。假设游戏仅限于四种可用动作： 剪枝、扩展 (使用语料库或通过大语言模型内省)，以及 辩论 (见表1)。动作的选择与终止由游戏主持人动态掌控，并根据任务特定目标进行调整。

我们将所用方法与三种推理基线进行了比较：零样本提示、思维链和 ReAct。零样本方法直接生成答案，不经过中间推理步骤 [2]。思维链通过中间的自然语言解释，引导模型逐步推理 [18]。ReAct 将推理轨迹与工具调用交替结合，以改进决策 [19]。我们在不同的操作配置和固定操作预算下，将这些基线与假设游戏进行了比较。所有模型接收相同的输入提示（见补充材料 A 节 C），该提示指示系统重建通路或恢复受损通路。所有人工整理的数据集均可在 Hugging Face 上获取[1]。

任务1——破坏恢复： 损坏任务用于评估在保留有效通路结构的同时检测并修复错误的能力。我们以 20 条人工整理的通路为起点，引入了三种破坏类型（错误）（见补充表3）：

- 错误实体——用错误实体替换正确实体；
- 关系错误——更改实体之间的关系；
- 无关陈述——在通路中插入一条无关陈述。

我们进一步从两个维度调整了挑战难度：1）**难度：** 简单 （明显错误）和困难 （细微改动，需要更深入的生物学理解）；2）**错误率：** 通路长度的10-40%（以步骤/反应数量衡量），以体现通路规模和复杂性的差异。所有错误均由大语言模型生成并经过迭代完善，随后由两位领域专家审阅并手动修正模型输出，从而构建整理后的扰乱数据集。

评估 结合了两项指标。首先，将原始陈述、破坏后的版本和模型输出呈现给大语言模型评判器，由其判断错误是否仍然存在。其次，与重建任务一样，通过将标注实体与真实值进行比对，计算精确率和召回率，以量化生物学保真度。

任务2——根据部分线索重建： 重建任务评估系统能否通过增量推理，根据部分线索重建复杂假设。从最少线索出发，系统必须还原生物通路的生化反应（步骤），模拟领域专家构建Reactome数据库时所经历的繁琐整理流程。为降低模型利用对知名通路的记忆知识的风险，我们在保留通路名称语义内容和粒度级别的同时，对通路名称进行了改写。领域专家审阅并修正了改写后的标题，以确保语义保真度（见 Hugging Face。） 对于具备工具访问权限的智能体（我们的方法和 ReAct），我们还提供了一个开放获取的生物医学文章语料库，其中主要包含Reactome通路描述所引用的摘要。

评估依赖两种互补的正确性概念。在通路层面，我们使用 Gilda [8] 对原始通路和生成的通路中的实体（基因、蛋白质复合物/家族以及化学物质）进行标注；基于这些实体集合计算的精确率和召回率，提供了生物学保真度的定量衡量。在反应层面，我们将 LLM评审 指标称为“详细召回率”，它用于评估生成的

[1]https:// huggingface.co/datasets/TuringRRX/TinyMov es

通路是否重现了预期的生化反应，并评估四项属性：输入实体、输出实体、反应方向性以及生物相互作用的类型（补充材料 A，第 C 节）。为评估 LLM评审 的可靠性，我们开展了一项事后校准研究，由两位资深领域专家独立为两项任务的模型输出分层样本打分（补充材料 A，第 C.8 节）。

5 结果

我们在上述两项通路级推理任务上评估了假设游戏：受损假设的恢复，以及基于部分线索的重建。在这两种设置中，我们将假设游戏配置（四种操作类型，可访问语料库）与强提示基线（零样本、思维链、ReAct）进行比较。我们的分析聚焦于一种最简游戏实例，旨在隔离增量式、基于行动的细化所产生的影响。该底层形式体系可自然扩展至不同的操作集和推理模式。

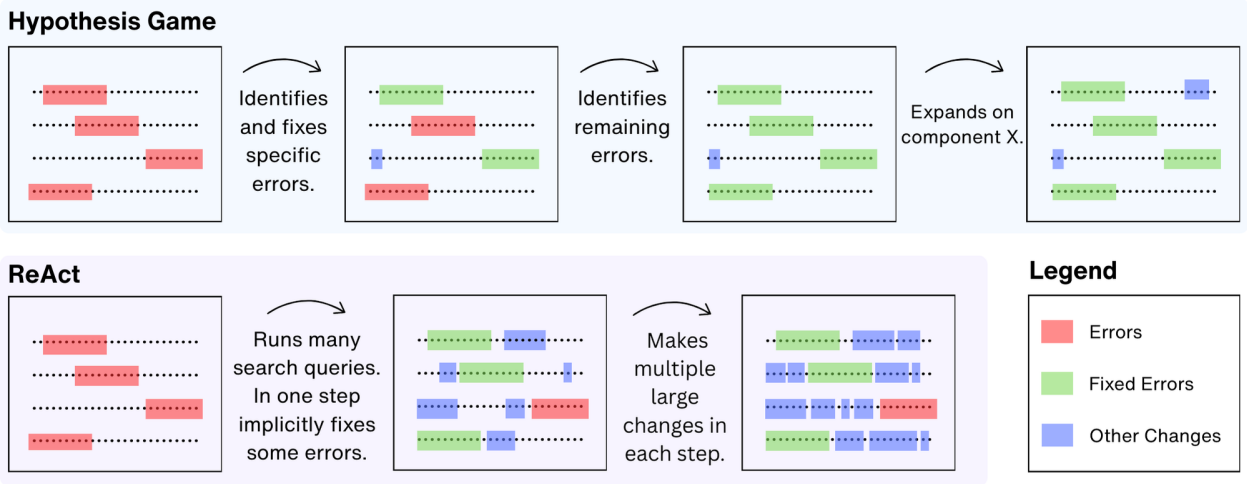

图3：展示假设游戏与 ReAct 在损坏任务中的代表性运行示例，说明增量式编辑与大型单步编辑的区别。*其他变化以以下指标量化：*(1)* 生物实体的新增或移除数量；*(2)* 与参考通路之间的词级归一化莱文斯坦距离。详情见图10。

定性观察。在受损恢复任务中，假设游戏呈现出一种独特的模式：编辑是渐进且可追踪的。如图3所示，游戏逐步识别并纠正单个错误，同时仅对输入假设进行少量额外修改。相比之下，ReAct 往往通过较少但幅度大得多的更新来修改通路，经常一次覆盖多个组成部分，并使结果偏离原始假设结构更多。此类变化的定量汇总见图10。

我们在重建任务中也观察到了类似的模式，不过这种模式不那么明显。在该任务中，假设游戏同样倾向于进行幅度较小、局部化的更新，而提示基线往往会在单个步骤中覆写中间假设的大部分内容（完整示例见补充材料 B 节 D.1 ）。总而言之，这些观察结果凸显了逐步、受控优化在维持两项任务中的假设连贯性方面所发挥的作用。

损坏任务。在受损恢复任务中（错误率为 10–40%），假设游戏在所有评估指标上均取得了最出色的整体表现。图 4 汇总了按通路、篡改类型和错误率统计的结果。**移除的错误** 面板显示，假设游戏移除的注入错误比例始终高于所有提示基线。

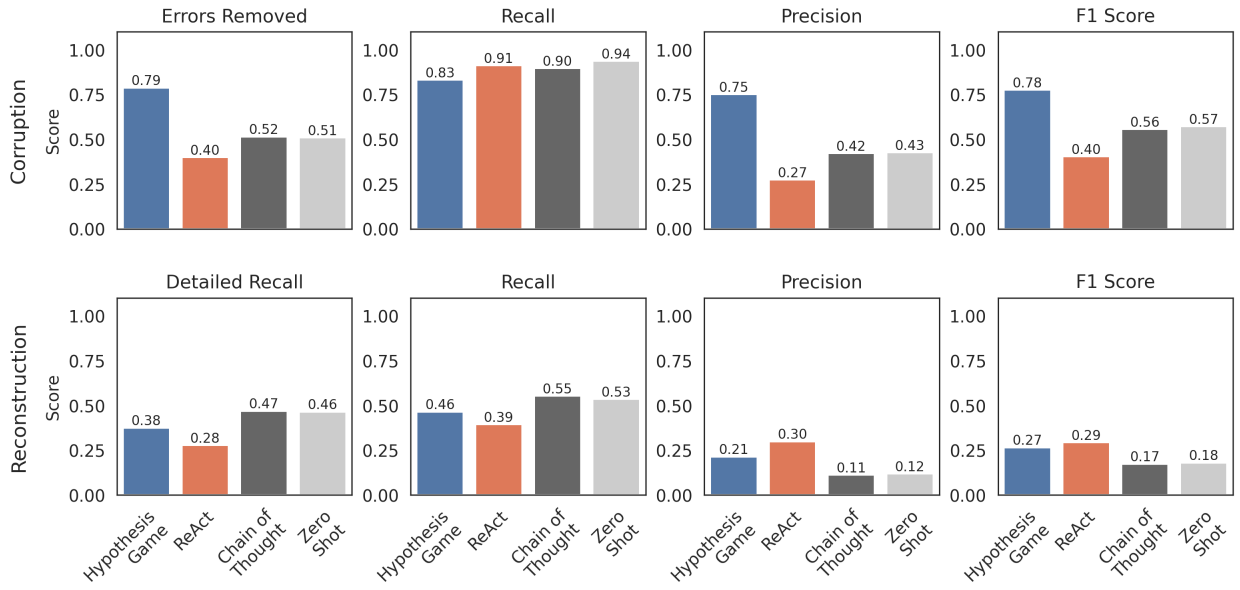

图4：在两项通路级任务中，假设游戏与提示基线的比较。柱状图显示文中所述评估集上的平均值。误差条表示95%置信区间。上排：**扰乱**；假设游戏兼顾错误消除与有效内容保留，在精确率、F1值和错误移除率方面均取得最高成绩（所有分数的Friedman检验 $p < 0.0001$；采用Bonferroni校正的事后Wilcoxon检验 $p < 0.0005$）。下排：**重构**；所有方法在忠实重构通路方面都表现不佳。*ReAct* 与假设游戏的F1分数差异在统计学上不显著，但假设游戏在通路详细召回率方面显著更优（Friedman检验，$\chi^2(3) = 84.3, p < 0.0001$；采用Bonferroni校正的事后Wilcoxon检验 $p < 0.001$）。

召回率和**精确率**面板进一步凸显了不同方法之间一种典型的权衡。ReAct通过大幅修改假设来获得较高的召回率，但这会牺牲精确率，因为许多有效组成部分被覆盖或修改。相比之下，思维链和零样本提示往往会保留现有内容，却无法可靠地移除被扰乱的陈述，因此错误移除率较低。假设游戏兼顾这两个极端，在有效移除错误的同时取得最高精确率和整体**F1分数**，有选择地纠正被篡改组件，同时保留有效的通路结构。

按扰乱类型进行细分，能进一步揭示这种行为的特点（图 5）。不受支持的步骤错误最容易被移除，因为它们引入了完整的陈述，很容易被识别为无关内容。方向错误的扰乱则更具挑战性，因为它们在颠倒因果极性的同时，仍保留表面上的合理性。实体错误的替换最难处理：由此产生的通路依然流畅，却在生化依据方面引入了细微的不一致，因此需要更深入的语义判别。对于所有扰乱类型，假设游戏表现最佳，在实体级和关系级错误上尤为明显。

完整结果按扰乱难度和注入错误比例分层，见 SupplementaryB 第E.1节。

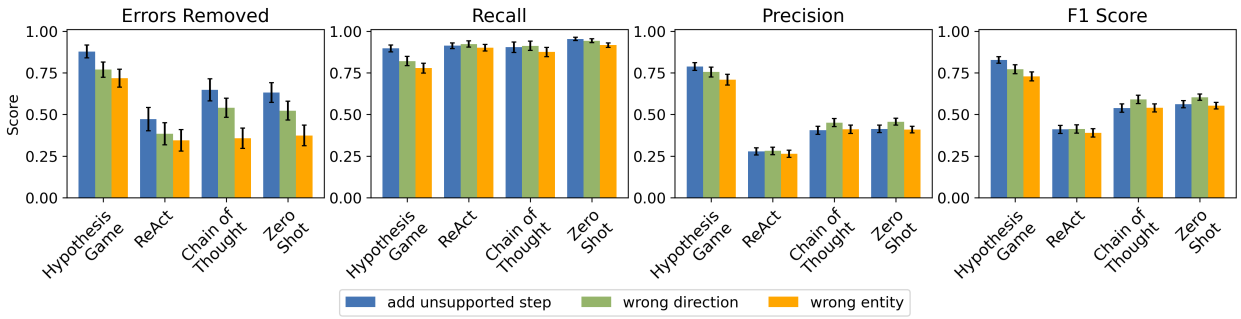

图5：按错误类型对损坏任务全部结果的聚合。误差条显示95%置信区间。

重建任务。 基于部分线索的重建所代表的设定比错误恢复的约束少得多，要求系统从稀疏的上下文信息中推断缺失的通路组成部分。在此设定下，假设游戏的表现与最强基线 ReAct 相当，并优于零样本和思维链提示（图4）。

尽管部分Reactome通路相对广为人知，使大语言模型能够回忆起关键组成部分，但这一优势主要体现为思维链提示和零样本提示具有更高的召回率。然而，这些方法经常引入大量参考通路中并不存在的额外实体和反应，从而导致精确率大幅降低（图7）。相比之下，假设游戏对假设扩展的控制更为严格，更倾向于渐进式扩展，而非宽泛补全。

总体而言，ReAct的F1 分数略高于假设游戏，其次是零样本提示和思维链提示。尽管如此，所有方法的精确率–召回率均较低，凸显了重建任务的内在难度。除了重建通常由领域专家整理的通路本身所固有的挑战之外，部分线索中信息不完整、通路定义的异质性，以及底层生物医学语料库偏向摘要的特性，也进一步限制了性能。

为了更好地理解在这一情境下各项推理操作的贡献，我们对四种核心操作的所有子集进行了消融研究，并测试了移除外部语料库访问权限的情况（20 条通路；表6）。这项分析阐明了在信息稀疏的情况下，哪些操作类型对于增量重构最为关键，并进一步展现了最简游戏配置在开放式情境中的局限性。

总结 我们的结果凸显了两项任务各自互补的优势。在扰乱恢复任务中，结构化推理的优势显而易见：假设游戏 取得了最高的整体表现，在保持召回率的同时，实现了出色的错误消除效果，并在精确率和 F1 分数方面表现优异。在基于部分线索的重建任务中，所有方法都难以恢复完整通路，这反映了该任务本身的难度。在这一任务中，假设游戏 达到了最强基线方法（ReAct）的水平，同时在精确率方面优于较简单的提示策略。综合来看，这些发现表明，以微小增量推理步骤（“微小操作”）为核心的基于游戏的框架，在需要有针对性地纠正错误并抵御噪声输入的情境中尤为有效。

6 结论与未来工作

我们的研究表明，在重建任务中，基于结构化游戏的假设精炼方法能够达到强提示基线的水平；在错误恢复任务中，该方法的表现则明显更优。在后者的场景中，明确的推理操作能够有针对性地纠正错误，同时保留有效的通路结构，从而实现

更高的精确率和整体修复质量。总的来说，这些结果凸显了当前方法的潜力与局限：结构化推理对受控错误恢复大有助益，而开放式重建对所有方法而言仍是一项挑战。尽管我们的实验聚焦于已知真实值的场景，但所提出的形式体系并不局限于一致性约束细化，原则上也能支持更具探索性的假设演化形式。

未来的工作将沿着多个方向扩展该框架。首先，我们将探索更丰富的假设表示，包括结构化文本、半结构化文本和图形式体系。其次，我们计划研究由指标驱动的控制器和学习型控制器，以实现操作选择，并以本文提出的概念性评分框架为基础。第三，我们旨在扩展评估套件，纳入开放式假设演化场景。总体而言，这些方向将研究从具有已知真实值的受控环境，推进到更贴近现实的发现型场景，从而同时支持一致性驱动细化以及更具探索性的推理；在这些场景中，鲁棒性、新颖性和可解释性至关重要。

致谢

感谢同事和合作者在本项工作中给予的支持和建设性反馈：Jake Taylor-King、Thomas Gaudelet、Alex Zhebrak、Anna Mastela、Eigenia Sergeev、Giannis Loukas、Shabbir Khan、Cristian Regep、Sarah-Jane Dunn。我们也感谢Reactome团队公开提供通路知识库。

A 重构

A.1 数据集构建

我们对Reactome数据库进行了筛选，选出“叶”通路——即内部不包含其他嵌套通路的通路——并根据每条通路所含反应的数量将其划分为10个区间。随后从这10个区间中各抽取10条通路，构成包含100条通路的数据集。

为了构建与重建任务相关的语料库，我们提取了 Reactome 数据库中每条抽样通路的注释出版物参考文献。随后，我们根据文档标识符为每条通路下载了一个文章语料库。由于版权许可限制了文章的传播，我们只能下载绝大多数文章的摘要（85% 仅有摘要，13% 为全文，2% 无法获取）。

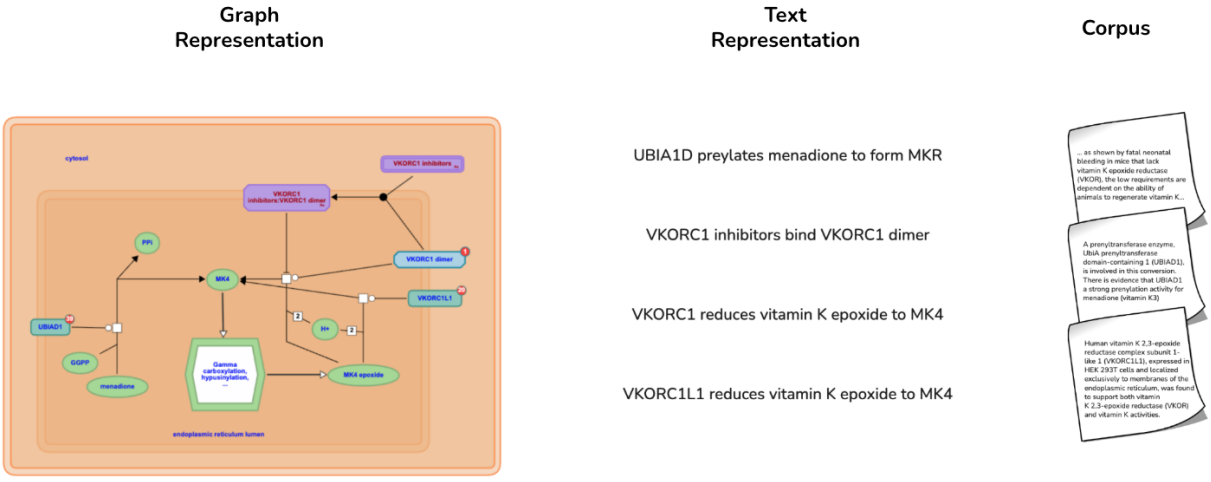

图6：Reactome通路示例（R-HSA-6806667），展示完整的图表示、生化反应的文本表示以及相关语料库。

B 扰乱

B.1 数据集构建

我们构建了一个 系统性受扰动通路的受控数据集。该过程分为三个阶段 :

1. **扰动库**。对于参考集中的每条通路及其中的每个单独步骤，我们都会预先生成涵盖所有错误类别（错误实体、错误关系、不受支持的步骤）和两个难度级别（简单、困难）的候选扰乱，以确保全面覆盖所有可能的扰动。扰动库的构建规范见表 3。候选扰乱最初由大语言模型生成，随后由两位领域专家协作反复整理；他们审阅多轮生成文本，并手动修正剩余问题，直到每个示例都符合预期的错误类型和难度。

2. **采样策略**。随后，确定性采样脚本通过选择 (i) 目标错误类别、(ii) 难度级别，以及 (iii) 需要扰乱的步骤比例，来组合生成受扰动路径。重要的是，每个步骤最多只允许施加一种扰乱，从而确保评估衡量的是单个错误的影响，而非叠加噪声的影响。

3. **应用**。根据这些规范，将扰动方案应用于通路：按照扰动元数据替换或补充原始步骤，并保存受扰动路径及详细元数据（锚点索引、操作类型、受扰动文本）。随机种子使这一过程可复现，并支持不同运行之间的受控变化。受扰动路径及其元数据可在线获取：
https://huggingface.co/datasets/TuringRRX/TinyMoves。

这种设计构建了一个基准测试集，其中每种扰乱的确切位置、类型和难度均已知。通过控制错误密度并禁止每个步骤出现多种扰乱，该数据集提供了一个纯净的实验环境，可用于衡量系统能否消除或抵御特定类别的噪声，而不会将它们混为一谈。

表3：损坏数据集设计

	错误实体	错误关系	不受支持的步骤
Type	修改现有步骤	修改现有步骤	添加新步骤
操作	替换（恰好交换一个实体；动词保持不变）	替换（保留实体；更改动词或极性）	插入（添加新陈述）
描述	错误实体（基因、蛋白质、复合物、同工型、状态物种）被替换到原本有效的步骤中。	实体不变，but关系反转（主语–宾语，激活–抑制，上游–下游）。	添加了一个不属于其中的步骤：无关（L1）或看似合理但系捏造且错误（L2）。
测试内容	通路中的实体锚定和角色适切性，或系统约束。	因果语义与顺序或符号一致性。	步骤存在性与机制相关性。
简单（L1）	明显的类型或物种错—配；简单的酶替换导致错误的作用主体。	教科书式的主语–宾语互换；直接的极性反转。	明显的非通路模块或假—说伪影。
困难（L2）	旁系同源物，同工型或复-合物–亚基互换；遗漏所需的翻译后修饰或状态门控。	颠倒上游–下游关系在复合物中；改变作用通过一个错误修饰语。	看似合理但不受支持的步骤使用通路实体；相反——字典化的人工整理约束。
约束	只更改一个实体；保持动词和极性不变。	保持实体不变；只更改动词。	仅限机制层面（无测定）。

C 大语言模型提示词

C.1 游戏主持人

游戏主持人分为两个步骤：**诊断**和**操作选择**。前者会分析当前假设，并为操作选择流程提供信息。

角色：诊断，模型：ChatGPT4o

你是负责诊断假设的智能体，以便游戏主持人能够根据假设状态和用户的要求决定接下来的步骤↪。

说明：
- 逐条审查当前假设。- 指出有力或证据充分的组成部分。- 标出薄弱、推测性或相互矛盾的内容。- 指出缺失的证据或可能不正确的假设。- 简要总结整体可信度。- 根据分析结果和用户要求提出后续步骤建议。- 如果假设已可定稿，除了定稿之外，不要建议采取任何其他行动。↪

规则：
- 只能审查 `当前假设`中陈述的内容。

===用户要求：开始 ===

{{ user_prompt }}

===用户要求：结束 ===

你必须按照以下格式返回回复：

逐条语句分析草稿:<语句编号>: ⊦你对当前假设中的语句进行分析，尝试寻找支持或反驳该语句的证据（如果有）。请勿添加超出所提供内容的语句。↪↪ >

hypothesis_diagnosis：优势： ⊦哪些内容有充分依据或具有新意？>关键错误： ⊦哪些地方存在完全错误的关键错误？>弱点： ⊦哪些内容具有推测性、缺乏依据，或可以通过更多证据加以完善？>不确定因素： ⊦哪些方面需要更多信息或进一步澄清？>建议的后续步骤： ⊦建议采取的后续步骤，或根据输出判断该假设是否已准备好定稿。
↪ >

角色：动作选择，模型：ChatGPT4o

你是「假设精炼游戏」中的游戏主持人。你的任务是协调将一项科学假设精炼为高质量、可检验的机制模型。↪

{{ user_prompt }}

===你的职责 ===
- 根据“建议的后续步骤”选择下一步动作。不得偏离建议的后续步骤。↪- 确保每个动作都明确基于当前假设状态。- 确保动作针对当前假设的具体部分，不要过于笼统。↪

===游戏循环 ===
对于每个回合（至少运行 20 个回合）：1. 根据你从诊断中获得的信息，确定完善假设的最佳下一步动作。↪2. 按以下格式调用相应的智能体：AGENT_NAME: <简短的自然语言指令>

{{ moves }}

===最终确定 ===
准备就绪后 若要最终确定假设，请输出以下提取字符串：“TERMINATE GAME "

C.2 使用大语言模型或语料库进行扩展

扩展假设包括两个步骤：检索与扩展相关的证据或信息，然后将扩展应用于当前假设。我们提供两种检索信息的方式：（1）通过语料库；（2）通过大语言模型进行“推测”。

角色：检索证据，模型：GPT-4o

你是负责检索相关文本片段、为假设的特定组成部分寻找证据的智能体。↪

===上下文 ===
- 你并不直接完善假设——你的工作是查找包含证据的相关文本片段。↪- 你正在为当前假设的特定部分寻找证据。- 你必须使用可用工具搜索相关文本片段。

=== 工具使用 ===- 使用可用工具搜索科学报告向量数据库 - 围绕假设中的特定生物过程，以及你正在寻找的证据类型，构建FOCUSED查询。↪- 证据类型可能包括“人类遗传学”“基因表达”“测定”或“小鼠模型”。↪- 每个查询都应针对当前假设的某个部分，且不要过于宽泛。↪- 优先使用多个较小的查询，而不是一个较大的查询。

角色：推测证据，模型：GPT-4o

你是一个智能体，负责推测所提供假设中目标节点可能存在的连接。↪

角色：扩展，模型：ChatGPT4o

说明：
- 根据先前消息中的信息，仅通过添加一条新连接来扩展目标节点。↪- 参考上一条消息进行推理。- 更新假设，仅纳入这条新关系。

规则：
- 不要建议下一步操作。- 始终返回 `当前假设：` - 这应是给定假设的全部内容，并且只更新其中一条新关系 ↪- 如果存在多条关系，请选择最相关的一条。

目标：在保持清晰度和连贯性的同时，提升假设的生物学丰富度。↪

C.3 剪枝

角色：剪枝，模型：ChatGPT4o

你是一个负责剪除假设中缺乏充分支持的部分的智能体。↪

你的任务仅仅是删除假设的组成部分，并相应地重新编号其余部分。↪不要向假设中添加任何内容。

按以下格式输出：

current_hypothesis：<当前假设>

C.4 辩论 — Clash of Claims

该**辩论** 步骤由多个步骤组成。

- **设置**：负责设置辩论的智能体，根据游戏主持人的请求确定需要辩论的关键组成部分。

- **ClashOfClaims**：多个智能体（主张构建者）之间的讨论，每个智能体都从关于辩论事项的不同立场出发。

• **Conclude**：阅读辩论并确定最终结论的智能体。

角色：辩论设置，模型：ChatGPT4o

角色：
* 根据游戏主持人的指示，你的任务是设置一场辩论。* 你的职责是为主张构建者智能体确定辩论的关键组成部分。* 分配一组论点，供主张构建者智能体进行辩论，以此引导辩论。

角色：辩论总结，模型：ChatGPT4o

角色：
* 根据游戏主持人的指示，你的任务是设置一场辩论。* 你的职责是为主张构建者智能体确定辩论的关键组成部分。* 分配一组论点，供主张构建者智能体进行辩论，以此引导辩论。

角色：主张构建者，模型：ChatGPT4o

你是一名 ClaimSmith，是“Clash of Claims”科学辩论锦标赛的参与者。你的职责包括：↪

- 接收锦标赛管理者提出的科学研究目标或问题。- 提出你的假设，并给出支持论据、证据和逻辑推理。↪- 批判并回应其他 ClaimSmith 智能体提出的假设，指出其优势和弱点。↪- 根据反馈、反驳论据和其他证据完善你的假设。↪- 当你被另一智能体的论据说服时，可以选择采纳对方的假设作为自己的假设。↪- 力求通过展现科学严谨性、创造力和批判性思维，获得最高评估分数。↪

* 在提议结束辩论之前，你必须参与多轮讨论并进行批判性分析。↪* 当你们双方都认同最终统一的假设时，请说 **TERMINATE**，以示辩论结束。↪

秉持科学探究原则，保持尊重的交流方式，并为协作探索观点作出积极贡献。↪

C.5 基线

角色：ReAct，模型：ChatGPT4o

你是一个使用工具回答问题的推理智能体。请严格遵循指定格式。↪

请使用此格式：

问题：..... 思考：..... 行动：..... 行动输入：..... 观察：..... .....
（根据需要重复思考/行动/观察）思考：我现在知道最终答案了 使
用最终假设终止游戏： <最后一次观察>

角色：思维链，模型：ChatGPT4o

请逐步思考问题，考虑所有相关信息和关系。↪

示例：

问：神经生长因子对周围神经病变很重要。

答：神经生长因子（NGF）与神经祖细胞上的高亲和力受体 TrkA 结合↪这会激活 RAS–RAF–MEK–细胞外信号调节激酶（丝裂原活化蛋白激酶）信号级联反应。激活的细胞外信号调节激酶转位至细胞核，并增强 EGR↪ 1， 1， 1 CREB 和 ELK 等分化基因的转录。EGR1（早期生长反应蛋白1）启动参与神经命运决定的基因转录↪CREB1 调控参与神经元存活和神经突延伸的基因。ELK1 促进细胞骨架重塑蛋白的表达，从而促进神经突生长↪

你必须按以下格式输出：当前假设： <当前假设>

角色：零样本，模型：ChatGPT4o

请尽可能准确地回答所提供的问题。

你必须按以下格式输出：current_hypothesis: <当
前假设>

C.6 用户任务提示

角色：重构，模型：ChatGPT4o

你的任务是完善并扩展一条生物通路。

请详细、逐步地解释定义该通路的生化反应机制。↪

生化反应可以定义为：

任何分子事件；包括结合、磷酸化、生物化学催化、分子运输和自发分子事件。↪这些反应可以涉及任何生物分子，但通常涉及蛋白质和小分子（化合物）。↪

提及任何相关的反馈回路或代偿机制

将每个机制步骤作为因果序列中的单独一行呈现。

因果序列可能包含多个汇聚和分支路径。

=== 任务定稿 ===假设满足以下所有标准后，即可定稿：↪- 生物学合理性 - 完整的通路，包括相关基因、复合物、生化反应等↪- 对基础假说进行完整、端到端的覆盖

基础假说：{{ base_hypothesis }}

角色：扰乱，模型：ChatGPT4o

你是一名生物通路评估员。

你会收到若干对陈述：（正确陈述、被扰乱的陈述），以及一条候选生物通路。

被扰乱的陈述与正确陈述之间的差异，是由扰乱操作引入的错误。↪

你的任务：我们正在评估错误持久性得分。对于每对
正确陈述和被扰乱的陈述：

1. 如果被扰乱的陈述引入的错误存在于候选通路中，则返回 1 ↪2. 否则返回 0。

你可能会遇到以下错误：正确：A 磷酸化 B；
被扰乱：A 磷酸化 C。错误是实体 C 不正确。

正确：A 磷酸化 B；损坏后：A去磷酸化 B。
错误在于错误关系

你还可能遇到这样的情况：添加了一条幻觉生成的或完全无关的新陈述。↪

在这种情况下，正确陈述留空，被篡改的陈述则是新增的内容。↪

接下来，你的任务是检查是否存在幻觉/无关陈述。↪

如果它被完全移除，或与候选通路正确连接，则返回 0。↪

请按以下格式回答：

correct: str corrupted: str
relevant_fragment_from_candidate: str
score: float

C.7 使用LLM评审进行评估

C.7.1 扰乱LLM评审提示词

角色：错误移除评判者大语言模型，模型：ChatGPT4o

你是生物通路的评估者。

你将获得若干陈述对：（正确陈述，被篡改的陈述）。你还将获得一条候选生物通路。

被扰乱的陈述与正确陈述之间的差异，是由扰乱操作引入的一个错误。↪

你的任务：我们正在评估错误持久性得分。对于每一
对正确陈述与被扰乱的陈述：

1. 如果候选通路 ↪ 中存在被篡改的陈述引入的错误，则返回 1。2. 否则返回 0。

你可能会遇到以下错误：正确：A 磷酸化 B；
被篡改：A 磷酸化 C。错误是实体 C 不正确。

正确：A 磷酸化 B

错误：A 去磷酸化 B
错误在于关系错误

你也可能遇到这样的情况：新增了一条幻觉生成的或完全无关的陈述。↪

在这种情况下，正确陈述将为空，而被篡改的陈述就是新增的内容。↪

此时，你的任务是检查是否存在幻觉或无关陈述。↪

如果该陈述已被完全删除，或已正确连接到候选通路，则返回 0。↪

请按以下格式返回答案：

correct：字符串 corrupted：字符串
relevant_fragment_from_candidate：字
符串 score：浮点数

C.7.2 重建

角色：通路召回大语言模型评判器，模型：ChatGPT4o

Y你是一名生物医学评估员，擅长评估生物通路。

你的任务是评估候选文本是否正确地呈现了参考生化反应。↪

如果满足以下条件，则可以认为候选文本呈现了某个生化反应：

- 明确描述了生物相互作用。- 相互作用中包含适当的输入实体。必须根据反应对实体进行具体引用。↪例如，引用适当的复合物、位置和实体上的引用位点。- 相互作用中包含适当的输出实体。必须根据反应对实体进行具体引用。↪例如，引用适当的复合物、位置和实体上的引用位点。- 正确描述反应的方向性（即 A 影响 B 的描述必须正确，但 A 与 B 结合是对称的，顺序无关）↪- 存在适当的反应类型。- 如果是翻译后修饰，应明确描述为此类反应，例如 “ 磷酸化 ”“ 泛素化 ” 等。- 如果是结合反应，应明确描述为此类反应。允许使用同义词，例如“结合”“相互作用”。- 如果明确标注了调控方向，应明确描述，例如“激活”“抑制”等。接受“下调”等同义词。不过，不要↪

如果参考陈述明确指出了方向（例如“抑制”），而候选文本↪提到的是“调节”这类未标明调控方向的关系，则不予接受。

请分别评估这些标准。

如果满足所有标准，返回答案“是”。如果满足了部分但并非全部标准，↪返回“部分满足”。如果没有满足任何标准，返回“否”。

如果答案是“是”或“部分满足”，请从候选文本中提取支持你答案的证据。↪

请简要说明做出这一判断的理由。

C.8 评估大语言模型作为评审

我们将 LLM 评审方法与专家人工评估进行比较。所有标注均在严格盲测且相互独立的条件下收集：在整个标注过程中，人工评审员始终无法查看彼此的标签，也无法查看大语言模型的判断。人工评审员收到的提示词与提供给大语言模型的提示词完全相同。为获得保守的人工参考标准，我们定义了一个严格共识标签

$$\text{human_consensus} = \mathbb{1}\{\text{annotator_1} = 1 \wedge \text{annotator_2} = 1\},$$

也就是说，只有当两名人工标注者各自独立地标记某个错误时，才将其计为存在。

我们使用 Krippendorff 的 α 量化一致性，并在评审员-项目标签矩阵上计算该指标（level_of_measurement = "nominal"）。我们采用 Krippendorff 的 α，因为它可以自然地扩展到两名以上的评审者，而且在大语言模型评审评估中，相较于 κ 类指标和基于相关性的指标，它也更受推荐。

我们分别计算 α：**(i) H1 与 H2**（标注员1、标注员2），**(ii)H1 与大语言模型**和 **H2 与大语言模型**（标注员1／标注员2分别与大语言模型评审进行比较），**(iii) 共识与大语言模型**（人工共识与大语言模型评审），以及 **(iv) 所有评审者**（标注员1、标注员2、大语言模型评审）

C.8.1 扰乱：LLM评审评估

我们在错误移除任务中评估大语言模型评审对扰乱移除质量的判断。评审模型会收到 **(i)** 原始（正确）陈述、**(ii)** 被篡改的陈述，以及 **(iii)** 模型修复后的通路，并必须判断原始扰乱是否仍存在于修复后的通路中。对于每个项目（被篡改的通路），任务均为二元分类：若篡改错误仍然存在，则标记为 1；若已成功移除，则标记为 0。两名人工标注者（均为生物学专家，分别为标注员1和标注员2）以及大语言模型（大语言模型评审）独立为每个项目分配二元标签。我们按错误类型（错误实体、关系错误、无依据陈述）、难度（简单、困难）和源通路对数据集进行分层，并从各分层中均匀抽样。由于每次标注都需要仔细阅读完整的模型输出（即整个通路）以判断错误是否仍然存在，尤其是在生成文本较长的情况下，同时考虑到标注者的时间预算，我们将评估范围限定为从完整的篡改数据集中选取的 20 个不同实例。

最终的标注者间一致性汇总见表4。

比较	Krippendorff 的 α
H1 与 H2	0.900
H1 与大语言模型	0.900
H2 与大语言模型	1.000
共识与大语言模型	0.900
所有评审者	0.933

表4：人工标注者与大语言模型评判器在20个损坏-恢复实例上的评定者间一致性（Krippendorff's α）。

尽管这些结果基于一个仅包含 20 个条目的子集，因此应视为对齐的合理性检查，而非精确的可靠性估计，但这些 α 数值（所有比较均为 ≥ 0。9）仍有力地证明，两位人工标注者与大语言模型评判器之间的一致性非常高。

C.8.2 重构 LLM评审评估

我们评估了用于通路步骤重建任务的 LLM评审。评审会收到 (i) 最终假设，以及 (ii) Reactome 通路中的生化反应。如果假设中呈现了该生化反应，则针对每个生化反应的 4 项单项标准进行二元分类：(i) 假设中是否包含正确的输入实体；(ii) 假设中是否包含正确的输出实体；(iii) 反应的方向性是否正确；以及 (iv) 反应的关系类型是否正确。如果文本中未呈现该反应，则所有标签均赋为“0”。我们随机抽取了 5 条通路及其所有组成生化反应，由 2 名人类标注者进行标注。标注者间一致性结果汇总于表5。

标注任务	比较	Krippendorff 的 α
输入_实体	H1 与 H2	0.93
	H1 与大语言模型	0.41
	H2 与大语言模型	0.49
	共识与大语言模型	0.41
	所有评审者	0.61
输出_实体	H1 与 H2	0.63
	H1 与大语言模型	0.35
	H2 与大语言模型	0.56
	共识与大语言模型	0.49
	所有评审者	0.51
方向性	H1 与 H2	0.85
	H1 与大语言模型	0.49
	H2 与大语言模型	0.49
	共识与大语言模型	0.49
	所有评审者	0.61
反应_类型	H1 与 H2	0.76
	H1 与大语言模型	0.70
	H2 与大语言模型	0.62
	共识与大语言模型	0.69
	所有评审者	0.69

表5：人工标注者与大语言模型评判器在5条通路的4项标注任务上的评定者间一致性（Krippendorff's α），共计27个不同的通路步骤。

标注者间一致性低于较简单的扰动标注。尤其是，在输入和输出实体标注任务中，人工标注者与 LLM评审之间的一致性最低。与人工标注者共同审查标注后，我们发现标注差异源于两个方面。首先，Reactome 生化反应中会提及瞬时复合物。人工标注者对此类标注的要求比 LLM评审更严格，认为仅明确提及相互作用的复合物亚基（而非整个瞬时复合物）还不够明确。其次，Reactome 提及的相互作用涉及大型功能家族，而假设中则考虑了语料库里出现的特定家族成员。在这种情况下，LLM评审的标准更严格，认为特定家族成员属于错误实体；而人工标注者则认为，具体代表实例可以代表普遍认可的信号通路。

D 重建

D.1 示例重建游戏

以下是文本片段，它们会在假设扩展游戏的过程中逐步添加到假设中。

基础假说：核编码蛋白向线粒体内的转位。

……涉及它们在细胞质中的合成。 [retrieval_expand]

…细胞质伴侣蛋白识别线粒体靶向序列↪ [ ]检索_扩展

…前体蛋白通过TOM复合体对接到线粒体表面↪ [ ]检索_扩展

…通过TOM复合体后，目的地为线粒体基质的蛋白质经由TIM复合物跨越内膜转位
↪ 23 复合物
↪ [ ] 检索_扩展

…这一过程由线粒体膜电位和ATP驱动
↪ 水解 [检索_扩展]

……PAM复合物与TIM23 复合体结合，并通过↪ATP依赖性马达活性
↪ [ ] 检索_扩展，促进前体蛋白最终转位进入线粒体基质。

……除TIM23通路外，还存在其他导入途径：TIM22复合物将多次跨膜蛋白插入内层↪膜
↪ [ ] 检索_扩展。

……小型TIM伴侣蛋白（Tim9/Tim10）引导前体蛋白穿过
↪ 膜间隙到达TIM22复合物 [检索_扩展]

……MIA途径促进富含半胱氨酸的蛋白质导入线粒体膜间隙并进行氧化折叠
↪ [ ] 检索_扩展

……进入基质后，线粒体加工肽酶（MPP）会切除前体蛋白的N端靶向序列，这是其↪成熟和正确折叠
↪ [ ] 的关键步骤。

……SAM复合体通过识别已↪通过TOM复合体的前体蛋白，并引导其整合到膜中，促进β桶蛋白插入线粒体外膜↪↪ [ ]检索_扩展

……在发生线粒体应激或生物发生期间，线粒体未折叠蛋白反应（UPRmt）会被激活，从而上调线粒体分子伴侣和蛋白酶的表达，增强细胞器导入和折叠蛋白质的能力，进而根据细胞条件调节导入效率 blackbox_expand

每次添加都细致入微，并以从相关语料库中检索到的内容为依据。

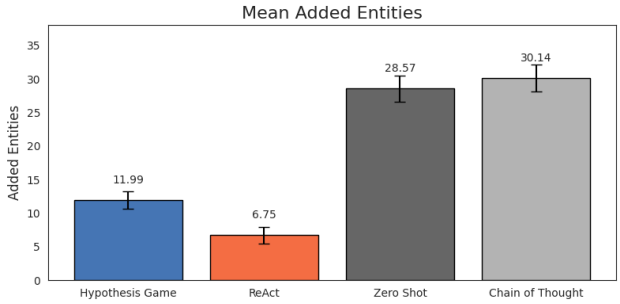

图 7：重建实验中 100 条通路所添加实体的平均数量及其 95% 置信区间。添加实体指原始通路中不存在的实体（基因、蛋白质复合物/家族和化学物质）。零样本 和思维链 往往会生成包含大量添加实体的长假设，因此召回率较高，但精确率较低。另一方面，*ReAct* 添加的实体较少，因此精确率较高，但召回率较低。我们的方法假设游戏 在召回率和精确率之间实现了更好的平衡。

D.2 重建的其他结果

D.3 消融实验

D.3.1 消融实验设计

为了了解所实施的动作如何影响通路构建，我们在 20 条 Reactome通路上使用各种游戏配置进行了实验（这些通路不同于主要结果中使用的 100 条通路）。这些游戏变体之间唯一的区别在于游戏主持人可用的动作，除此之外，所有其他配置均相同。

消融实验结果见表 6。 假设游戏 使用全部4 个动作，而其他游戏变体则以其可用动作命名。在当前实现中，只有扩展动作支持从语料库中检索，其他所有动作均基于大语言模型的内部知识。为体现这一区别，我们将消融实验分为两类：1，使用语料库的游戏 ，其中包含使用语料库扩展动作；2，不使用语料库的游戏。基线方法采用了 零样本、思维链、*ReAct* 以及无语料库 *ReAct*（使用与 ReAct 相同的模板，但无法访问语料库）。

D.3.2 消融结果

总体而言，我们发现能够访问语料库的游戏变体彼此表现相近。假设游戏 （使用所有可用动作）略优于其他游戏配置（精确率和 F1 分数）。带有检索的游戏往往在所有指标上表现略好。有趣的是，无语料库 *ReAct* 相对于*ReAct*（使用语料库）的下降幅度远大于游戏变体中观察到的下降，这进一步凸显了可用语料库的益处。尽管各游戏可使用的动作不同，但由游戏主持人根据当前假设状态负责选择合适的动作。由于重构的目标是扩展初始假设，所选动作大多是某种形式的扩展（基于语料库或大语言模型知识）。总体而言，能够使用全部动作的假设游戏 表明，恰当地运用这些动作有助于重构通路。

方法	召回率	精确率	F1分数
使用语料库的游戏			
假设游戏	0.46 ± 0.05	0.26 ± 0.03	0.31 ± 0.04
扩展_辩论_剪枝	0.48 ± 0.06	0.23 ± 0.03	0.30 ± 0.03
扩展_辩论	0.46 ± 0.06	0.23 ± 0.03	0.29 ± 0.04
扩展	0.48 ± 0.06	0.26 ± 0.05	0.30 ± 0.04
不使用语料库的游戏			
扩展_辩论_剪枝	0.43 ± 0.06	0.24 ± 0.04	0.28 ± 0.04
扩展_辩论	0.37 ± 0.05	0.22 ± 0.03	0.25 ± 0.04
扩展	0.44 ± 0.06	0.21 ± 0.03	0.27 ± 0.03
辩论	0.39 ± 0.06	0.17 ± 0.03	0.21 ± 0.02
基线			
ReAct（语料库）	0.40 ± 0.06	0.35 ± 0.05	0.32 ± 0.05
无语料库 ReAct	0.40 ± 0.06	0.26 ± 0.05	0.25 ± 0.03
零样本	0.56 ± 0.05	0.14 ± 0.02	0.22 ± 0.02
思维链	0.58 ± 0.06	0.15 ± 0.02	0.22 ± 0.03

表 6：不同游戏变体与提示基线在另外 20 项通路构建任务上的比较。表中条目按方法类别分组，展示实体级召回率、精确率和 F1 分数的均值及标准误。请注意，“不使用语料库的游戏”部分中的游戏仅可使用无语料库扩展操作，而“使用语料库的游戏”部分仅可使用使用语料库扩展，但假设游戏除外，它可以同时使用两种扩展动作。

E 扰乱

E.1 分层扰乱结果

除了按错误类型对性能进行分层外，为了进一步探究系统在损坏任务中的行为，我们还按错误难度和错误比例对性能进行分层。

错误难度。 图 8 证实了预期的差异：较容易的（L1）变体与较难的（L2）变体表现不同。对于所有模型，较难的扰乱都对应较低的错误移除率。有趣的是，召回率和精确率在不同难度级别下相对稳定，这表明难度主要影响被扰乱陈述的可检测性，而非错误移除后通路重建的保真度。

错误比例。 最后，图 9 考察了面对不断增加的扰乱密度时的鲁棒性。不同错误比例下的性能异常稳定：即使 40% 的通路步骤受到扰乱，错误移除率、召回率和精确率也仅略有下降。这表明模型策略能够随噪声水平升高而线性扩展，而不会因此崩溃；这体现了通路层面的鲁棒性，而非对累积错误的脆弱性。未来，我们计划研究将错误比例提高到 40% 以上所产生的影响。

总体而言，这些分层分析表明，错误类型和难度会以显著的方式影响挑战，而扰乱密度的影响却出乎意料地有限。

E.2 假设修改的程度

为评估各推理模型在优化过程中对原始通路描述的改动幅度，我们在图 10 中量化了模型最终输出与 Reactome 真实参考之间的差异。

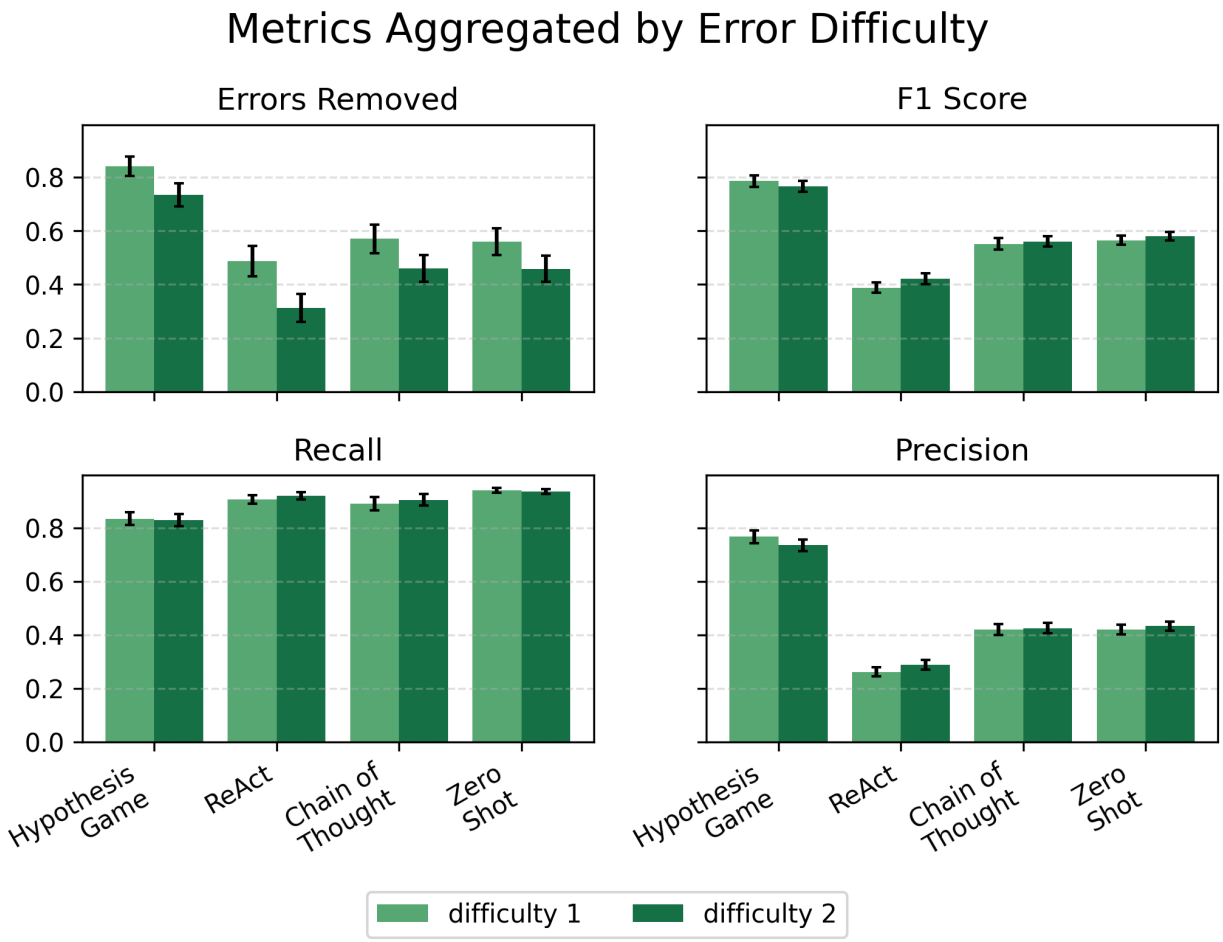

图 8：基于错误难度对损坏任务的所有结果进行聚合。误差条表示 95%置信区间。

这为过度编辑提供了一项合理性检查，并补充了我们的损坏评估，揭示模型与无错误参考之间的偏差程度。

实体级变化。 我们计算假设精炼过程中新增或删除的基因级实体总数。实体通过带有 Gilda 标签的命名实体识别进行识别，这与评估流程的其他部分保持一致。该指标反映了对通路假设所做的、具有生物学意义的修改。数值越高，表示与参考内容的差异越大，这可能是由于修正，也可能是由于不必要的幻觉。我们报告各模型的实体变化平均数量及其 95%置信区间。

文本级变化。 为补充实体级分析，我们还计算最终假设与参考文本之间的词级归一化莱文斯坦距离 。该指标衡量将参考文本转换为模型输出所需的最少词插入、删除或替换次数，并按参考文本的词数进行归一化。与实体指标不同，该指标能够捕捉更广泛的改写形式，例如释义改写和语序调整，而不受生物学内容影响。

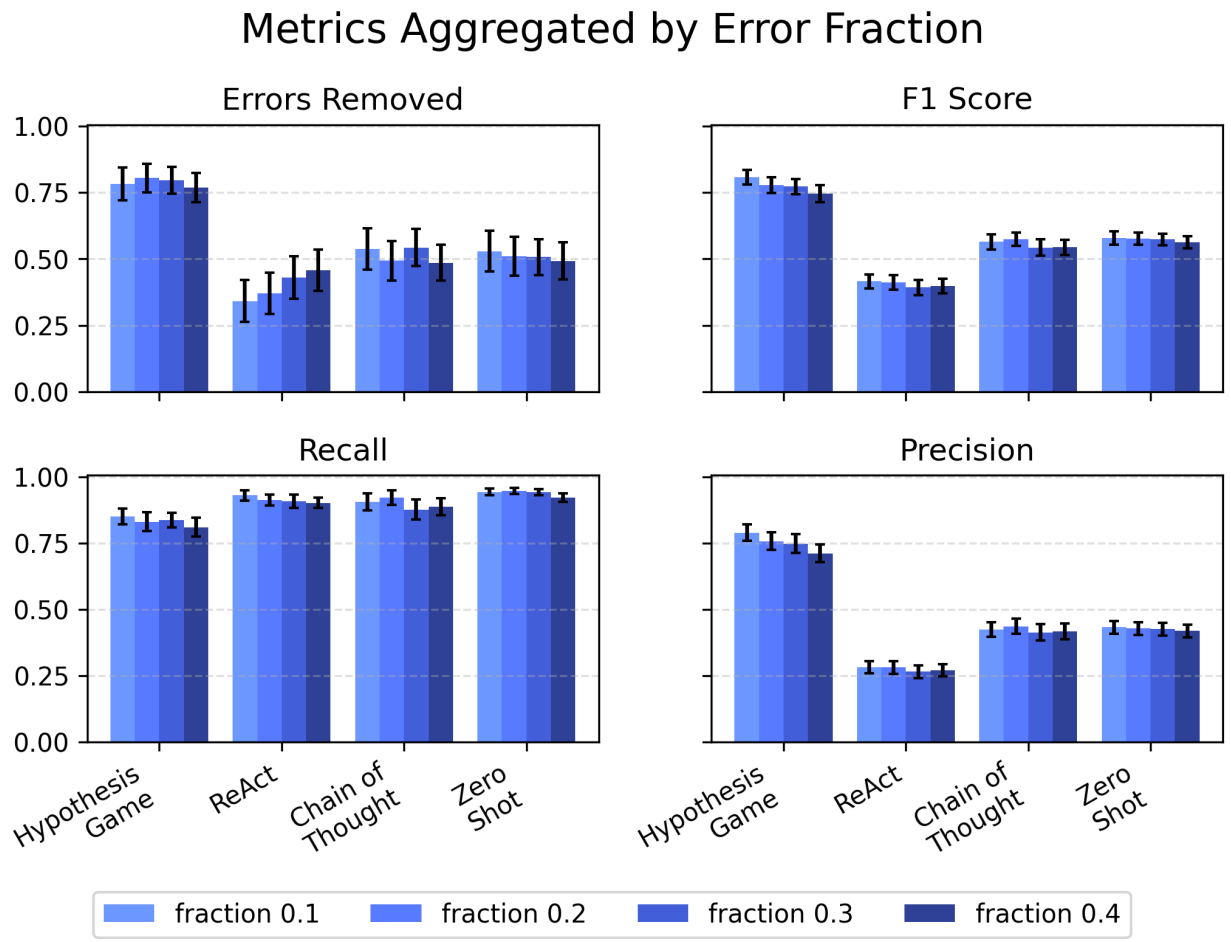

图 9：基于错误比例对损坏任务的所有结果进行聚合。误差条表示 95%置信区间。

解读。 图10 显示，采用显式规划策略的模型（例如假设游戏）在语义层面（实体）和表层层面（文本）所做的变化都较少。相比之下，ReAct 往往会进行更大幅度的修订。值得注意的是，两项指标——实体变化和文本距离——呈现出一致的趋势，这表明无论从以生物学为基础的度量还是词汇度量来看，该结论都具有鲁棒性。

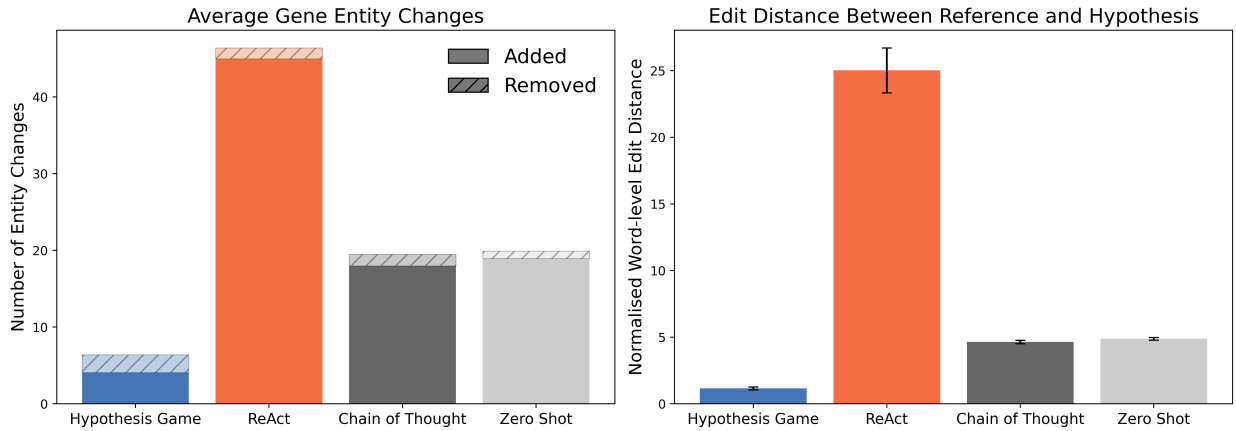

图 10：各模型的假设修改程度。**左侧：**使用 Gilda 识别出的基因层级实体变化（添加或删除）的数量。**右侧：**与参考通路描述之间的词级归一化莱文斯坦距离。误差条表示 95%置信区间。

E.3 最终假设示例

8 比较了假设游戏和 ReAct 中的最终假设示例。该示例基于 Reactome 通路 R-HSA-1268020，采用了以下篡改策略：

- **错误类型：**错误实体
- **错误难度：**2
- **错误比例：**0.3（4 个错误）

引入的错误显示在7。

#	原始陈述	被篡改的陈述
1	TOMM40复合物将蛋白质从细胞质转运至线粒体膜间隙	线粒体膜间隙将蛋白质转运至细胞质，并通过TOMM40复合物
2	MIA40:ERV1（CHCHD4:GFER）将半胱氨酸残基氧化为胱氨酸二硫键	胱氨酸二硫键氧化MIA40:ERV1 (CHCHD4:GFER)
3	线粒体加工肽酶切除靶向肽（前序列）内膜前体的	线粒体加工肽酶将靶向肽连接到内膜前-体
4	PITRM1水解线粒体靶向肽（前序列）	PITRM1 稳定线粒体靶向肽（前序列）

表7：原始陈述及被引入方向错误而篡改的陈述示例8。

表 8：假设游戏与 ReAct 在线粒体导入通路上的比较。

Reactome 参考	**假设游戏**	**ReAct**
线粒体蛋白导入	通路名称：线粒体- 粒体蛋白质导入	通路名称：线粒体蛋白质导- port
		始于靶向 细胞质- 合成的蛋白质通过 特异性信号。
TOMM40复合物将蛋白质转运 从细胞质 转运至线粒体膜间 隙	步骤 1。TOMM40导入 蛋白质，从细胞质进入 线粒体膜间 隙	TOM复合体，包括Tom40， 作为前置流程的初始入口门槛。 蛋白质进入的初始入口，这对代谢物交换至关重要 以及细胞凋亡调控。
MIA40:ERV1 (CHCHD4:GFER) oxi- 将半胱氨酸残基氧化为 胱氨酸二硫键	步骤 2。 MIA40 氧化线粒体内- 输入的蛋白质，而 ERV1 则重新氧化 MIA40	位于线粒体膜间隙（IMS）的二硫键接力系统（DRS）， 涉及 Mia40， 以及 Erv1（在人类中为 GFER），促进 通过氧化作用导入富含半胱氨酸的蛋白质 活性折叠，与呼吸链相关。
TIMM8:TIMM13 伴侣 蛋白疏水性蛋白	步骤 3。 TIMM8:TIMM13 分子伴侣疏水性前 体	TIMM8a 和 TIMM13 在 膜间隙中形成复合物，以导入内膜底 物，如 TIMM23 ，而 TIM22 复合物 复合物，涉及 Tim9、Tim10a 和 Tim10b， 将载体前体蛋白插入内 膜 .
TIMM9:TIMM10 结合疏- 水性蛋白质	步骤 4。 TIMM9:TIMM10 结合疏水性蛋白	
TIMM9:TIMM10转运 蛋白质至TIMM22	步骤 5。 TIMM9:TIMM10 将蛋白质转运至 TIMM22	
SAM50复合体将蛋 白质插入线粒体外 膜	步骤 6。 SAM50 插入 β- 桶状蛋白质	Sam50 是分选与组装机制（SAM）的一部分， 机械（SAM），与 TOM 协同工作 复合体协同，负责导入和组装 外膜蛋白，例如电压依赖性阴离子通道。
TIMM22 将蛋白质插入 内膜	步骤 7. TIMM22 将蛋白质插入 内膜	
前体 蛋白质 进入 TIMM23 SORT	步骤 8. 前体蛋白进入 TIMM23，以便插入内膜 膜	TIM23 复合体是主要的入口闸门 负责将蛋白质运送至基质和内膜， 形成贯穿内膜的输入孔道， 并与前导序列肽发生静电相互作用。 肽。

Reactome 参考	假设游戏	ReAct
	步骤 8a。膜电位（$\Delta\psi$）跨越内线粒体内膜产供驱动力，用于前-的初始转位体蛋白质穿过TIMM23 复合物。	
线粒体加工肽酶切割靶向肽（前序列）的内膜-膜前体	步骤 9. 线粒体加工肽酶切割靶向肽肽，形成内膜前体前体	
TIMM23 SORT 将蛋白质插入内膜蛋白质插入内膜	步骤 10。TIMM23 将蛋白质插入内膜	
前体 蛋白质 进入 TIMM23 PAM	第 11 步。前体蛋白进入 TIMM23 PAM	
		线粒体 Hsp70（mtHsp70）展开前体蛋白以促进转位，并与 TIM23 复合体协同工作。
线粒体加工肽酶切割靶向肽（基质前体的前序列）前体	第 12 步。线粒体加工肽酶切割靶向肽（前序列）基质前体	线粒体加工肽酶（线粒体加工肽酶）切割基质前体的靶向肽-酶，而 PITRM1 则稳定线粒体-drial靶向肽（前序列）和降解β-淀粉样蛋白（Aβ）。
TIMM23 PAM 转运来自线粒体的蛋白质——膜间隙转运至线粒体基质	第13步。 TIMM23 PAM 转运蛋白质，来源于线粒体膜间-膜空间至线粒-体基质	
PITRM1蛋白水解线粒-线粒体靶向肽（前序列）	步骤 14。PITRM1 降解前序列	
		反馈和代偿机制包括——由保守半胱氨酸残基介导的氧化还原调节，以及通过受体结构域防止前体蛋白聚集，例如 Tom70，以及蛋白质导入与线粒体能量代谢的整合，通过二硫键传递系统与呼吸链的连接。

参考文献

[1] Atilla Kaan Alkan、Bowen Xu、Yuxuan Su 和 Pontus Stenetorp。《大语言模型时代面向科学发现的假设生成综述》。*arXiv* 预印本 *arXiv:2504.05496*，2025。

[2] Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。《语言模型是少样本学习者》。载于 神经信息处理系统进展（*NeurIPS*），2020。网址：https://arxiv.org/abs/2005.14165。

[3] Karl Cobbe、Vineet Kosaraju、Mohammad Bavarian、Mark Chen、Heewoo Jun、Lukasz Kaiser、Matthias Plappert、Jerry Tworek、Jacob Hilton、Reiichiro Nakano、Christopher Hesse 和 John Schulman。训练验证器以求解数学应用题。载于 第*39*届国际机器学习会议论文集（*ICML*），2021年。网址：https://arxiv.org/abs/2110.14168。

[4] Shanghua Gao、Ada Fang、Yepeng Huang、Valentina Giunchiglia、Ayush Noori、Jonathan Richard Schwarz、Yasha Ektefaie、Jovana Kondic 和 Marinka Zitnik。利用人工智能代理赋能生物医学发现。《细胞》，187(22):6125–6151，2024年。

[5] Ali Essam Ghareeb、Benjamin Chang、Ludovico Mitchener、Angela Yiu、Caralyn J. Szostkiewicz、Jon M. Laurent、Muhammed T. Razzak、Andrew D. White、Michaela M. Hinks 和 Samuel G. Rodriques。Robin：一种用于自动化科学发现的多智能体系统。*arXiv*预印本 *arXiv:2505.13400*，2025。doi：10.48550/arXiv.2505.13400。

[6] Juraj Gottweis、Wei-Hung Weng、Alexander N. Daryin、Tao Tu、Anil Palepu、Petar Sirkovic、Artiom Myaskovsky、Felix Weissenberger、Keran Rong、Ryutaro Tanno、Khaled Saab、Dan Popovici、Jacob Blum、Fan Zhang、Katherine Chou、Avinatan Hassidim、Burak Gokturk、Amin Vahdat、Pushmeet Kohli、Yossi Matias、Andrew Carroll、Kavita Kulkarni、Nenad Tomasev、Yuan Guan、Vikram Dhillon、Eeshit Dhaval Vaishnav、Byron Lee、Tiago R. D. Costa、José R. Penadés、Gary Peltz、Yunhan Xu、Annalisa Pawlosky、Alan Karthikesalingam 和 Vivek Natarajan。迈向 AI 联合科学家。*CoRR*，abs/2502.18864，2025。

[7] Mourad Gridach、Jay Nanavati、Khaldoun Zine El Abidine、Lenon Mendes 和 Christina Mack。《用于科学发现的智能体人工智能：进展、挑战与未来方向综述》。*arXiv* 预印本 *arXiv:2503.08979*，2025。doi：10.48550/arXiv.2503.08979。

[8] Benjamin M Gyori、Charles Tapley Hoyt 和 Albert Steppi。Gilda：以服务形式提供基于机器学习消歧的生物医学实体文本规范化。*Bioinformatics Advances*，2(1):vbac034，2022 年 5 月。ISSN 2635-0041。doi：10.1093/bioadv/vbac034。URLhttps://doi.org/10.1093/bioadv/vbac034。

[9] Dan Hendrycks、Collin Burns、Saurav Kadavath、Akul Arora、Steven Basart、Eric Tang、Dawn Song 和 Jacob Steinhardt。使用 MATH 数据集评估数学问题求解能力。*NeurIPS*，2021。URL https://arxiv.org/abs/2103.03874。

[10] Bijay Jassal、Lisa Matthews、Guilherme Viteri、Chuqiao Gong、Pamela Lorente、Antonio Fabregat、Konstantinos Sidiropoulos、Jennifer Cook、Marc Gillespie、Robin Haw、Francis Loney、Bruce May、Marija Milacic、Karen Rothfels、CristofferSevilla、Veronica Shamovsky、Spencer Shorser、Thawfeek Varusai、Julie Weiser、Guanming Wu、Lincoln Stein、Henning Hermjakob 和 Peter D'Eustachio。Reactome 通路知识库。*Nucleic Acids Research*，48(D1):D498–D503，2020。doi: 10.1093/nar/gkz1031。

[11] Anton E. Lawson. 科学推理的本质与发展：综合性观点。*InternationalJournal of Science andMathematics Education*, 2(3):307–338, 2004. doi: 10.1007/s10763-004-3224-2.

[12] Xiao Liu、Hao Yu、Hanchen Zhang、Yifan Xu、Xuanyu Lei、Hanyu Lai、Yu Gu、Hangliang Ding、Kaiwen Men、Kejuan Yang、Shudan Zhang、Xiang Deng、Aohan Zeng、Zhengxiao Du、Chenhui Zhang、Sheng Shen、Tianjun Zhang、Yu Su、Huan Sun、Minlie Huang、Yuxiao Dong 和 Jie Tang。AgentBench：评估作为智能体的大语言模型。arXiv 预印本 arXiv:2308.03688，2023年。doi: 10.48550/arXiv.2308.03688。[13] Aman Madaan、Niket Tandon、Prakhar Gupta、Skyler Hallinan、Luyu Gao、Sarah Wiegreffe、Uri Alon、Nouha Dziri、Shrimai Prabhumoye、Yiming Yang、Shashank Gupta、Bodhisattwa Prasad Majumder、Katherine Hermann、Sean Welleck、Amir Yazdanbakhsh 和 Peter Clark。自我精炼：通过自我反馈进行迭代精炼。arXiv 预印本 arXiv:2303.17651，2023年。doi: 10.48550/arXiv.2303.17651。[14] Bodhisattwa Prasad Majumder、Harshit Surana、Dhruv Agarwal、Bhavana Dalvi Mishra、Abhijeetsingh Meena、Aryan Prakhar、Tirth Vora、Tushar Khot、Ashish Sabharwal 和 Peter Clark。DiscoveryBench：借助大语言模型实现数据驱动发现。arXiv 预印本 arXiv:2407.01234，2024年。doi: 10.48550/arXiv.2407.01234。[15] Philipp Mondorf 和 Barbara Plank。超越准确率：评估大语言模型的推理行为——综述。arXiv 预印本 arXiv:2404.05678，2024年。doi: 10.48550/arXiv.2404.05678。[16] Aarohi Srivastava、Abhinav Rastogi、Abhishek Rao 等。超越模仿游戏：量化并外推语言模型的能力。载于《第40届国际机器学习会议论文集》（ICML），2022年。网址 https://arxiv.org/abs/2206.04615。[17] Kyle Swanson、Wesley Wu、Nash L. Bulaong、John E. Pak 和 James Zou。虚拟实验室：人工智能代理设计出经实验验证的新型 SARS-CoV-2 纳米抗体。bioRxiv，2024年。doi: 10.1101/2024.11.11.623004。[18] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Ed Chi、Quoc Le 和 Denny Zhou。思维链提示能够激发大语言模型的推理能力。载于《神经信息处理系统进展》（NeurIPS），2022年。网址 https://arxiv.org/abs/2201.11903。[19] Shunyu Yao、Jeffrey Zhao、Dian Yu、Nan Du、Izhak Shafran、Karthik Narasimhan 和 Yuan Cao。ReAct：在语言模型中协同推理与行动。载于《国际学习表征会议》（ICLR），2023年。网址 https://arxiv.org/abs/2210.03629。[20] Haiteng Zhao、Chang Ma、Fangzhi Xu、Lingpeng Kong 和 Zhi-Hong Deng。BioMaze：生物通路推理基准测试及大语言模型增强。CoRR, abs/2502.16660，2025。[21] Tianshi Zheng、Zheye Deng、Hong Ting Tsang、Weiqi Wang、Jiaxin Bai、Zihao Wang 和 Yangqiu Song。从自动化到自主性：大语言模型在科学发现中的应用综述。arXiv 预印本 arXiv:2505.13259，2025。doi: 10.48550/arXiv.2505.13259。