FUTUREMIND：通过自适应知识蒸馏为小型语言模型赋予战略性思维模式先验

Shaoxiong Yang[1], **Junting Li**[2*], **Mengyuan Zhang**[1],**Chao Li**[1],**Wei Liu**[1],**JianLuan**[1]
MiLM Plus，Xiaomi Inc.[2]伦敦帝国学院计算机科学系
{yangshaoxiong,zhangmengyuan7,lichao75,liuwei40,luanjian}@xiaomi.com
jl523@ic.ac.uk

arXiv:2602.01222v1 [cs.AI] 1 Feb 2026

摘要

小型语言模型（SLM）凭借高效、低延迟的推断能力，在成本敏感且资源有限的场景中颇具吸引力。然而，它们在需要结构化推理和有效检索的复杂知识密集型任务上往往表现不佳。为解决这些局限，我们提出 FutureMind，这是一种模块化推理框架，通过从大语言模型（LLM）进行自适应知识蒸馏，为小语言模型赋予战略性思维模式先验。FutureMind 引入了一种动态推理流程，由四个关键模块组成：问题分析、逻辑推理、策略规划和检索引导。该流程结合了三种不同的检索范式，将复杂查询拆解为可处理的子问题，确保检索执行高效且准确。在包括 2WikiMultihopQA、MuSiQue、Bamboogle 和 Frames 在内的多跳问答基准上进行的大量实验表明，FutureMind 表现优异。在无训练条件下，它在不同的小语言模型架构和规模上持续超越 Search-o1 等强基线方法，取得了当前最优结果。除了实证收益之外，我们的分析还揭示，思维模式蒸馏过程受到教师模型（LLM）与学生模型（SLM）之间认知偏差瓶颈的限制。这为推理能力的可迁移性提供了新的视角，也为开发兼具效率与真正认知能力的小语言模型铺平了道路。

1 引言

近年来，在海量数据集和可扩展计算能力的推动下，大语言模型（LLM）通过预训练（Raffel 等人，2020）、指令微调（Wei 等人，2022a）以及基于人类反馈的强化学习（RLHF）（Touvron 等人，2023；OpenAI，2023），在常识推理（Yang 等人，2025）、代码生成（Guo 等人，2024）和数学推理（Shao 等人，2024）等广泛的通用任务中，展现出卓越的问题理解和解决能力。然而，当问题具有时效性，或需要特定领域知识时（Peng 等人，2023；Li 等人，2023b），模型性能会受到其固有静态参数的限制，暴露出知识过时和领域覆盖不足等缺陷。这一局限凸显了在推理过程中引入外部知识源的必要性。在此背景下，检索增强生成（RAG）（Gao 等人，2023；Xiong 等人，2025）应运而生：通过在推断前向模型提供检索文档，它有效提升了语言模型的准确性和领域适应性。然而，单步检索往往难以应对知识密集型的多跳推理任务（Yang 等人，2018；Ho 等人，2020b）。针对这一问题，近期研究提出了“深度搜索”范式（Li 等人，2025c；Alzubi 等人，2025），强调推理与检索之间的动态交互：在解决问题的过程中，模型持续拆解问题、迭代检索信息并汇总证据，直至答案收敛。

*在小米实习期间完成的作品。

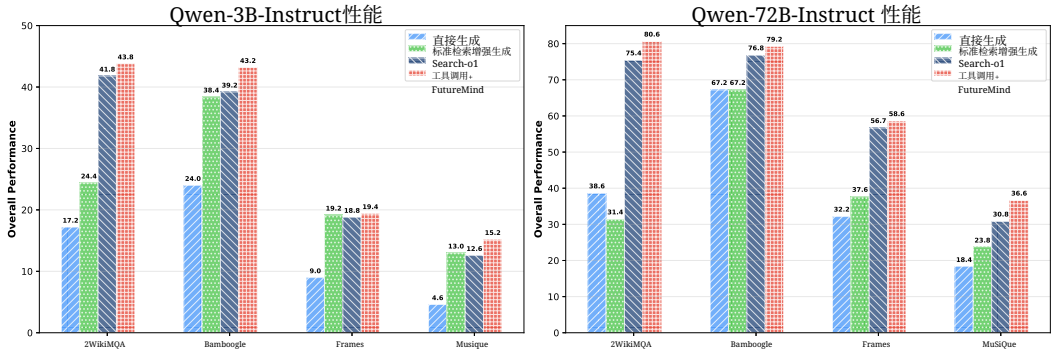

图1：FutureMind 与其他方法在四个多跳问答基准上的整体性能对比。左图展示了30亿参数小语言模型（小语言模型，SLM）上的性能，右图则展示了720亿参数大语言模型（大语言模型，LLM）上的性能。

随着问题复杂度不断增加，仅仅扩大模型规模或增加记忆已不足以应对；有效推理还需要明确的“检索逻辑”，以确定何时、检索什么，以及如何检索相关证据（Schick 等人，2023；Zhang 等人，2024）。Search-o1（Li 等人，2025a）将检索整合进思维链，而 ReAct（Yao 等人，2023）则将“推理—行动—观察—推理”范式形式化，以获取特定的外部信息（Li 等人，2025b）。这些方法体现了一项长期共识：应在推断过程中动态、按需激活大语言模型的能力（Wang 等人，2024b；Jin 等人，2024b）。对于自主智能体而言，这意味着目标应从直接作答转向系统性推理——分析、拆解并整合证据，从而获得更深入的见解并制定更稳健的策略。

然而，实现显式检索逻辑对模型能力提出了很高要求。大语言模型擅长多轮推理和检索，但会带来难以承受的延迟和计算成本（Wan 等人，2023；Wang，2024）。相比之下，小语言模型在效率、成本和隐私方面具有显著优势，但其记忆有限、上下文保持能力较弱，且结构化推理能力受限，难以有效完成问题分解、迭代证据检索和多跳聚合（Wang 等人，2024a；Xu 等人，2025年）。因此，如何在推理效果与计算效率之间实现最佳平衡，仍是一项亟待解决的关键挑战（Bai 等人，2024），尤其对于部署在实时或隐私敏感场景中的资源受限模型而言。

为此，我们提出**FutureMind**，一种免训练模块化推理框架。它借助自适应思维模式蒸馏策略，无需梯度更新即可实现低延迟、高精度的复杂推理。FutureMind这一名称体现了我们对未来人工智能系统的愿景：即便在资源受限的情况下，模型也能借助蒸馏得到的思维模式先验，在免训练条件下泛化到高难度及未见过的问题。FutureMind将推理分解为四阶段流水线——**问题分析**、**逻辑推理**、**策略规划**和**检索引导**——依次解决是否检索、检索什么、如何整合检索证据，以及如何生成连贯答案。为进一步降低检索开销，我们基于复杂问题检索逻辑的分解设计了三种检索范式：（1）**前向逐步推理** ——逐步扩展子查询；（2）**反向约束聚焦** ——从答案约束出发并缩小搜索范围；（3）**并行交集推理**——并行执行子搜索并求证据交集。通过在单轮内完成推理策略，该框架赋予模型清晰的规划、检索与知识综合能力，弥合推理深度与效率之间的差距。本文的贡献总结如下：

1. **一种免训练模块化推理框架：**我们提出了 FutureMind，这是一种四阶段流程（问题分析、逻辑推理、策略规划、检索引导），并辅以动态思考模块，为结构化推理提供明确的知识支持。该框架适用于大语言模型和小语言模型，兼顾准确性与效率，如图 1所示。

2. **可组合的检索策略：**我们设计了三种自适应检索范式（前向逐步推理、反向约束聚焦、并行交集推理），将复杂的多跳问题分解为易于处理的子查询，并高效地进行证据整合。

3. **系统性实验与认知洞见：** 在四个多跳问答基准上进行的实验表明，FutureMind 在不同架构和规模的模型上均能持续提升性能，其中小语言模型的提升幅度最大，并在免训练方法中达到新的最先进水平。此外，我们发现了“认知偏差瓶颈”：一旦教师模型的计划超出学生模型的能力，蒸馏就会产生信息损失，导致推理链断裂并放大噪声。这凸显了教师—学生兼容性比原始模型规模更为重要，并为设计轻量且可扩展的推理系统提供了指导。

2 相关工作

大语言模型与检索。大语言模型（Achiam 等人，2023；Team，2024）展现出强大的推理和代码生成能力（Guo 等人，2025；2024），但由于依赖静态的参数化知识，仍容易产生幻觉（Zhang 等人，2023）。为缓解这一问题，广泛采用外部搜索，主要有两种方式：（i）检索增强生成（RAG）（Gao 等人，2023），将检索到的证据整合到生成过程中；（ii）搜索即工具（Schick 等人，2023），大语言模型通过提示（Trivedi 等人，2022；Schick 等人，2023）或微调（Schick 等人，2023）与搜索引擎进行显式交互。然而，RAG 的静态单阶段检索——即不具备自适应能力的单次查找，忽略了查询复杂度和中间生成信号——可能会返回无关或信息量不足的段落，从而妨碍组合推理和多跳推理（Jin 等人，2024a）；基于工具的方法虽然交互性更强，但在复杂的多步推断中，仍难以检索到相关性和准确性都足够高的证据。

小型语言模型与认知迁移。小语言模型在成本敏感、低延迟和隐私保护应用中颇具吸引力，但在记忆、上下文传播以及结构化多步推理方面存在明显缺陷（Fu 等人，2023；Hsieh 等人，2023）。为弥合这一差距，认知迁移技术试图将推理行为从较大型模型迁移过来。思维链蒸馏可迁移逐步推理轨迹（Wei 等人，2022b；Wang 等人，2023；Fu 等人，2023），但适应性有限，且在分布偏移或风格偏移下可能较为脆弱。提示蒸馏通过提取紧凑提示来减少风格不匹配（Li 等人，2023a；Chen 和 Feng，2023），但通常主要编码静态知识模板，无法支持动态规划。检索增强迁移（如 Meta-RAG）借助外部知识提升效率（Mombaerts 等人，2024），但通常将检索视为固定的非自适应流水线，因而未能将检索与自适应推理充分整合。总体而言，这些方法仅部分缓解了小语言模型的推理缺陷，且缺乏稳健问题分解、迭代检索和多跳聚合所需的泛化能力与动态适应性——这促使人们研究能够为小语言模型赋予轻量级结构化推理流程和战略性检索策略的方法。

3 方法

3.1 概述

FutureMind 是一种模块化推理框架，采用**自适应知识蒸馏**将教师模型的结构化推理和检索策略迁移到学生模型中。与主要侧重于压缩知识表示的传统蒸馏方法不同，FutureMind 致力于蒸馏**系统化思维模式**。具体而言，它捕捉从问题定义到检索引导的完整逻辑链，并将这些模式提炼为轻量、可复用的战略性思维模式先验。这一设计使学生模型，尤其是小语言模型，能够进行自适应推理和深入、结构化的检索规划，从而即使在资源受限环境中也能实现更优的检索性能。

如图所示2，FutureMind 由**思维模块**协调，该模块会根据任务特征、数据可用性和效率动态生成最优检索策略

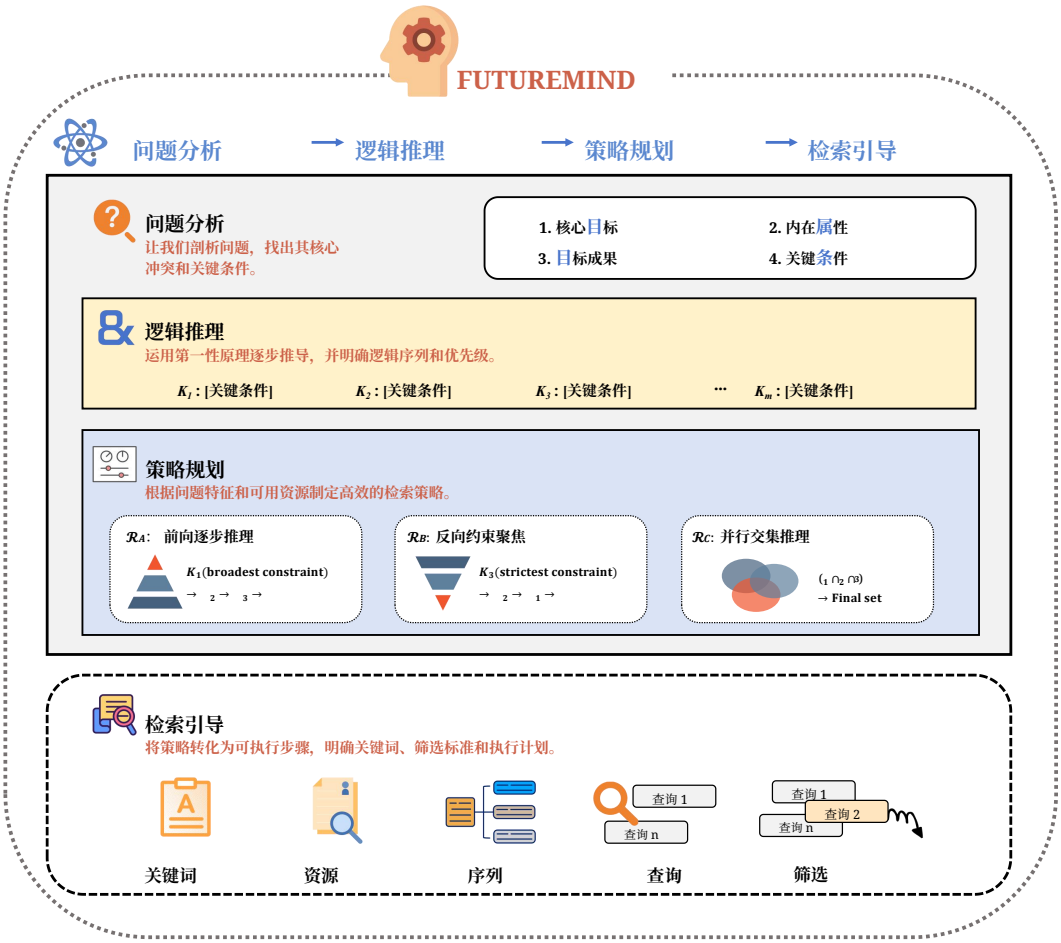

图2：FutureMind 框架概览。

约束条件。它由四个核心模块组成：**问题分析**、**逻辑推理**、**策略规划**和**检索引导**，实现模块化、可解释性和端到端优化。

形式上，FutureMind 是一个由思维模块 $\mathcal{M}$ 协调的四阶段流程：

$$F = \mathcal{M}\langle \mathcal{P}, \mathcal{L}, \mathcal{S}, \mathcal{R} \rangle, \tag{1}$$

其中， $\mathcal{P}$、 $\mathcal{L}$、 $\mathcal{S}$和 $\mathcal{R}$分别表示问题分析、逻辑推理、策略规划和检索引导。

为清晰起见并便于复现，我们在附录中进一步提供了各模块的指令与执行示例E.5.1- E.5.4，以说明各组件如何在整体框架中运作。

随后，我们全面概述FutureMind的四个核心模块。

3.2 模块定义

3.2.1 问题分析 $\mathcal{P}$

问题分析模块通过将输入查询x 分解为其基本组成部分来启动推理流程。该分解产生一种结构化表示，从而支持后续的推理与决策过程。具体而言，该模块识别以下关键要素：

$$\mathcal{P}(x) \rightarrow (\mathcal{O}, \mathcal{A}, \mathcal{T}, \mathcal{C}), \tag{2}$$

其中：

- $\mathcal{O}$ 表示核心目标，它们定义了问题求解过程的主要方向和预期结果。
- $\mathcal{A}$ 表示内在属性，用于描述问题固有的性质和条件。
- $\mathcal{T}$ 规定目标结果，指出问题解决后预期的结果或输出类型。
- $\mathcal{C} = \{C_1, C_2, \ldots, C_n\}$ 捕捉关键维度，即问题中的关键条件或因素，其中每个 C_i 代表一个特定维度。

通过将输入查询系统地分解为这些组成部分，问题分析模块为后续推理阶段奠定了清晰、结构化的基础。

3.2.2 逻辑推理 $\mathcal{L}$

逻辑推理模块采用第一性原理方法来推导核心问题机制。它识别最基本的原理，并运用演绎推理构建连贯的抽象。不同于启发式推理或基于类比的推理，该方法将推断建立在因果结构之上，减少对不完整先验知识的依赖。形式化表示为：

$$\mathcal{L}(\mathcal{O}, \mathcal{A}, \mathcal{T}, \mathcal{C}) \to (\mathcal{M}, \mathcal{K}), \tag{3}$$

其中：

- $\mathcal{O}$、 $\mathcal{A}$、 $\mathcal{T}$和 $\mathcal{C}$ 分别表示从问题分析中提取的核心目标、内在属性、目标结果和关键维度。
- $\mathcal{M}$ 表示机制理解，涵盖支配系统的因果关系和基本原理。
- $\mathcal{K} = \{K_1, K_2, \ldots, K_m\}$ 是关键条件的有序集合，依据逻辑依赖和判别重要性进行排序。

将推理建立在第一性原理之上，可以提升可解释性和适应性，尤其适用于复杂或知识密集型任务，因为在此类任务中，启发式方法往往不足。通过将问题分解为基本组成部分、识别关键条件，并按照逻辑优先级排列这些条件，该模块将问题分析中的结构化输入转换为机制抽象（$\mathcal{M}$， $\mathcal{K}$）。这一抽象为策略规划和检索引导奠定基础。

3.2.3 策略规划 $\mathcal{S}$

策略规划模块连接了逻辑推理得出的机制洞见与检索指导的规范层。它会根据机制理解 $\mathcal{M}$ 和优先级条件序列 $\mathcal{K} = \{K_1, K_2, \ldots, K_m\}$，动态确定最优检索策略 $\mathcal{R}^*$。形式化表示如下：

$$\mathcal{S}(\mathcal{M}, \mathcal{K}) \to \mathcal{R}^*, \qquad \mathcal{R}^* = \arg\min_{\mathcal{R} \in \mathcal{P}_{\text{cand}}} \mathcal{F}(\mathcal{R}; \mathcal{M}, \mathcal{K}), \tag{4}$$

其中：

- $\mathcal{M}$：从逻辑推理中得出的机制理解。
- $\mathcal{K} = \{K_1, K_2, \ldots, K_m\}$：按优先级排列的条件序列。
- $\mathcal{P}_{\text{候选}} = \{\mathcal{R}_A, \mathcal{R}_B, \mathcal{R}_C\}$：推理策略的候选池。
- $\mathcal{F}(\cdot)$：一种用于评估效率、约束条件、相互依赖关系和数据可用性的成本函数。在本作品中， $\mathcal{F}$ 特指教师模型为初步评估检索计划而实现的函数，旨在生成最优检索策略的建议。

基于知识密集型任务的洞见，我们设计了三种不同的检索范式（图 4），以实现更高效、准确的执行。StrategyPlanning 模块根据条件集拓扑 $\mathcal{K}$ 动态选择其中一种。

令 $\mathcal{U}$ 表示候选空间。对于每个条件 K_i，我们定义一个评估函数 $\phi(K_i,\ x) \in\{0,1\}$，用于检查候选项 $x \in \mathcal{U}$ 是否满足 K_i。中间集合定义如下：

$$X_i = \{x \in \mathcal{U} \mid \phi(K_i, x) = 1\}, \qquad X^* = \bigcap_{i=1}^{m} X_i. \tag{5}$$

策略 A：前向逐步推理（$\mathcal{R}_A$）适用于早期条件范围较广、但能有效缩小筛选范围的情况。约束条件按从一般到具体的顺序依次应用：

$$X_1 = \{x \in \mathcal{U} \mid \phi(K_1, x) = 1\},\ \ X_j = \{x \in X_{j-1} \mid \phi(K_j, x) = 1\},\ \ X^* = \bigcap_{i=1}^{m} X_i. \tag{6}$$

策略 B：反向约束聚焦（$\mathcal{R}_B$）适用于下游条件具有很强筛选性的情况。推理从最严格的约束条件开始，再逐步放宽：

$$X_m = \{x \in \mathcal{U} \mid \phi(K_m, x) = 1\},\ \ X_j = \{x \in X_{j+1} \mid \phi(K_j, x) = 1\},\ \ X^* = \bigcap_{i=1}^{m} X_i. \tag{7}$$

策略 C：并行交集推理（$\mathcal{R}_C$）最适用于相互独立或正交的条件。所有约束条件并行处理，然后取交集：

$$X_i = \{x \in \mathcal{U} \mid \phi(K_i, x) = 1\}, \qquad X^* = \bigcap_{i=1}^{m} X_i. \tag{8}$$

通过深入理解知识密集型问题，该模块能够自适应地选择最优检索策略，确保检索执行既高效又精准。

3.2.4 检索引导 $\mathcal{R}$

检索引导模块作为规范层，将抽象推理和所选检索策略转化为用于执行的结构化指令。与直接检索不同，该模块生成规定性指南，明确检索应如何执行。

给定机制理解 $\mathcal{M}$、优先条件 $\mathcal{K}$以及所选策略 $\mathcal{R}^*$，该模块输出一组检索指南：

$$\mathcal{R}(\mathcal{M}, \mathcal{K}, \mathcal{R}^*) \to \Gamma, \tag{9}$$

其中 $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_q\}$ 是一组指导检索过程的规范原则（例如，优先级顺序、来源偏好、评估标准）。

该引导被结构化为五个关键且互补的阶段：

- **关键词引导**。 从 $\mathcal{K}$ 中提取核心实体、属性和关系，并指定检索应优先考虑的词汇与语义变体。该引导概述了查询可以变化的维度，从而在保持与底层推理结构一致的同时，实现跨领域的自适应检索。
- **资源引导**。 指明按相关性 $\mathcal{M}$ 和可信度排序的信息来源类别（例如，学术数据库、行业报告、政策文件），从而将检索导向可靠的知识领域。
- **序列指导**。 根据 $\mathcal{R}^*$ 对检索步骤的顺序提出建议。例如，前向逐步策略先从范围广、召回率高的存储库入手，建立初步覆盖，然后再转向领域特定集合。相反，后向约束策略从筛选性很强的监管资源入手，以高精度证据为搜索锚点，再根据需要向外扩展。
- **查询引导**。 为查询构式提供结构模板（例如布尔模式、语义扩展、层次约束），强调初始检索的包容性以及后续阶段的逐步收窄。该引导提供可调整的设计原则，而非固定查询字符串。

- **筛选引导**。 界定评估检索结果的原则，包括其与 $\mathcal{M}$的相关性、来源可信度以及方法严谨性。该模块从概念上规定评估标准。

通过围绕关键词、资源、排序、查询构式和评估来组织指导，该模块在保持执行独立性的同时，将认知策略与检索衔接起来。

4 实验

4.1 实验设置

数据集。 我们在四个广泛使用的多跳问答（QA）基准上评估我们的方法：**2WikiMultihopQA (2WikiMQA)**（Ho 等人，2020a），**Bamboogle**（Press 等人，2022），**MuSiQue**（Tang 与 Yang，2024），以及 **FRAMES**（Krishna 等人，2024）。具体而言，我们从 2WikiMQA 和 MuSiQue 的验证集中随机抽取 500 个实例，同时在 Bamboogle 和 FRAMES 的完整测试集上进行评估。

指标。 在评估中，遵循先前作品（Sun 等人，2025年），我们采用两个互补的指标：**基于覆盖率的精确匹配（ACC$_E$）**和**大语言模型评审（ACC$_L$）**。ACC$_E$ 衡量预测答案是否完全覆盖标准答案，同时允许释义变体；其详细计算公式见附录 G.1。相比之下，ACC$_L$ 使用 GPT-4o-mini 作为自动评估器，以判断预测答案相对于标准答案的语义正确性。ACC$_L$ 的完整评估提示见附录 F.1。

基线方法。 我们考虑三类基线方法：（1）**朴素生成**：在不进行检索的情况下生成答案。（2）**标准检索增强生成**（Zhao 等人，2024）：使用原始问题作为查询来检索文档。（3）**Search-o1**（Li 等人，2025年a）：使用提示执行自主发起的检索。

实现细节。 我们评估所提出的 **FutureMind**方法的有效性，使用不同架构和规模的模型（Qwen-2.5-3B/7B/14B/32B/72B-Instruct 和 Llama3.1- 8B-Instruct）。在生成时，我们将最大序列长度设置为 32768 个 token，温度设为= 0.0，top-p 设为 = 0.8，top-k 设为 = 20，重复惩罚设为 = 1.05，并应用于所有模型。在检索时，我们采用 Google Web Search API，检索排名前 k = 10的结果。在实验中，FutureMind 利用增强版的 **工具调用（TC）**，即一种 ReAct风格编排框架 *。我们将原始 TC 框架的单次搜索过程替换为并行搜索，从而更高效、更稳健地聚合检索证据。该配置称为 **TC+FM**。实现细节见附录 E。

4.2 主要结果

表 1 比较了不同模型架构和规模在四个多跳问答基准上的表现，涵盖四种方法：朴素生成、标准检索增强生成、Search-o1 和基于工具调用的 FutureMind。结果呈现出几项关键发现。受篇幅限制，更多基准测试结果见附录 6 中的 B表。

1. **基线方法的固有限制**。 朴素生成（仅依赖内部知识）的准确率最低，凸显了其整合外部证据的能力不足。标准检索增强生成（启用检索）无法可靠地进行多步推理；若推理整合失败，其表现甚至可能不如朴素生成。Search-o1（基于文档进行推理）提升了检索质量，但仍有局限：小型模型获益甚微，而较大型模型也受制于其内在的摘要和整合能力。

2. **自适应知识蒸馏对 FutureMind 的有效性**。 与 Search-o1 的固定提示设计不同，FutureMind 采用自适应知识，使问题求解过程更加灵活有效，并带来稳定的性能提升。例如，Qwen-3B 在 Frames 上的表现有所提升（ACC$_E$: 11.77 → 18.84），Llama3.1-8B 在 Bamboogle 上的表现有所提升（ACC$_L$: 52.00 → 64.00），Qwen-72B 在 2WikiMQA 上的表现有所提升（ACC$_L$: 75.40 → 80.60）。通过提供自适应外部

*https://github.com/QwenLM/Qwen-Agent/

表 1：四个多跳问答基准上的主要结果。**粗体** 表示最佳性能，下划线表示第二佳。蓝色背景的行对应我们的方法TC+FM*，其中 FM* 为每个基础模型选择性能最佳的 FutureMind 增强变体。

模型	方法	2WikiMQA		Bamboogle		Frames		MuSiQue		AVG	
		ACC_E	ACC_L	ACC_E	ACC_L	ACC_E	ACC_L	ACC_E	ACC_L	ACC_E	ACC_L
Qwen-3B	朴素生成	16.80	17.20	20.80	24.00	5.94	8.98	3.60	4.60	11.79	13.70
	标准检索增强生成	24.00	24.40	26.40	38.40	12.01	19.17	10.20	13.00	18.15	23.74
	Search-o1	41.00	41.80	34.40	39.20	11.77	18.81	10.40	12.60	24.39	28.10
	TC+FM*	**56.40**	**43.80**	**39.20**	**43.20**	**18.84**	**19.42**	**14.20**	**15.20**	**32.16**	**30.41**
Qwen-7B	朴素生成	29.40	25.20	34.40	37.60	11.29	16.87	7.60	10.80	20.67	22.62
	标准检索增强生成	30.20	29.80	42.40	52.80	15.78	24.76	13.20	16.80	25.39	31.04
	Search-o1	57.80	59.80	43.20	51.20	24.63	**38.34**	**20.80**	**23.80**	36.61	43.29
	TC+FM*	**62.00**	**64.00**	**58.40**	**64.80**	**25.12**	34.71	20.00	23.80	**41.38**	**46.83**
Qwen-14B	朴素生成	30.40	30.80	48.80	55.20	14.81	22.82	8.80	12.40	25.70	30.30
	标准检索增强生成	27.40	28.40	44.80	56.00	17.96	28.40	14.00	18.60	26.04	32.85
	Search-o1	66.80	68.40	43.20	55.20	30.46	46.48	20.60	25.60	40.27	48.92
	TC+FM*	**71.60**	**75.20**	**70.40**	**72.80**	**34.83**	**49.51**	**24.00**	**28.20**	**50.21**	**56.43**
Qwen-32B	朴素生成	30.80	31.30	54.40	60.80	15.66	24.51	10.80	15.20	27.91	32.95
	标准检索增强生成	24.60	24.40	52.80	61.60	19.78	30.95	16.20	19.60	28.35	34.14
	Search-o1	68.60	71.60	60.80	67.20	34.34	**54.12**	22.80	27.80	46.63	55.18
	TC+FM*	**74.40**	**77.80**	**68.80**	**72.80**	**37.15**	53.86	**26.00**	**30.40**	**51.59**	**58.71**
Qwen-72B	朴素生成	38.20	38.60	60.00	67.20	21.12	32.16	12.80	18.40	33.03	39.09
	标准检索增强生成	31.00	31.40	59.20	67.20	25.97	37.62	19.00	23.80	33.79	40.01
	Search-o1	72.60	75.40	67.20	72.80	37.37	56.67	24.60	30.80	50.44	58.92
	TC+FM*	**74.20**	**80.60**	**75.20**	**79.20**	**41.38**	**58.59**	**28.40**	**36.60**	**54.80**	**63.75**
Llama3.1-8B	朴素生成	38.20	38.60	**60.00**	**67.20**	21.12	32.16	12.80	18.40	33.03	39.09
	标准检索增强生成	29.20	30.40	39.20	47.20	15.05	22.82	12.20	15.20	23.91	28.90
	Search-o1	54.00	56.00	46.40	52.00	24.88	37.62	15.40	18.20	35.17	40.95
	TC+FM*	**55.20**	56.80	**58.40**	64.00	**27.43**	**39.92**	**21.80**	**25.20**	**40.71**	**46.48**

表 2：教师模型规模对多跳问答基准测试中学生性能的影响。**粗体**表示最佳性能，下划线表示次佳性能。蓝色背景的行对应每种学生模型的最佳教师模型规模。

模型	方法	2WikiMQA		Bamboogle		Frames		MuSiQue		Avg	
		ACC_E	ACC_L	ACC_E	ACC_L	ACC_E	ACC_L	ACC_E	ACC_L	ACC_E	ACC_L
Qwen-3B	TC	54.20	42.40	37.60	40.00	17.96	**22.94**	13.00	12.60	30.69	29.49
	TC+FM（3B）	42.00	30.80	28.00	30.40	11.53	10.32	7.40	8.60	22.23	20.03
	TC+FM（7B）	53.00	39.60	30.40	36.00	15.78	16.26	11.20	11.40	27.60	25.82
	TC+FM（14B）	55.20	**45.60**	**40.80**	42.40	17.11	18.46	12.20	**12.80**	**31.33**	**29.82**
	TC+FM（32B）	49.20	37.00	36.00	37.60	12.26	12.01	10.20	10.40	26.92	24.25
	TC+FM（72B）	**56.40**	43.80	39.20	**43.20**	**17.84**	19.42	**14.20**	15.20	31.91	30.41
Qwen-7B	TC	56.80	56.20	49.60	54.40	23.78	32.28	16.40	18.80	36.65	40.42
	TC+FM（3B）	60.20	60.00	49.60	50.40	23.09	30.83	17.00	19.60	37.97	40.21
	TC+FM（7B）	60.20	61.80	53.60	57.60	24.39	33.55	17.60	21.20	38.95	43.04
	TC+FM（14B）	**62.00**	**64.00**	**58.40**	**64.80**	25.12	**34.71**	**20.00**	**23.80**	**41.38**	**46.83**
	TC+FM（32B）	57.80	57.60	52.00	58.40	22.57	29.98	15.20	19.40	36.89	41.35
	TC+FM（72B）	60.40	60.00	56.80	61.60	**26.58**	**34.71**	18.20	21.20	40.50	44.38

策略赋能，而非仅依赖内部能力，FutureMind 实现了更强且更具可扩展性的推理，尤其是在资源受限的环境中。

3. **FutureMind 在不同模型架构和规模下的普适性**。 TC+FM*几乎在所有设置中都取得了最先进的成果，在不同模型架构和参数规模下均带来可扩展的提升。这表明 FutureMind 能通过外部策略迁移，广泛有效地增强多跳推理能力，缓解资源受限模型的能力瓶颈，同时在较大型模型中保持强劲表现。

表3：多跳问答基准上模块化组件的消融研究。**粗体** 表示启用所有模块时的性能。

模型	方法	2WikiMQA		Bamboogle		Frames		MuSiQue		Avg	
		ACC_E	ACC_L	ACC_E	ACC_L	ACC_E	ACC_LACC_E	ACC_LACC_E		ACC_L	
Qwen-7B	所有模块	**62.00**	**64.00**	**58.40**	**64.80**	**25.12**	**34.71**	**20.00**	**23.80**	**41.38**	**46.83**
	- 无问题分析	58.20	57.40	56.00	62.40	24.93	34.57	16.00	21.00	38.78	43.84
	- 不含逻辑推理	60.40	59.00	49.60	53.60	24.59	34.21	19.20	23.00	38.45	42.45
	- 不含策略规划	56.40	57.20	49.30	51.20	22.63	33.10	17.60	19.80	36.48	40.33
	- 不含检索引导	59.20	60.40	57.60	60.00	23.39	33.34	18.40	20.60	39.65	43.59

表 4：多跳问答基准上检索策略的消融研究。**粗体** 表示从 Qwen-14B 教师模型蒸馏全部三种检索策略后的性能。

模型	方法	2WikiMQA		Bamboogle		Frames		MuSiQue		Avg	
		ACC_E	ACC_LACC_E	ACC_L	ACC_LACC_E	ACC_E	ACC_L	ACC_E	ACC_LACC_E	ACC_L	
Qwen-7B	所有策略	**62.00**	**64.00**	**58.40**	**64.80**	**25.12**	**34.71**	**20.00**	**23.80**	**41.38**	**46.83**
	- 无策略 A	57.40	57.80	54.20	60.00	24.64	32.16	16.60	21.20	38.21	42.79
	- 无策略 B	58.80	58.00	57.60	61.60	25.09	34.47	18.40	22.60	39.97	44.67
	- 无策略 C	60.40	59.00	54.40	60.80	24.72	33.37	17.40	21.40	39.23	43.64

4.3 教师模型设计对师生认知对齐的影响 NT

我们系统分析了知识蒸馏中教师模型设计对学生性能的影响（表 2）。若干一致的规律随之显现：

1. **小规模教师模型会降低学生性能**。在 TC+FM 中，为 Qwen-3B 使用 3B 教师模型会降低平均性能（ACC_E：30.69 → 22.23，ACC_L：29.49 → 20.03），表明低容量教师模型可能产生噪声或误导性规划信号，从而阻碍迁移效果。

2. **中等规模的教师模型提供最佳对齐效果**。两个学生模型从 14B 教师模型中获益最多（Qwen-3B：ACC_E：31.33，ACC_L：29.82；Qwen-7B：ACC_E：41.38，ACC_L：46.83），其表现优于 32B 版本，平均而言与 72B 模型相当或更佳。

3. **认知兼容性比原始规模更重要**。尽管 72B 教师模型在某些子任务上表现出色，但其平均表现并未始终超过 14B 教师模型（Qwen-7B 学生模型：72B 教师模型 ACC_E：40.50，ACC_L：44.38 < 14B 教师模型 ACC_E：41.38，ACC_L：46.83），这表明，教师模型与学生模型之间的认知对齐，对知识蒸馏效果的影响比原始规模更为关键。

“认知偏差瓶颈”进一步表明，过于复杂的教师模型规划可能无法将推理能力传递给更小的学生模型，因为策略信息丢失可能破坏关键推理链或放大噪声。因此，在知识蒸馏过程中，优先考虑教师模型与学生模型之间的兼容性，比关注原始模型规模更为重要。未来的工作应系统量化规划质量，并评估其在不同任务和对齐策略上的泛化能力，以确保轻量级模型具备可扩展的推理能力。

除了规模之外，教师模型的架构、推理方向和指令微调也必须与学生模型相兼容。附录中的实验B 表明，架构与学生模型高度匹配的教师模型能够持续实现更有效的思维模式蒸馏，从而提升多跳推理性能。

4.4 消融研究

我们首先评估 FutureMind 四个核心模块的贡献：问题分析、逻辑推理、策略规划和检索引导。如表 3所示，移除任何一个模块都会导致性能明显下降，证实了它们具有互补作用。其中，策略规划的影响最大，凸显了它在将结构化推理转化为有效检索行动方面的核心作用。

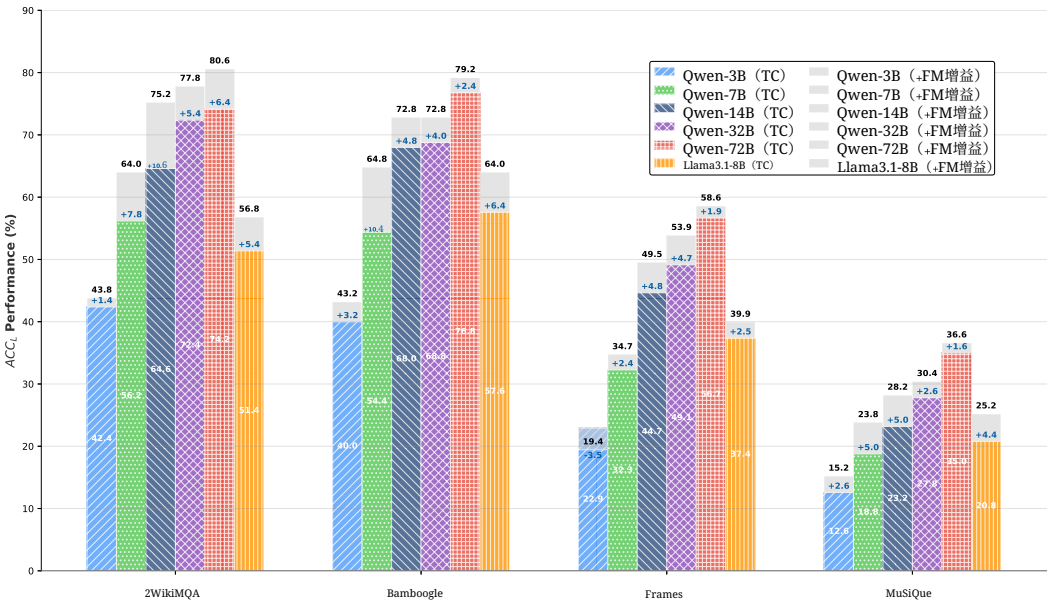

图 3：不同模型上增强型工具调用的消融研究。

接下来，我们分析策略规划模块中的三种检索策略。表 4显示，移除任何一种策略——前向逐步推理（$\mathcal{R}_A$）、反向约束聚焦（$\mathcal{R}_B$）或并行交集推理（$\mathcal{R}_C$）——都会降低性能。总体而言，$\mathcal{R}_A$ 最为关键，$\mathcal{R}_C$ 在条件彼此独立的数据集上贡献更大，而 $\mathcal{R}_B$ 则有助于处理以约束为重点的案例。

最后，我们通过比较单独使用增强型工具调用与使用工具调用驱动的 FutureMind，评估 FutureMind 的整体影响。图 3 显示，移除 FutureMind 模块会持续降低性能，这证明该模块在协调多跳推理和检索规划方面发挥着不可或缺的作用。详细数值结果见表 7。

这些消融结果共同表明，FutureMind及其模块和检索策略构建了一个有效且连贯的推理框架。

5 结论

我们提出了**FutureMind**，一种免训练模块化推理框架，使大型和小型语言模型都能够进行高效、准确且结构化的推理。FutureMind 将推理分解为**问题分析**、**逻辑推理**、**策略规划**和**检索引导**，并利用自适应检索策略，明确指导何时、检索什么以及如何检索证据，从而有效平衡推理深度与效率。

在四个多跳问答基准上的评估表明，不同模型规模与架构均取得了稳定提升，其中小语言模型的改进最为显著。我们进一步发现，师生思维模式蒸馏中存在认知偏差瓶颈：过于复杂的教师模型计划可能超出学生能力，导致策略信息丢失并削弱推理。有效的蒸馏既需要规模，也需要架构对齐，因为结构兼容且推理方向一致的教师模型能够实现更有效的策略迁移。

总而言之，FutureMind 表明，即使对于小语言模型，也可以实现结构化的自适应推理，通过战略性思维模式先验将其转化为具备认知能力的智能体。

伦理声明

我们严格遵守 ICLR 道德准则。本作品仅使用四个公开可用的多跳问答数据集，不涉及任何个人或敏感信息，也未招募人类受试者。本研究不存在利益冲突或可预见的伦理风险。除标准多跳问答任务范围内的问题外，我们的研究未引入其他伦理方面的担忧。

可复现性声明

为确保研究结果可复现，我们将在论文被接收后公开发布所有实验代码和数据处理脚本。本文使用的四个数据集均可公开获取（例如 2WikiMultihopQA、MuSiQue、Bamboogle、Frames），附录详细介绍了数据预处理、检索配置和相关超参数。正文对模型架构、推理流程和评估指标进行了充分说明，以便其他研究人员复现我们的结果。

参考文献

Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat 等。GPT-4 技术报告。*arXiv* 预印本 *arXiv:2303.08774*，2023。

Salaheddin Alzubi、Creston Brooks、Purva Chiniya、Edoardo Contente、Chiara von Gerlach、Lu-cas Irwin、Yihan Jiang、Arda Kaz、Windsor Nguyen、Sewoong Oh 等。Open deep search：借助开源推理智能体，让搜索惠及大众。*arXiv* 预印本 *arXiv:2503.20201*，2025年。

Guangji Bai、Zheng Chai、Chen Ling、Shiyu Wang、Jiaying Lu、Nan Zhang、Tingwei Shi、Ziyang Yu、Mengdan Zhu、Yifei Zhang 等。超越效率：资源高效型大语言模型的系统综述。*arXiv* 预印本 *arXiv:2401.00625*，2024。

Feng Chen 和 Yujian Feng。用于多模态命名实体识别和多模态关系抽取的思维链提示蒸馏。*arXiv* 预印本 *arXiv:2306.14122*，2023。

Yao Fu、Hao Peng、Litu Ou、Ashish Sabharwal 与 Tushar Khot。使较小语言模型专门面向多步推理。载于国际机器学习会议，第10421–10430页。PMLR，2023。

Yunfan Gao、Yun Xiong、Xinyu Gao、Kangxiang Jia、Jinliu Pan、Yuxi Bi、Yi Dai、Jiawei Sun、Haofen Wang 与 Haofen Wang。面向大语言模型的检索增强生成：综述。*arXiv*预印本 *arXiv:2312.10997*，2，2023。

Daya Guo、Qihao Zhu、Dejian Yang、Zhenda Xie、Kai Dong、Wentao Zhang、Guanting Chen、Xiao Bi、Yu Wu、YK Li 等。Deepseek-coder：当大语言模型遇见编程——代码智能的兴起。*arXiv*预印本 *arXiv:2401.14196*，2024。

Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang、Xiao Bi 等。Deepseek-r1：通过强化学习激发大语言模型的推理能力。*arXiv* 预印本 *arXiv:2501.12948*，2025年。

Xanh Ho、Anh-Khoa Duong Nguyen、Saku Sugawara 和 Akiko Aizawa。构建用于全面评估推理步骤的多跳问答数据集。*arXiv* 预印本 *arXiv:2011.01060*，2020a。

Xanh Ho、Anh-Khoa Duong Nguyen、Saku Sugawara 和 Akiko Aizawa。构建用于全面评估推理步骤的多跳问答数据集。载于第*28*届国际计算语言学会议论文集，第6609–6625页，2020b。

Cheng-Yu Hsieh、Chun-Liang Li、Chih-Kuan Yeh、Hootan Nakhost、Yasuhisa Fujii、Alexander Ratner、Ranjay Krishna、Chen-Yu Lee 和 Tomas Pfister。逐步蒸馏！以更少的训练数据和更小的模型规模超越更大的语言模型。*arXiv* 预印本 *arXiv:2305.02301*，2023。

Bowen Jin、Jinsung Yoon、Jiawei Han 和 Sercan O Arik。长上下文大语言模型与检索增强生成相遇：克服检索增强生成中长输入带来的挑战。发表于第十三届国际学习表征会议，2024a。

Haolin Jin、Linghan Huang、Haipeng Cai、Jun Yan、Bo Li 和 Huaming Chen。《从大语言模型到用于软件工程的大语言模型智能体：现状、挑战与未来综述》。*arXiv* 预印本 *arXiv:2408.02479*，2024b。

Satyapriya Krishna、Kalpesh Krishna、Anhad Mohananey、Steven Schwarcz、Adam Stambler、Shyam Upadhyay 和 Manaal Faruqui。《事实、检索与推理：对检索增强生成的统一评估》。*arXiv* 预印本 *arXiv:2409.12941*，2024。

Lei Li、Yongfeng Zhang 和 Li Chen。《面向高效基于大语言模型的推荐的提示蒸馏》。载于第 *32* 届 *ACM* 信息与知识管理国际会议论文集，第 1348–1357 页，2023a。

Xiaoxi Li、Guanting Dong、Jiajie Jin、Yuyao Zhang、Yujia Zhou、Yutao Zhu、Peitian Zhang 和 Zhicheng Dou。Search-o1：通过智能体式搜索增强的大型推理模型。*CoRR*，abs/2501.05366，2025a。doi：10.48550/ARXIV.2501.05366。URLhttps://doi.org/10.48550/arXiv.2501.05366。

Xiaoxi Li、Guanting Dong、Jiajie Jin、Yuyao Zhang、Yujia Zhou、Yutao Zhu、Peitian Zhang 和 Zhicheng Dou。Search-o1：通过智能体式搜索增强的大型推理模型。*arXiv* 预印本 *arXiv:2501.05366*，2025b。

Xiaoxi Li、Jiajie Jin、Guanting Dong、Hongjin Qian、Yutao Zhu、Yongkang Wu、Ji-Rong Wen 和 Zhicheng Dou。Webthinker：赋予大型推理模型深度研究能力。*CoRR*，abs/2504.21776，2025c。doi：10.48550/ARXIV.2504.21776。URLhttps://arxiv.org/abs/2504.21776。

Yinheng Li、Shaofei Wang、Han Ding和Hang Chen。金融中的大语言模型：综述。载于第四届*ACM*金融人工智能国际会议论文集，第374–382页，2023b。

Laurent Mombaerts、Terry Ding、Adi Banerjee、Florian Felice、Jonathan Taws和Tarik Borogovac。面向检索增强大语言模型的元知识。*arXiv*预印本 *arXiv:2408.09017*，2024。

OpenAI。GPT-4技术报告。*CoRR*，abs/2303.08774，2023。doi: 10.48550/ARXIV.2303.08774。URL https://doi.org/10.48550/arXiv.2303.08774。

Cheng Peng、Xi Yang、Aokun Chen、Kaleb E Smith、Nima PourNejatian、Anthony B Costa、Cheryl Martin、Mona G Flores、Ying Zhang、Tanja Magoc等。面向医学研究与医疗保健的生成式大语言模型研究。*NPJ*数字医学，6(1):210，2023。

Ofir Press、Muru Zhang、Sewon Min、Ludwig Schmidt、Noah A Smith 和 Mike Lewis。衡量并缩小语言模型中的组合性差距。*arXiv* 预印本 *arXiv:2210.03350*，2022。

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J. Liu。探索采用统一文本到文本转换器的迁移学习极限。机器学习研究杂志，21:140:1–140:67，2020。网址：https://jmlr.org/papers/v21/20-074.html。

Timo Schick、Jane Dwivedi-Yu、Roberto Dessì、Roberta Raileanu、Maria Lomeli、Eric Hambro、Luke Zettlemoyer、Nicola Cancedda 和 Thomas Scialom。Toolformer：语言模型可以自学使用工具。神经信息处理系统进展，36:68539–68551，2023。

Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、YK Li、Y Wu 等。Deepseekmath：推动开放语言模型中数学推理的极限。*arXiv* 预印本 *arXiv:2402.03300*，2024。

Zhongxiang Sun、Qipeng Wang、Weijie Yu、Xiaoxue Zang、Kai Zheng、Jun Xu、Xiao Zhang、Yang Song 和 Han Li。Rearter：结合可信过程奖励的检索增强推理。载于第*48*届国际*ACM SIGIR*信息检索研究与发展会议，第1251–1261页，2025年。

Yixuan Tang和Yi Yang。Multihop-RAG：面向多跳查询的检索增强生成基准评测。*arXiv*预印本 *arXiv:2401.15391*，2024。

Gemini团队。Gemini 1.5：在数百万词元的上下文中解锁多模态理解。*arXiv*预印本 *arXiv:2403.05530*，2024。

Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、Baptiste Rozière、Naman Goyal、Eric Hambro、Faisal Azhar、Aurélien Rodriguez、Armand Joulin、Edouard Grave和Guillaume Lample。Llama：开放且高效的基础语言模型。*CoRR*，abs/2302.13971，2023。doi: 10.48550/ARXIV.2302.13971。URL https://doi.org/10.48550/arXiv.2302.13971。

Harsh Trivedi、Niranjan Balasubramanian、Tushar Khot 和 Ashish Sabharwal。将检索与思维链推理交替用于知识密集型多步问题。*arXiv* 预印本 *arXiv:2212.10509*，2022。

Zhongwei Wan、Xin Wang、Che Liu、Samiul Alam、Yu Zheng、Jiachen Liu、Zhongnan Qu、Shen Yan、Yi Zhu、Quanlu Zhang、Mosharaf Chowdhury 和 Mi Zhang。高效大语言模型：综述。*CoRR*，abs/2312.03863，2023。doi：10.48550/ARXIV.2312.03863。网址：https://doi.org/10.48550/arXiv.2312.03863。

Fali Wang、Zhiwei Zhang、Xianren Zhang、Zongyu Wu、Tzuhao Mo、Qiuhao Lu、Wanjing Wang、Rui Li、Junjie Xu、Xianfeng Tang 等。大语言模型时代的小型语言模型综合综述：技术、增强方法、应用、与大语言模型的协作及可信性。*ACM*智能系统与技术汇刊，2024a。

Lei Wang、Chen Ma、Xueyang Feng、Zeyu Zhang、Hao Yang、Jingsen Zhang、Zhiyuan Chen、Jiakai Tang、Xu Chen、Yankai Lin等。基于大语言模型的自主智能体综述。计算机科学前沿，18(6):186345，2024b。

Peifeng Wang、Zhengyang Wang、Zheng Li、Yifan Gao、Bing Yin和Xiang Ren。Scott：自洽思维链蒸馏。*arXiv*预印本 *arXiv:2305.01879*，2023。

Zeyu Wang。CausalBench：用于评估大语言模型因果推理能力的综合基准。见第*10*届*SIGHAN*中文信息处理研讨会（*SIGHAN-10*）论文集，第143–151页，2024。

Jason Wei、Maarten Bosma、Vincent Y. Zhao、Kelvin Guu、Adams Wei Yu、Brian Lester、Nan Du、Andrew M. Dai和Quoc V. Le。微调语言模型是零样本学习器。见第十届国际学习表征会议，*ICLR 2022*，线上活动，*2022*年*4*月*25*日至*29*日。OpenReview.net，2022a。网址https://openreview.net/forum?id=gEZrGCozdqR。

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou 等。思维链提示能够激发大语言模型的推理能力。神经信息处理系统进展，35:24824–24837，2022b。

Guangzhi Xiong, Qiao Jin, Xiao Wang, Yin Fang, Haolin Liu, Yifan Yang, Fangyuan Chen, Zhixing Song, Dengyu Wang, Minjia Zhang 等。RAG-Gym：通过过程监督优化推理与搜索智能体。*arXiv* 预印本 *arXiv:2502.13957*，2025年。

Weihong Xu, Haein Choi, Po-kai Hsu, Shimeng Yu 和 Tajana Simunic。SLIM：一种通过自适应阈值化实现稀疏大语言模型边缘推理的异构加速器。*ACM*嵌入式计算系统汇刊，2025年。

Meng Yang、Yihao Wang 和 Yu Gu。面向常识问答的基于语言的推理图神经网络。《神经网络》，181:106816，2025年。

Zhilin Yang、Peng Qi、Saizheng Zhang、Yoshua Bengio、William Cohen、Ruslan Salakhutdinov 和 Christopher D Manning。Hotpotqa：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，2018 年。

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》。发表于国际学习表征会议（*ICLR*），2023。

Yue Zhang、Yafu Li、Leyang Cui、Deng Cai、Lemao Liu、Tingchen Fu、Xinting Huang、Enbo Zhao、Yu Zhang、Yulong Chen 等。《人工智能海洋中的塞壬之歌：大语言模型幻觉综述》。*arXiv* 预印本 *arXiv:2309.01219*，2023。

Zihan Zhang、Meng Fang 和 Ling Chen。《Retrievalqa：评估短文本开放域问答的自适应检索增强生成》。载于 Lun-Wei Ku、Andre Martins 和 Vivek Srikumar（编），计算语言学协会研究成果，*ACL 2024*，泰国曼谷及线上会议，*2024*年*8*月*11*日至*16*日，第 6963–6975 页。计算语言学协会，2024。doi：10.18653/V1/2024.FINDINGS-ACL.415。网址：https://doi.org/10.18653/v1/2024.findings-acl.415。

Penghao Zhao、Hailin Zhang、Qinhan Yu、Zhengren Wang、Yunteng Geng、Fangcheng Fu、Ling Yang、Wentao Zhang 和 Bin Cui。检索增强生成在人工智能生成内容中的应用：一项综述。*CoRR*，abs/2402.19473，2024。doi：10.48550/ARXIV.2402.19473。网址：https://doi.org/10.48550/arXiv.2402.19473。

大语言模型（LLM）的使用

在撰写本文的过程中，我们使用大语言模型（LLM）作为辅助工具，仅用于润色论文的语法和表达，以提升其规范性和可读性。论文的研究构想、核心内容、实验设计和结论均由作者独立完成，大语言模型未参与研究或创作过程中的任何实质性工作。

B 补充实验结果

如表 5 所示，在 TC+FM 框架下，教师模型的架构对学生模型的适配效果有决定性影响：在所有规模（3B–72B）上，Qwen2.5-72B 在 ACC_E 和 ACC_L方面均持续优于 Llama3.1-70B。这一规律表明，架构对齐是多跳推理中实现有效认知迁移的关键。

C 工具设计

为实现高效、准确的检索驱动推理，我们采用 Qwen-Agent 中实现的 ReAct 框架，并开发了两个互补工具。第一个是 **并行搜索工具** C.1，它基于 Google 搜索 API，能够执行并行检索并做出检索决策，从而实现高效搜索。第二个是 **FutureMind 工具** C.2，这是一个由更大型语言模型实例化的模块，能够增强结构化推理和检索逻辑，从而实现更精准、更有针对性的检索。这些组件共同将复杂的检索过程与推理逻辑解耦，从而高效、准确地执行检索，进而促进更好的证据整合并得出更准确的最终答案。

表 5：教师模型架构对学生性能的影响。我们在不同的教师模型架构下评估学生模型（3B–72B），具体使用 Qwen2.5-72B-Instruct（记作 Q2.5）和 Llama3.1-70B-Instruct（记作 L3.1），并采用工具调用驱动的 FutureMind（TC+FM）。**粗体**表示最佳性能，下划线表示次佳性能。蓝色背景的行对应每个学生模型的最佳教师模型架构。

模型	方法	2WikiMQA		Bamboogle		Frames		MuSiQue		Avg	
		ACC_E	ACC_L	ACC_E	ACC_L	ACC_E	ACC_L	ACC_E	ACC_L	ACC_E	ACC_L
Qwen-3B	TC	54.20	42.40	37.60	40.00	17.96	22.94	13.00	12.60	30.69	29.49
	TC+FM (L3.1)	47.80	34.00	27.20	27.20	11.53	9.83	10.20	10.40	24.18	20.36
	TC+FM（Q2.5）	56.40	43.80	39.20	43.20	17.84	19.42	14.20	15.20	**31.91**	**30.41**
Qwen-7B	TC	56.80	56.20	49.60	54.40	23.78	32.28	16.40	18.80	36.65	40.42
	TC+FM（L3.1）	53.60	50.20	48.80	51.20	19.54	26.58	14.20	17.20	34.03	36.30
	TC+FM (Q2.5)	57.80	57.60	52.00	58.40	22.57	29.98	15.20	19.40	**36.89**	**41.34**
Qwen-14B	TC	64.40	64.60	64.00	68.00	34.22	44.66	20.40	23.20	45.76	50.11
	TC+FM（L3.1）	63.40	63.80	63.60	67.20	28.16	37.86	22.80	25.20	44.49	48.52
	TC+FM (Q2.5)	70.00	71.60	64.00	68.00	35.92	48.06	25.80	28.20	**48.93**	**53.96**
Qwen-32B	TC	71.20	72.40	66.40	68.80	36.28	49.15	24.60	27.80	49.62	54.54
	TC+FM（L3.1）	63.40	61.40	66.40	68.00	28.79	38.74	23.20	26.20	45.45	48.59
	TC+FM（Q2.5）	74.40	77.80	68.80	72.80	34.15	47.86	26.00	30.40	**50.84**	**57.21**
Qwen-72B	TC	71.60	74.20	68.80	76.80	40.04	56.67	27.40	35.00	51.96	60.67
	TC+FM（L3.1）	67.40	69.80	69.60	76.00	33.98	47.33	23.20	29.60	48.54	55.68
	TC+FM (Q2.5)	74.20	80.60	75.20	79.20	41.38	58.59	27.40	34.60	**54.55**	**63.25**

表 6：四个多跳问答基准上的补充结果。

模型	方法	2WikiMQA		Bamboogle		Frames		MuSiQue		AVG	
		ACC_E	ACC_L	ACC_E	ACC_L	ACC_E	ACC_L	ACC_E	ACC_L	ACC_E	ACC_L
Qwen-14B	TC	64.40	64.60	64.00	68.00	34.22	44.66	20.40	23.20	45.76	50.12
	TC+FM（3B）	68.40	70.00	58.40	64.00	32.89	43.33	22.80	27.00	45.62	51.08
	TC+FM（7B）	69.00	71.20	63.20	72.00	33.37	46.84	20.60	24.40	46.54	53.61
	TC+FM（14B）	71.60	75.20	70.40	72.80	34.83	49.51	24.00	28.20	50.21	56.43
	TC+FM（32B）	71.80	74.40	64.80	72.00	35.25	46.24	22.20	27.20	48.51	54.96
	TC+FM（72B）	70.00	71.60	64.00	68.00	35.92	48.06	25.80	28.20	48.93	53.96
Qwen-32B	TC	71.20	72.40	66.40	68.80	36.28	49.15	24.60	27.80	49.62	54.54
	TC+FM（3B）	70.40	70.00	60.80	67.20	32.77	44.53	22.60	24.80	46.64	51.63
	TC+FM（7B）	71.80	73.20	61.60	67.80	34.83	45.14	22.90	25.80	47.28	53.49
	TC+FM（14B）	74.60	76.40	70.40	75.20	35.19	47.33	23.20	29.80	50.85	57.18
	TC+FM（32B）	69.60	70.00	64.00	73.60	31.10	42.89	21.00	24.60	46.43	52.77
	TC+FM（72B）	74.40	77.80	68.80	72.80	37.15	53.86	26.00	30.40	51.59	59.71
Qwen-72B	TC	71.60	74.20	68.80	76.80	40.04	56.67	27.40	35.00	51.96	60.67
	TC+FM（3B）	71.20	75.20	65.60	72.00	38.23	53.64	24.60	29.40	49.91	57.56
	TC+FM（7B）	72.80	76.80	71.20	76.00	37.13	55.70	24.60	30.40	51.93	59.73
	TC+FM（14B）	71.00	73.40	74.40	77.60	35.56	50.12	23.60	30.00	51.14	57.78
	TC+FM（72B）	74.20	80.60	75.20	79.20	41.38	58.59	28.40	36.60	54.80	63.75

C.1 并行搜索工具

并行搜索工具（Google）介绍

名称： 并行_搜索（Google）

描述：
当用户的查询属于以下任一类别时，你应调用 Parallel_Search 工具（google）：
1. 你的内部知识库和训练数据不足以准确回答问题。2. 用户询问某个具体示例、产品或信息，而你可以通过网络获取更详细的内容。

表 7：不同模型上的增强型工具调用消融研究。

方法	模型	2WikiMQA		Bamboogle		Frames		MuSiQue		Avg	
		ACC_E	ACC_L	ACC_E	ACC_L	ACC_E	ACC_L	ACC_E	ACC_L	ACC_E	ACC_L
工具调用	Qwen-3B	54.20	42.40	37.60	40.00	17.96	22.94	13.00	12.60	30.69	29.49
	Qwen-7B	56.80	56.20	49.60	54.40	23.78	32.28	16.40	18.80	36.65	40.42
	Qwen-14B	64.40	64.60	64.00	68.00	34.22	44.66	20.40	23.20	45.76	50.11
	Qwen-32B	71.20	72.40	66.40	68.80	36.28	49.15	24.60	27.80	49.62	54.54
	Qwen-72B	71.60	74.20	68.80	76.80	40.04	56.67	27.40	35.00	51.96	60.67
	Llama3.1-8B	51.20	51.40	55.20	57.60	24.51	37.38	17.40	20.80	37.08	41.80

3. 问题涉及最新数据、动态信息，或训练截止时间之后的知识，需要实时更新。4. 答案存在于你无法直接访问的外部知识源中；你必须搜索才能获取。5. 虽然你掌握该主题的一般知识，但在线搜索能提供更详细或更新的信息（例如当前流行的行话或热门话题）。6. 你遇到不熟悉的术语或概念，为避免编造信息，必须通过搜索工具核实。7. 你需要查阅产品手册或官方规范来支持回答。

搜索工具支持并行和顺序查询：1. 如果多个搜索彼此独立，可以并行发起。2. 如果查询相互依赖（即需要按顺序执行），则依次进行。

参数：

```
{
  "type": "object",
  "properties": {
    "查询": {
      "type": "数组",
      "项目": { "类型": "字符串" },
      "描述": (
        "搜索关键词列表："
        "- 并行搜索：一次提供多个关键词；"
        "- 迭代搜索：提供一个单元素数组。"
      ),
      "examples": [
        {
          "queries": [
            "小米SU7 Ultra 官方价格",
            "特斯拉Model S 最新价格"
          ]
        },
        {
          "queries": ["Mishi wolffin鱼命名者"]
        }
      ]
    }
  },
  "required": ["查询"]
}
```

C.2 FUTUREMIND 工具

FutureMind 工具介绍

名称： FutureMind

描述：
收到查询后，首先调用 FutureMind 工具，以获取**针对该问题的系统化思维模式和解决方案路线图**。

对于检索类问题：1. 它会生成结构化的问题解决工作流和检索策略。2. 它会明确 delineate 从“问题定义”经由“条件分解”到“结论推导”的逻辑链。3. 它会提供可执行的搜索序列和组合查询条件，作为检索引导，从而提高信息获取效率并确保检索准确性。

参数：

{ "type": "object", "properties": { "query": { "type": "string", "description": "需要进行系统化问题分析并制定检索策略的查询" } }, "required": ["query"]}

D 工具调用模板

工具调用模板

工具
您可以调用一个或多个函数来协助处理用户查询。

提供给你的函数签名位于 <tools></tools> XML 标签中：

<tools>
{tool_descs}
</tools>
对于每次函数调用，请返回一个包含函数名称和参数的 JSON 对象，位于 <tool_call></tool_call> XML 标签中：

<工具 –调用>
{"name": <函数名>, "arguments": <参数 JSON 对象>}
</工具 –调用>

E FUTUREMIND：描述和实现细节

E.1 FutureMind 概述

如图 2所示，FutureMind是一个轻量级的免训练推理框架，可将系统化思维模式从教师模型迁移到更小的学生模型，经由**自适应思维-**

模式蒸馏。它将任务分解为四个分阶段模块——问题分析、逻辑推理、策略规划与检索引导——并输出简洁、可审计的计划，明确是否检索、检索什么以及如何整合证据。为降低检索开销，该框架支持三种可组合范式（图4）：前向逐步推理、反向约束聚焦与并行交集推理。总体而言，FutureMind使资源受限模型能够进行结构化的低延迟检索与推理，并提升准确性与可解释性。

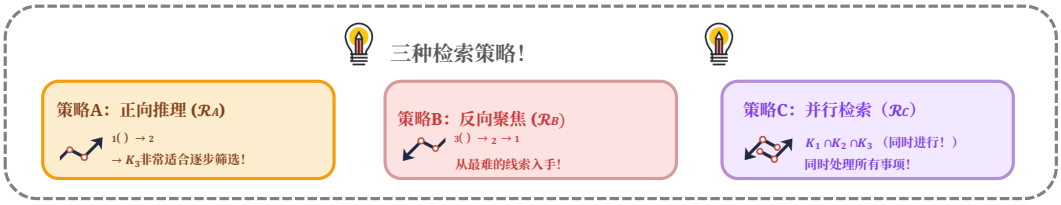

图4：FutureMind 中采用的三种自适应检索范式。

E.2 指令用于增强工具调用使用并行搜索工具C.1

使用并行搜索工具增强工具调用的指令

系统：
你是一名乐于助人的助手。

工具：
你可以调用一个或多个函数来协助处理用户查询。 你将获得函数签名，位于 <工具></工具> XML 标签内：

```
<工具>
        {
            "type": "function",
            "函数":
                {
                    "name": "google_search",
                    "描述": "并行搜索工具的
                        描述",
                    "参数": "并行搜索工具的
                        参数",
                }
        }
</tools>}
```

对于每次函数调用，返回一个包含函数名称和参数的 JSON 对象 <tools></tools>XML 标签：

<工具 –调用>
{"name": <函数名>, "参数": <参数 JSON 对象>}
</工具 –调用>

用户：
问题：{user question}
你首先将推理过程作为内心独白进行思考，然后提供最终答案。推理过程必须包含在 <think> </think> 标签内。最终答案必须放在 <answer> </answer> 标签内。

E.3 指令 工具调用-驱动的 未来思维 使用 并行搜索工具 C.1

使用并行搜索工具的工具调用驱动的 FutureMind 指令

系统：
你是一位乐于助人的助手。

工具：
你可以调用一个或多个函数来协助处理用户查询。我们在 <tools></tools> XML 标签中提供了函数签名：

```
<tools>
        {
            "type": "function",
            "function":
                {
                    "name": "并行搜索" ,
                    "description": "描述的并行搜索工具
                        描述",
                    "parameters": "参数的并行搜索工具
                        参数",
                }
        }
</工具>}
<工具>
        {
            "type": "function",
            "function":
                {
                    "name": "futuremind" ,
                    "description": "Futuremind 工具的
                        描述",
                    "parameters": "FutureMind 工具的参数",
                }
        }
</tools>}
```

对于每次函数调用，请在<tools></tools>XML 标签内返回一个包含函数名称和参数的 JSON 对象：

<工具调用>_{ "name": <函数名>, "arguments": <
参数 JSON 对象>}</tool_调用>

用户：
问题：{user question}
你首先将推理过程作为内心独白进行思考，然后给出最终答案。推理过程必须置于 <think> </think> 标签内。最终答案必须放在 <answer> </answer> 标签内。首先，调用“futuremind”工具，针对给定问题（问题原文逐字保留）获取系统化的思维策略与解题路线图。之后，你可以根据需要多次调用搜索工具来收集或核实信息，直到拥有足以作答的材料。一旦确认所有必要信息，请使用简洁、聚焦的语言给出最终答案，避免不必要的赘述。

E.4 相关方法：增强型工具调用（TC）和工具调用驱动的 FutureMind（TC+FM）

增强型工具调用（TC）。 增强型工具调用（TC）在 ReAct 框架之上实现，并遵循工具调用模板 D。为支持高效的信息获取，我们采用**并行搜索工具** C.1，它支持并行检索，同时将推理过程限制在最多 10 步。该设计兼顾效率与可控性。推理过程遵循所提供的预定义指令，见 E.2。

工具调用驱动的 FutureMind（TC+FM）。工具调用驱动的 FutureMind（TC+FM）通过将**FutureMind（FM）方法**融入增强型工具调用框架，扩展了这一设计。为了更高效、稳健地聚合检索到的证据，我们采用**并行搜索工具**C.1和**FutureMind 工具**C.2，从而实现并行检索，同时将推理过程限制在最多 10 个步骤内。这一设计在保持推理可控的同时，进一步提升了检索鲁棒性。推理过程遵循E.3中提供的预定义指令。

E.5 FutureMind 模块：问题分析—逻辑推理—策略规划—检索引导

本节首先介绍各模块使用的指令，然后通过一个具体示例进行详细说明。具体而言，我们采用如下查询来逐步展示 FutureMind 框架中各组件的执行过程：“一位画家的父亲死于心脏病，他有一个姐姐，并在婚姻破裂前与妻子育有五个孩子。之后，他又经历了三段恋情。后来，有人根据此人创作了一部文学作品。那部作品的标题是什么？”

E.5.1 问题分析模块

问题分析模块指令

模块名称：问题分析**指令**：请首先识别问题的核心冲突、内在属性和目标结果；提取关键维度 $C_1, C_2, C_3, C_4, \ldots$（例如：因果关系、约束条件、目标函数）。

问题分析模块执行示例

1. 核心目标
精确识别满足指定传记约束的画家，并报告有文献记载、以该画家为基础的唯一文学作品（标题及作者），同时附上简明的支持性证据。

2. 内在属性
需核查的可验证、原子化事实：(a) 职业 = 画家；(b) 父亲的死因 = 心脏病；(c) 恰好有一位姐妹；(d) 与妻子育有五个孩子；(e) 这段婚姻后来解体；(f) 此后经历了三段彼此 distinct 的重要恋爱关系；(g) 存在一部与此人有明确关联的著名文学作品。

3. 目标结果
输出：(1) 该文学作品的标题和作者；(2) 一份简明的证据映射（3 个要点），说明上述哪些内在属性（按标签列出）得到了将这位画家与该作品联系起来的权威来源的佐证。

4. 关键条件
将四个关键条件定义如下：

- **C_1 （身份／职业）：** “画家”——将候选范围限定为艺术家／画家。

• $\mathbf{C}_2$ **（家庭／人生事件）：** 父亲死于心脏病；有一个姐妹；与妻子育有五个孩子；两人的婚姻后来破裂。这些都是鲜明且可检索的生平事实。

• $\mathbf{C}_3$ **（恋爱史）：** 后来经历了三段重要的恋爱关系（有多位伴侣），这可能意味着曾搬迁或有跨文化生活经历（类似案例中较为常见，但文中并未明确说明）。

• $\mathbf{C}_4$ **（文学改编线索）：** 一部文学作品以此人为原型创作——目标是该作品的标题（及其作者），且这部作品很可能广为人知。

E.5.2 逻辑推理模块

逻辑推理模块指令

模块名称：逻辑推理**指令**：运用第一性原理逆向推导底层逻辑，确定关键条件要素（记为 $\mathbf{K_1}, \mathbf{K_2}, \mathbf{K_3}, \mathbf{K_4}$），并按重要性或逻辑顺序排列。

逻辑推理模块执行示例

运用第一性原理进行分解，梳理约束条件并确定最高效的求解序列。

• $\mathbf{K_1}$：与妻子育有五个孩子，最终婚姻破裂。这些非常具体的人生经历可以排除大多数画家。

• $\mathbf{K_2}$：此后有过多段（约三段）重要的恋爱关系。结合$\mathbf{K_1}$，候选范围便大幅缩小。

• $\mathbf{K_3}$：父亲死于心脏病（更多时间和地点细节能提高区分度）。可作为辅助佐证。

• $\mathbf{K_4}$：明确指出其是某部著名文学作品的原型或灵感来源，为最终验证提供了关键依据。

首先搜索满足$\mathbf{K_1}$的个人；并行验证$\mathbf{K_2}$ 与$\mathbf{K_3}$；最后确认其与一部著名文学作品（$\mathbf{K_4}$）的关联，以识别唯一匹配。

E.5.3 策略规划模块

策略规划模块指令

模块名称：策略规划**指令**：思维策略（根据情况选择一项或多项）。

策略 A：F orward 逐步推理（渐进式 Filte 环）

• 按从基础条件到决定性条件的顺序逐步筛选：a. 找出满足基础条件 $\mathbf{K_1}$ 的集合 A；b. 从 A 中筛选出满足次要条件 $\mathbf{K_2}$ 的子集 B；c. 从 B 中筛选出满足关键条件 $\mathbf{K_3}$ 的子集 C；d. 在 C 中验证满足决定性条件 $\mathbf{K_4}$ 的候选项。

策略B：反向约束聚焦（反向缩小范围）

• 从严格约束反向推导至基础兼容性：a. 优先检索满足最严格约束的集合 A（例如，$\mathbf{K_3}$），以快速缩小范围；

b. 从 A 中筛选满足关键特征条件的集合 B（例如，$\mathbf{K_4}$）；c. 从 B 中验证满足前置条件的集合 C（例如，$\mathbf{K_2}$）；d. 在 C 中确认与基线条件兼容的最终候选项（例如，$\mathbf{K_1}$）。

策略 C：并行交集推理（并行筛选）
• 将条件视为相互独立的维度并行处理：a. 分别检索满足条件$\mathbf{K_1}$、$\mathbf{K_2}$、$\mathbf{K_3}$、$\mathbf{K_4}$ 的集合 A、B、C、D；b. 计算交集（$A \cap B \cap C \cap D$），提取同时满足所有条件的解。

根据问题特征、数据可用性和效率要求选择策略（或组合）。

策略规划模块执行示例

已选策略：策略 B — 反向约束聚焦（推荐）R
ationale: The "li文学作品的d改编ion" condition ($\mathbf{K_4}$) is the most powerful and least ambiguous constraint. Starting from it enables the fastest, most discriminative narrowing of the search space, since only a few artists have been explicitly used as namesakes for famous literary works.

策略B：反向约束聚焦（反向缩小范围）
• 从最严格的约束（$\mathbf{K_4}$）进行反向推导：a. 从$\mathbf{K_4}$开始：检索或列举明确基于真实画家或受其启发的知名文学作品。b. 对于每一部已识别的文学作品，提取被记载或声称作为原型的画家。c. 对于每一位候选画家，验证其与$\mathbf{K_1}$（职业）、$\mathbf{K_2}$（家庭／生活事件）以及$\mathbf{K_3}$（恋爱史）的传记兼容性。d. 剔除任何未能满足一项或多项条件的候选者；若无候选者留存，则返回步骤（a）并扩大文学作品集合。e. 满足$\mathbf{K_1}$–$\mathbf{K_3}$ 且匹配$\mathbf{K_4}$ 的画家即为最终解。

仅应用策略 B。从最强约束（文学改编）出发，识别候选作品及其原型，然后按$\mathbf{K_1}$–$\mathbf{K_3}$ 进行筛选，以获得唯一匹配。

E.5.4 检索指导模块

检索引导模块指令

模块名称：检索引导**指令**：思维策略（根据需要选择一项或多项）。

关键词引导
从问题分析中提取核心实体、属性和关系（例如："人工智能 +伦理风险 +监管政策"）。

资源引导
列出与该领域相关的权威来源：学术数据库、行业报告、政府机构或标准机构等（例如 PubMed、IEEE Xplore、世界银行数据）。

序列指导
按优先级或逻辑设计检索序列（例如，先行业标准，然后实证研究，最后政策文件）。

查询引导
Formulate Layer 1 组合搜索表达以支持迭代筛选：例如，("人工智能伦理风险" AND "监管政策") OR ("自动驾驶安全标准" AND "欧盟法规")；或者先用关键词 A 检索，再用关键词 B 进一步筛选结果。

筛选引导
定义初步的纳入/排除标准和相关性评估，以支持后续深入分析。

检索引导模块执行示例

1. 关键词设计（分层）

- 第一级（直接定位作品/原型）：“以画家为原型的小说”、“受画家启发的小说”
- 第二级（结合传记细节）：“画家有五个孩子、与妻子离婚”、“画家的父亲死于心脏病”、“画家的姐妹”
- 第三级（核实人物与作品的组合）：“<作品标题>原型”、“<画家姓名>被描绘为 <作品标题>”

2. 资源选择（优先级）

- 权威传记与学术出版物（艺术家传记、艺术史书籍）
- 文学研究与注释（关于小说原型的研究论文、作者回忆录）
- 公认的百科全书与数据库（例如，初步筛选时使用维基百科，最终参考专业传记）
- 文化与历史期刊、作家研究及书评（用于核实作品以真实人物为原型的说法）

3. 检索顺序（推荐）

- 首先检索知名小说或以画家为原型的作品，以生成符合文学改编这一关键条件的候选作品
- 检查这些作品的注释、作者注释和学术评论，以确定与之相关的原型画家
- 查阅候选画家的传记，核实子女数量、婚姻状况、父亲的死因和恋爱史等传记细节
- 通过独立权威来源进行交叉验证，最终确认

4. 查询示例（促进迭代筛选）

- “受画家启发的小说”或“以画家为原型的小说”
- “画家 AND 五个孩子 AND 离异妻子”
- “’<作品标题>’ 原型 <画家姓名>” （用于核实已知候选对象）

5. 筛选标准（初步纳入／排除）

- 纳入：传记、学术论文、权威书评和作者陈述等可靠来源；多个独立来源引用相同的原型和事实
- 排除：未经核实的博客或没有来源的二手说法；缺乏原始证据的循环引用；传记细节与关键事实相矛盾的候选人（例如子女数量、婚姻状况）

F 指令模板

F.1 评估指令

在本作品中，我们使用大语言模型评审（GPT-4o-mini）来评估多跳问答（QA）基准：2WikiMultihopQA、Bamboogle、MuSiQue 和 FRAMES。具体指令如下。

评审指令

给定一个问题及其标准答案，验证预测答案是否正确。若预测与标准答案的含义和关键信息完全一致，则该预测正确。若预测正确，回答 True，否则回答 False。

[问题：] **{user question}**[标准答案]
{reference answer}[预测答案] **{assistant'**
s answer}

F.2 朴素生成指令

朴素生成指令

请回答以下问题。你应该逐步思考并解决问题。最终答案必须放在 <answer> </answer> 标签中。

[问题：] **{user question}**

F.3 标准检索增强生成指令

标准检索增强生成指令

你是一位知识渊博的助手，会利用提供的文档准确回答用户问题。

指南：
- 分析提供的文档，提取相关信息。综合这些信息，形成连贯且准确的回答。- 确保回答直接针对用户问题，并使用文档中的信息。

[问题：] **{user question}**[
文档：] **{documents}**

F.4 Search-o1 指令

Search-o1 指令

你是一名推理助手，能够执行网络搜索，以帮助你准确回答用户的问题。你拥有专用工具：要执行搜索，请写入 <|begin_search_query|> 在此处输入你的查询 <|end_search_query|>。随后，系统将搜索并分析相关网页，然后以以下格式向你提供有用信息： <|begin_search_result|> ……搜索结果…… <|end_search_result|>。如有必要，你可以多次重复搜索过程。搜索尝试的最大次数限制为 {MAX_SEARCH_LIMIT}。获取所需的全部信息后，继续进行推理。

示例：问题：“.....” 助手思考步骤：- 我可能需要查找有关.....的详细信息 助手：<|begin_search_query|>.....<|end_search_query|>（系统返回从相关网页中处理的信息）助手根据新信息继续推理.....请记住：- 使用 <|begin_search_query|> 来请求网络搜索，并以 <|end_search_query|>结束。- 完成搜索后，继续你的推理。

G 公式描述

G.1 ACC_E公式说明

公式：

$$\text{ACC}_E = \begin{cases} 1 & \exists g \in G, \text{norm}(pred) \supseteq \text{norm}(g) \\ 0 & \text{otherwise} \end{cases}$$

其中，$pred$是预测答案；$G = \{g_1, \ldots, g_k\}$是标准答案集合；norm(·)表示归一化过程，用于处理释义改写、标点符号及相关表层变体；关系 $A \supseteq B$ 表示 A 完全涵盖了 B的核心语义。因此，当归一化后的预测在语义上涵盖至少一个归一化后的标准答案时，$\text{ACC}_E = 1$；否则为 0。

H GPT-4o-mini 和 Google 搜索 API 的总体令牌消耗与 API 成本

H.1 大型模型评估成本分析

对于 GPT-4o-mini 模型，评分过程中消耗的词元总数可分为三个主要部分：评分提示F.1、问题、标准答案、预测答案以及 GPT-4o-mini 模型的输出（True/False）。

我们对不同数据集的输入 token 用量进行了统计分析，如表8所示。四个数据集（2WikiMQA、Bamboogle、Frames 和 MuSiQue）的输入令牌数总计约为 772,000 个，而输出令牌数总计约为 2,000 个。

根据 GPT-4o-mini 的官方定价——每百万个输入令牌 0.15 美元，每百万个输出令牌 0.60 美元——所有数据集完成一次实验设置的 API 成本估计约为 0.117 美元。我们总共进行了 84 次这样的实验运行，因此 GPT-4o-mini 评分过程的整体评估成本约为 9.828 美元。

表 8：GPT-4o-mini 在单一实验设置下于不同数据集上的总 token 消耗量。所有数值均以*10,000* 个 *token*（*w*）为单位报告。

数据集	**消耗的词元（w）**
2WikiMQA	16.5 w
Bamboogle	3.7 w
Frames	38.0 w
MuSiQue	19.0 w
总计	**77.2 w**

表 9：在基线和 FutureMind 设置下，每个教师–学生模型对跨数据集的搜索查询总数。

数据集	Qwen-72B → 3B		Qwen-32B → 3B		Qwen-14B → 3B		Qwen-7B → 3B	
	TC	TC+FM	TC	TC+FM	TC	TC+FM	TC	TC+FM
Wiki	1032	1297	1028	1120	987	1322	1005	1259
Bamboogle	208	235	203	184	196	286	202	214
Frames	2030	1923	2000	1210	1693	2031	2016	1492
MuSiQue	1056	1217	1032	821	1021	821	1040	997
总计	4326	4672	4263	3335	3897	4460	4263	3962

H.2 搜索工具调用的成本分析

此成本估算仅关注增强型工具调用方法和工具调用驱动的 *FutureMind* 方法所产生的直接费用。对于这两种方法，大部分成本来自并行搜索工具所进行的 API 调用。根据 Custo m Google Search JSON API 的官方定价，每 1,000 次查询收费 5 美元。

在四个数据集（*2WikiMQA*、*Bamboogle*、*Frames*和*MuSiQue*）中，增强型工具调用和工具调用驱动的 FutureMind 方法平均各执行约 4.2k 次搜索引擎查询。因此，使用任一方法在这四个数据集上执行所有实验的预计成本约为 21 美元。考虑到实验运行总数，与搜索相关的累计支出预计约为 1,365 美元。

此外，我们提供一个具体示例来说明这一估算。在此案例中，学生模型为 *Qwen2.5-3B*，教师模型分别为 *Qwen2.5-72B*、*Qwen2.5-32B*、*Qwen2.5-14B*和*Qwen2.5-7B*。详细统计数据见表 9，其中报告了在 增强型工具调用（TC）和工具调用驱动的 *FutureMind*（TC+FM）设置下，各教师–学生模型对在不同数据集上的搜索查询总数。

I 讨论：FUTUREMIND 在纠正逻辑推理路径方面的灵活性

本节讨论 **FutureMind** 如何灵活地纠正和完善逻辑推理路径。这项能力体现了 FutureMind 的核心设计目标之一：实现动态自适应推理，在结构化规划与实时灵活性之间取得平衡。

FutureMind 并非一次性生成固定计划，而是被设计为一种**可动态调用的工具**（详见附录 E.3），可根据需要反复调用。这种设计在保留结构化规划能力的同时，确保推理过程具备灵活性，从而支持更具适应性的推理策略。其核心机制包括：

- 按需触发。当学生模型检测到候选空间过大、检索噪声过高或中间状态无效等问题时，可以按需调用 FutureMind。在这些情况下，FutureMind 会提供经过完善或替代性的推理策略，以指导后续探索。
- 策略多样性与去中心化控制。FutureMind 提供动态且多样的推理策略，而推理路径的最终选择和调整则由学生模型根据实时检索结果和可用计算预算作出，从而保持灵活性。

总体而言，这种交互构成了一个由**执行瓶颈 → 按需调用 → 策略优化**组成的闭环自适应循环，映照出人类专家迭代式的问题解决逻辑。

示例说明。以下案例通过一个知识密集型推理问题展示了这一机制：谁是《百年孤独》作者的朋友，同时也是诺贝尔奖得主的诗人?

- **首次调用 FutureMind**。 在初步探索期间，模型意识到，检索“朋友”可能会产生大量重复候选项。因此，它首次调用 FutureMind，并获得了**前向逐步推理策略**：确定作者→ 检索知名文学界友人 → 筛选诗人 → 核实诺贝尔奖得主身份。
- **第二次调用 FutureMind**。 执行过程中，即使将“友人”的范围缩小，仍有数十个候选项，这表明后续推理难度较高。模型第二次调用 FutureMind，并获得了建议的**反向约束聚焦策略**：列出诺贝尔奖得主中的诗人 → 筛选有文献记载与加西亚·马尔克斯存在交往的候选人 →通过可靠来源确认友谊。
- 收到这一指导后，模型将第二步细化为“筛选与加西亚·马尔克斯有公开可查交往记录的候选人”，随后继续执行。它首先确定**巴勃罗·聂鲁达**（1971年诺贝尔文学奖得主、诗人）是核心候选人，并通过马尔克斯的文章《**加西亚·马尔克斯论巴勃罗·聂鲁达**》确认了两人长期的友谊。最终，通过查阅文学档案**聂鲁达**，得出了正确答案。

J 案例研究

我们通过一个案例研究，展示所提出的 FutureMind 方法 2的完整流程。该案例选自 MuSiQue 基准测试，其中的问题几乎无法由 Qwen2.5-7B-Instruct 模型单独正确回答。接着，我们展示工具调用驱动的 FutureMind 如何使用 Qwen2.5-14B-Instruct 作为教师模型，引导 Qwen2.5-7B-Instruct 解决该问题，并给出推导出最终答案的完整推理轨迹。

MuSiQue 中的案例轨迹

问题： 加泰罗尼亚自治区政府宫是在位于西班牙、Berdejo 所在地区中名为 Martin 的男子去世的那座城市里，何时建成的?

步骤 1

<思考>
我对这个问题不太熟悉，因此我会先调用 FutureMind 工具，获取系统化的思维模式和路线图。**</think>**

<工具调用>
{"name": "FutureMind", "arguments": {"query": ["加泰罗尼亚自治区政府宫是在西班牙贝尔德霍所在地区的马丁去世的那座城市中，何时建成的？"]}}
</工具调用>

<工具_响应>
1. **问题分析** - 确定加泰罗尼亚自治区政府宫的建造日期。该宫位于马丁去世的城市，而马丁来自西班牙贝尔德霍所在的地区。

2. **逻辑推理** - 关键条件要素
   K1（正确识别 Martin 及其所在的西班牙地区）K2（Martin 与死亡城市之间的联系）K3（Berdejo 在西班牙的位置）K4（所识别城市中 Palau delaGeneralitat 的建造日期）

3. **策略规划**- 选择策略 A（前向逐步推理） a. 检索西班牙境内 Berdejo 所在地区的 Martin 信息（K1） b. 根据这些信息，确定 Martin 去世的城市（K2）

c. 验证 Berdejo 与西班牙境内所确定地区之间的联系（K3） d. 在所确定的城市中，查找 Palau de la Generalitat 的建造日期（K4）

4. **检索引导**

关键词： 西班牙 [地区的Martin],Berdejo、加泰罗尼亚自治区政府宫、建造date

资源： 维基百科、不列颠百科全书、西班牙历史档案馆、所识别地区和城市的地方政府网站

序列： 历史传记 → 地区信息 → 地方历史记录→ 建筑或政府网站

查询： “西班牙 [地区的Martin]传记“；“Berdejo在西班牙的位置“；“Palau 位于所识别城市的加泰罗尼亚自治区政府宫建造日期

筛选： 排除无关的传记或历史记录；仅限包含可核实建造日期信息的来源。

</工具_响应>

步骤 2

<思考>

要回答这个问题，目标是确定位于某座城市的加泰罗尼亚自治区政府宫的建造日期。这座城市是 Martin 去世的地方；Martin 来自西班牙的一个地区，而 Berdejo 就位于该地区。关键步骤如下：1. 首先，收集有关 Martin 以及 Berdejo 所在地区的信息，以便确定 Martin 去世的城市。2. 然后，核实 Berdejo 所在地与所确定的西班牙地区之间的关联。3. 确定 Martin 去世的城市后，查找该市加泰罗尼亚自治区政府宫的建造日期。选定的策略是前向逐步推理，推理过程包括检索历史和地区资料，然后查找加泰罗尼亚自治区政府宫的建筑信息。下一步是使用以下查询搜索相关资料。**</思考>**

<工具调用>_

{"name": "google_search", "arguments": {"query": ["Martin的 [西班牙地区]传记", "Berdejo在西班牙的位置", "加泰罗尼亚自治区政府宫建造日期"]}}**</工具调用>_**

<工具_响应>

Martin（1356年生于赫罗纳，加泰罗尼亚 [西班牙]——卒于1410年5月31日，巴塞罗那）自1395年起担任阿拉贡国王，并自1409年起担任西西里国王（称Martin二世）（https://www.britannica.com/biography/Martin-king-of-Aragon-and-Sicily）

2024年3月2日至14日，瓦伦西亚自治区政府宫敞开大门，邀请您探索这座设有主席府的历史建筑（https://www.visitvalencia.com/en/events-valencia/palau-de-la-generalitat-opens-its-doors-celebrate-fallas）

...

</工具_响应>

步骤 3

<思考>

搜索结果显示如下：1. 西班牙不同地区有几位名叫 Martin 的人物，但阿拉贡的马丁一世（1392年至1409年）似乎最相关。不过，还需要进一步了解他的逝世地点。2. Berdejo 是西班牙阿拉贡萨拉戈萨省的一个城镇。3. 巴塞罗那的加泰罗尼亚自治区政府宫可追溯至15世纪。鉴于 Martin 一世与 Berdejo 之间缺乏直接关联，我们将进一步уточ细搜索范围，重点查明 Martin 一世去世的城市，以及该城市与 Berdejo 可能存在的联系。下一条搜索查询为：阿拉贡的马丁一世逝世地点。**</think>**

<工具_调用>
{"name": "google_search", "arguments": {"query": ["阿拉贡的马丁一世逝世地点"]}}</tool call>

<工具_响应>
Matino是阿拉贡国王马丁一世之子。在1390年，Matino迎娶了西西里女王兼表妹玛丽亚，并成为共同国王。Matino于1409年去世后，他的父亲成为西西里国王，称为 Matino（https://www.wikiwand.com/zh-cn/articles）
...
</工具_响应>

第 4 步

<思考>
阿拉贡的马丁一世于1410年5月31日在巴塞罗那去世。问题询问的是建造巴塞罗那加泰罗尼亚自治区政府宫的日期，马丁去世于此，并将其与 Berdejo 联系起来，位于萨拉戈萨，阿拉贡。由于Berdejo处于不同地区（阿拉贡），而Palau de la Generalitat位于巴塞罗那（加泰罗尼亚），我们需要关注Palau de la Generalitat在巴塞罗那的建造。该建筑于15世纪完工，而由于Martin一世于1410年去世，该建筑当时已经建成。
</think>

<答案>位于西班牙巴塞罗那的加泰罗尼亚自治区政府宫建于15世纪，远早于阿拉贡的马丁一世于1410年去世。 <答案> ✓