**实验流程框架**

1. **四个阶段**

**阶段一：灵感检索与图构建（GraphRag）**

1. **研究问题拆解与扩充（是否有必要？后续再定）**

* **输入**：研究兴趣：“Transformer 在多义隐喻识别中的局限”
* **操作**：调用 LLM（temperature=0.7，max\_tokens=64）生成 10 条检索短句。

**核心概念分解**: Agent首先对输入进行语义分解，识别出核心实体和概念。

* 主题 (Topic): "多义隐喻识别"
* 技术 (Technology): "Transformer"
* 问题 (Problem): "局限性 (Limitations)"
* **输出**：["Transformer 隐喻 分类 方法", …]
* problem\_centric: "Transformer metaphor ambiguity", "challenges in Transformer context modeling for metaphors" (深入挖掘问题本身)
* solution\_centric: "symbolic reasoning for NLP ambiguity", "knowledge graph enhanced language models" (寻找潜在的解决方案，即使它们当前未与主题直接关联)
* comparative: "graph neural networks vs transformers for context understanding", "symbolic AI vs deep learning for metaphor" (寻找替代性技术路径)
* outlier\_queries: "cognitive science of metaphor processing", "computational creativity theories" (寻找可能带来颠覆性创新的“远关系”领域)

1. **文献检索（外部知识库补充，基于已有的数据构建完向量数据库之后再说）**

* **接口**：Semantic Scholar API，top\_k=100
* **结果处理**：提取 title+abstract，并用 Grobid 解析 PDF 中的段落文本。

1. **灵感节点提取(抽取，当前已完成的数据准备)**

* **方法**：LLM 抽取显式“灵感句”
* 参数：temperature=0.3，max\_tokens=128
* 示例输出：["在 Transformer 中引入符号表示以增强多义性捕获","基于图网络的上下文感知方法"]
* **节点数**：每篇文献提取 2–3 条，汇总得约 250 个节点。

1. **构建关联边(构建，当前重点任务之一)**

* **知识图谱**：设计点权、边权、语义（先构建向量数据库）

1. **输出**：一个加权无向图，|V|≈250，|E|≈600。

**阶段二：图上搜索（Graph‑MCTS）**

① Graph‑MCTS：科研灵感的结构化推演引擎





② EA/GA：灵感子图的结构重组器



③ Bayesian Optimization：超参数调优师



1. **初始状态**：不仅选择高中心性节点，优先选择**类型为Conceptual\_Challenge或与初始问题高度相关的节点**作为根节点。
2. **Action 定义**：

* a\_follow\_relation(relation\_type): 从当前节点出发，沿着指定类型的边（如extends\_method\_of）进行扩展。这是MCTS进行语义导航的核心。
* a\_prune\_path\_end: 如果路径的末端节点导致后续探索的奖励持续走低，则回溯一步，剪掉路径末端。
* a\_mergePaths: 合并两条子图路径；
* a\_synthesize\_cluster: 当路径进入一个密集的节点簇时，调用LLM将该簇总结为一个“元概念”，并将其作为一个抽象节点加入路径，实现更高层次的推理。

1. **Reward 设计**：

* 结合节点质量（文献引用度等综合考虑的点权）、子图连通性、知识覆盖度，计算图上子路径的综合分,**奖励 = w₁ \* 连贯性 + w₂ \* 新颖性 + w₃ \* 重要性 + w₄ \* 问题解决度**

1. **连贯性 (Coherence)**:

* **计算**: 路径是否逻辑自洽。可以基于ANALYZE\_RELATIONSHIP动作的结果。例如，如果关系是SUPPORTS\_METHOD，则加分；如果是CONTRASTS\_WITH，可以暂时不加分，但如果最终形成的假说是整合对立观点的，则在FORMULATE\_HYPOTHESIS时给予高分。
* **LLM评估**: 可以让一个“评估者”LLM读取current\_hypothesis和path上的摘要，为路径的逻辑连贯性打分。

1. **新颖性 (Novelty)**:

* **计算**: **FIND\_BRIDGE动作应给予最高的新颖性奖励！** 如果连接的两个Main Paper属于不同的学科分类或关键词簇，新颖性得分更高。
* 可以计算路径在KG中的“罕见度”。如果一条路径连接了两个在图上距离很远的节点簇，则认为它是新颖的。

1. **重要性 (Significance)**:

* **计算**: 路径上所有Main Paper和来源论文的引用量的平均值或加权平均值。一条由高影响力论文构成的路径更有可能产生重要的发现。

1. **问题解决度 (Problem-Solving Score)**:

* **计算**: 当FORMULATE\_HYPOTHESIS或REFINE\_HYPOTHESIS动作被执行后，评估current\_hypothesis在多大程度上回应或解决了路径起点的背景问题。
* **LLM评估**: 可以让LLM判断“假说”是否是“初始问题”的一个潜在答案，并给出0-1的评分。如果路径成功将一个问题和一个来自“桥梁论文”的解决方案联系起来，并形成了假说，则此项得分应该最高。

1. **Graph‑MCTS 流程**：

* **SELECT**：基于 UCT 在子图动作空间中选取最优操作；
* **EXPAND**：执行上述动作生成新子图:

**核心动作定义：**

1. **EXPAND\_INSPIRATION (扩展灵感)**:

* **当前节点**: Main Paper 或 背景问题。
* **动作**: 遍历当前论文提出的2-4个“问题-灵感”对，选择一个灵感节点作为下一个目标。
* **目的**: 深入研究当前论文的解决方案。

1. **TRACE\_SOURCE (追溯来源)**:

* **当前节点**: 灵感。
* **动作**: 跳转到该灵感对应的来源论文（Source Paper）。
* **目的**: 寻找灵感的源头，这是知识溯源的关键一步。

1. **FIND\_BRIDGE (寻找桥梁)**:

* **当前节点**: 来源论文。
* **动作**: 在整个KG中，查找**还有哪些其他Main Paper**（尤其是在不同领域的）也引用了这篇来源论文作为它们的灵感。选择其中一篇Main Paper作为下一个目标。
* **目的**: **这是产生创新和交叉学科灵感的核心动作！** 它将两个原本不相关的Main Paper通过一个共同的来源论文连接起来，构成了知识的“桥梁”。

1. **ANALYZE\_RELATIONSHIP (分析关系)**:

* **当前节点**: 来源论文。
* **动作**: 检视并评估这篇来源论文与路径上前一个Main Paper之间的关系（您定义的 每个灵感论文对main论文的关系）。
* **目的**: 理解知识是如何被借鉴和应用的。如果关系是“提供了新方法”，则可能是一条有价值的路径；如果是“观点对立”，则可能导向一个辩证的、颠覆性的假说。

1. **FORMULATE\_HYPOTHESIS (形成假说)**:

* **触发条件**: 当路径达到一定长度（例如，包含了一个FIND\_BRIDGE动作后）或在特定类型的节点上。
* **动作**: 这不是一个图上的移动动作。它调用一个LLM，输入整个path上的节点信息（特别是初始的背景问题和路径上所有灵感），生成一个初始的current\_hypothesis文本。
* **目的**: 将结构化的路径探索，转化为可理解的、具体的科学断言。

1. **REFINE\_HYPOTHESIS (精炼假说)**:

* **触发条件**: 每当路径扩展一步（尤其是在FIND\_BRIDGE之后）。
* **动作**: 调用LLM，输入已有的current\_hypothesis和新加入的节点信息，让LLM对假说进行迭代、修正和补充。
* **目的**: 保证假说与探索路径的实时同步和逻辑一致性。
* **EVALUATE**：用评估Agent（住后面阶段中的 Judge）给当前子图打分；
* **BACKPROPAGATE**：更新子图节点的价值和访问次数。

1. **剪枝机制**：

* 在图上执行a\_pruneNode：若删点后子图仍连通且分数不下降，则永久删除；
* 控制子图大小、防止指数级爆炸。

1. **输出**：一组高分子图，每个子图代表一组高度关联的“灵感集合”。

**阶段三：假说生成与多智能体博弈**

1. **Students Agent (LLM1–LLM3)**：

* 并行地，对每个子图生成不同风格的初始假说（方法、实验设计等）。

1. **Mentor Agent (LLM4)**：

* 读取 Students 的所有假说，提出改进建议（针对结构不清晰、可行性欠缺等）。

1. **Judge Agent (LLM5)**：

* 对每轮更新后的假说进行细粒度打分（原创性、可行性、伦理性等维度）。

1. **博弈与迭代**：

* Students 根据 Mentor 和 Judge 的反馈，迭代更新假说；
* 反复若干轮（例如 3–5 轮），直至 Judge 打分收敛或达到预设阈值。

1. **最终选优**：

* 按照 Judge 平均得分或 ELO 排名，选取 top‑K 假说。



**点击图片可查看完整电子表格**

**博弈流程**

1. Students 同时生成 K=3 条假说；
2. Mentor 针对每条给建议，并要求修改；
3. Students 修改后提交 Judge 评分；
4. Socratic Loop：每轮 Judge 得分提升 ≥5% 则继续，否则终止；最多3 轮；
5. **最终结果**：选择 Judge 评分最高的一条。

**阶段四：实证验证与评估**

1. **实验证明设计**：

* 对选优假说，设计小规模实验或案例研究（可在模拟环境或小样本数据上验证）。

1. **评估指标**：

* **人类评分**：邀请 5–10 位专家盲评假说质量；
* **自动指标**：Judge Agent 得分、假说中方法可执行性自动检测（如代码/伪代码可运行检查）；
* **Novelty Index**：计算与现有文献的重合度（n‑gram 重合、引文相似度）；
* **连通性度量**：最终子图的平均聚类系数、覆盖率。

1. **对比实验**：

* Baseline A：纯 LLM（GPT-4）一次性生成假说；
* Baseline B：LLM + 传统 MCTS（树结构）；
* 我们的方法：Graph‑MCTS + 多 Agent 博弈。

1. **统计分析**：

* 用 ANOVA / t‑test 检验各组差异显著性；
* 用可视化（箱线图、热力图）展示指标分布与演进过程；
* 举例展示典型成功和失败案例。

**4.1 自动量化指标**



**点击图片可查看完整电子表格**

**4.2 专家盲评**

* **参与者**：7 位领域专家
* **评价表**：Likert 1–5，5 维度（原创性、可操作性、表达清晰、文献支持、整体价值）
* **统计**：计算均值±标准差；组间对比用 paired t‑test（α=0.05）

**4.3 小规模实证**

* **实验设置**：针对选优假说中的“实验方法部分”，在公共隐喻数据集上复现一个核心实验（如“符号表示增强器”），报告准确率提升。
* **指标**：Baseline Transformer + Proposed Method；对比准确率、F1 分数

2. **实验设计详述**



**点击图片可查看完整电子表格**

3. **对比与消融实验**

1. **Baseline**

* A1: 纯 LLM 一步生假说（GPT‑4）
* A2: 仅 MCTS（树结构）＋LLM

1. **全系统**：完整“五大图模型＋Graph‑MCTS＋EA＋BO＋Multi‑Agent”
2. **消融组**：依次去掉一个模块，检验性能下降：

* 去掉 Cross‑Domain Fusion
* 去掉 Hole Detector
* 去掉 EA/GA 进化
* 去掉 Mentor Agent
* 去掉 Socratic Loop

**指标对比**：Novelty ↑、Feasibility ↑、专家打分 ↑、迭代收敛速度 ↑。



**点击图片可查看完整电子表格**