量化预算约束的智能体大语言模型搜索中设计决策对准确率与成本的影响

Kyle A. McCleary 和 James M. Ghawaly计算机科学与
工程系，路易斯安那州立大学，美国路易斯安那州巴吞鲁日
jghawaly@lsu.edu，kmccl24@lsu.edu

摘要

智能体检索增强生成（RAG）系统结合了迭代搜索、规划提示词和检索后端，但实际部署环境会对工具调用次数和生成令牌施加明确限制。我们开展了一项受控测量研究，考察在固定约束下，搜索深度、检索策略和补全预算如何影响准确率与成本。我们使用预算约束智能体搜索（*BCAS*）这一与模型无关的评估工具，它会显示剩余预算并限制工具使用；基于此，我们在六个大语言模型和三个问答基准上进行了比较。在不同模型和数据集上，增加搜索次数能在达到较小上限前提升准确率；在消融实验网格中，混合词法与稠密检索并结合轻量级重排序带来的平均收益最大；而较大的补全预算对 HotpotQA 式综合任务最有帮助。这些结果为配置预算约束下的智能体检索管道提供了实用指导，并附有可复现的提示词和评估设置。

关键词：检索增强生成、智能体搜索、预算感知评估

arXiv:2603.08877v1 [cs.AI] 9 Mar 2026

1. 引言

检索增强生成（RAG）的格局在短短一年内发生了重大变化。最初的静态检索与生成流水线，已演变为复杂的智能体系统：大语言模型（LLM）能够自主规划并执行多步骤检索策略（Singh 等人，2024）。如今，智能体式检索能力已集成到主流 AI 工具中。OpenAI 的 ChatGPT 采用函数调用进行迭代搜索（OpenAI，2023），Anthropic 的 Claude 提供专门的研究模式（Anthropic，2024），而 Google 的 Gemini 2.0 则内置工具使用 API，可动态收集信息（Gemini Team，2025）。

然而，尽管这些进展主要致力于提升检索质量和答案准确率，一个关键维度仍未得到充分探索：计算预算约束。在实际部署中，多次 API 调用的成本、连续搜索的延迟，以及处理大量检索文档所带来的计算开销，都构成了重大挑战。

在这种设定下，我们将问题界定为一个测量缺口。检索增强生成如今已成为为大语言模型提供依据的默认策略，许多系统会在“智能体搜索”中结合迭代查询、中间推理、混合词法 + 稠密搜索和重排序。现有研究通常会报告单项方法带来的准确率提升，且往往允许较宽松地使用工具；但较少有研究量化这些设计参数在固定搜索和令牌预算下，如何共同影响多个模型和数据集上的质量与成本。

在本研究中，我们量化了常见的智能体设计决策（包括混合搜索、轻量级重排序、预先规划、反思和完成限制）如何在搜索次数和令牌受限时改变准确率与成本。我们在六种大语言模型和三个多跳问答基准上进行对照比较，在保持预算不变的同时每次只切换一项决策。

为开展这些实验，我们采用**预算约束智能体搜索（BCAS）**，这是一种考虑预算的评估工具，会将剩余搜索额度和令牌配额作为明确信号呈现，并据此限制工具调用。BCAS 保持了简单的测量方式：复用通用提示词，避免使用定制 API，并记录每个问题的搜索和令牌消耗，以便从业者根据自己的定价模型重新解读结果。

1.1. 研究问题

为量化预算感知规划的实际价值，我们使用六个大语言模型和三个多跳基准测试对 BCAS 开展了一项大规模研究，并围绕以下三个问题展开：

RQ1：不同模型规模下的性能表现
模型规模（参数量）如何影响检索性能？**研究问题2：预算感知的组件调优**

在搜索预算固定的情况下，智能体检索增强生成（*RAG*）的各个组件和超参数如何影响性能？能否利用这些影响，在预算范围内优化性能？

RQ3：准确率-预算权衡
当我们同时收紧迭代搜索步骤和生成输出令牌的预算时，准确率如何变化

智能体式 *RAG* 流程？

近期由强化学习驱动的系统，例如 Search-R1（Jin 等人，2025）和 DeepRe- trieval（Jiang 等人，2025），展现出显著的性能提升，但也带来了高昂的训练和推理开销，因此，智能体检索增强生成中的成本感知行为值得进一步研究。本文其余部分将在第 2 节回顾相关工作，在第 3 节介绍评估协议和预算控制，在第 4 节呈现结果，并在第 5 节和第 6 节通过讨论与总结作结。

2. 相关工作

短短几年间，智能体式检索已从静态流程发展到自主研究智能体，为 BCAS 奠定了背景。

静态流程。早期系统，如 RAG（Lewis 等人，2020）和 FiD（Izacard 和 Grave，2021）确立了先检索后生成的范式，并展示了如何有效结合参数化记忆和非参数化记忆。

提示驱动的智能体。轻量级提示很快便实现了无需新训练的迭代推理。ReAct（Yao 等人，2023）提出了用于搜索规划的“思考、行动、观察”循环；Self-Ask（Press 等人，2023）将问题分解为子查询；FLARE（Jiang 等人，2023）则按需触发检索。Self-RAG（Asai 等人，2024）和 CRAG（Yan 等人，2024）等反思驱动方法通过批判中间结果，进一步提升了鲁棒性。

自主研究智能体。近期研究通过结构化推理和强化学习，进一步深化了外部规划。Self-Reasoning(Xia etal., 2025)训练模型分析自身的轨迹，而 Search-R1 (Jinet al., 2025) 和 DeepRetrieval (Jiang et al.,2025) 则学习直接针对搜索引擎构造并安排查询。

预算感知方面的空白。这些方法通常假设可以大量使用工具，而且许多方法依赖专门的训练或微调。BCAS 提供了一个正交的评估层：它在推理时明确纳入计算约束，统一检索接口，并记录成本遥测数据，以便在共同预算下比较设计选择。我们在这一受控环境中量化了搜索深度、组件选择和令牌限制如何在不同大语言模型之间相互影响。

这一点会影响论文的阅读方式。我们并未推出经过训练的新型搜索智能体，也没有声称在任何基准上达到新的最先进水平。相反，我们采用通用框架，来分离出在搜索次数和生成长度受限时哪些设计选择会产生影响。因此，这项工作的贡献更接近于受控系统测量而非智能体排行榜优化。

3. 方法

本节介绍 BCAS 框架的实现细节，描述 BCAS 组件的消融研究，以及迭代搜索次数和输出上下文大小的扩展研究。我们仅将这一循环用作受控评估框架，以量化在明确预算限制下，搜索深度、检索组件和令牌限制如何影响准确率和成本。图 1展示了 BCAS 的运行示例：DeepSeek V3 (0324) 在严格的搜索预算内解答了一个复杂的 HotpotQA 问题。我们的实现和评估代码见 https://github.com/kmccleary3301/BCAS_检索增强生成。

3.1. 架构原则

BCAS 建立在三项设计原则之上，这些原则与我们的研究问题相契合：

显式预算管理（RQ3）。每一步都会向模型提供剩余的搜索次数和令牌额度，促使它们有条理地安排查询顺序。主要调节项是最大搜索次数和最大令牌总数。_ _ _

可组合的组件流水线（研究问题2）。通过配置启用或禁用检索工具、规划钩子和反思例程，从而能够对各项能力进行受控消融实验。

模型无关设计（RQ1）。提示词以通俗语言描述工具，而非定制 API，因此任何指令遵循型大语言模型都可以采用该框架，无需微调。

3.2. 提示词策略

对于每种评估方法，我们都使用一个适用于所有模型系列的提示模板，不会针对单个模型调整提示词。这样可以固定交互协议，从而直接比较组件效应和预算效应。因此，本研究中的跨模型差异应被解读为模型对共享框架的敏感性，而非各模型经过调优后所能达到的最佳性能。

3.3. 预算核算

BCAS 在每条轨迹中跟踪两种资源：剩余搜索调用次数和累计补全预算。搜索上限作为硬性工具约束执行。一旦额度耗尽，搜索工具就会从可用动作列表中移除，模型必须利用已经收集到的信息作答。每轮结束后，根据提供商报告的用量字段更新令牌计数；当配置的最大_总计_令牌

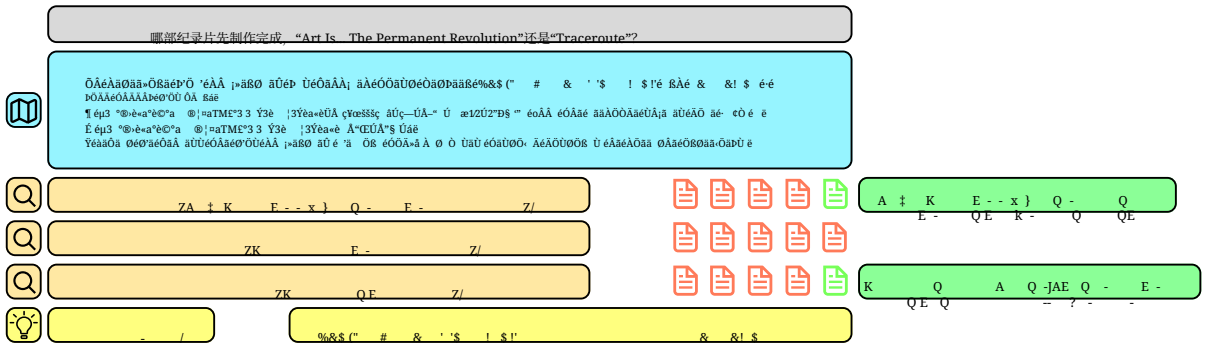

图 1：DeepSeek V3 (0324) 使用 BCAS，通过 4 次搜索和规划解决 HotpotQA“困难”问题的示例。这展示了 BCAS 的核心原则：模型在严格的 4 次搜索预算内运行（RQ1、RQ3），同时利用规划、迭代搜索等系统组件高效定位答案（研究问题2）并提前终止。

阈值达到后，循环停止。这种计量方案刻意保持简单。它不对排队延迟、批处理效应或服务提供商特定定价层级进行建模，但确实能够一致地展示每种配置如何将固定的交互预算转化为答案准确率。

3.4. BCAS 执行循环

BCAS 算法以有状态循环的形式实现，引导大语言模型进行预算感知推理和检索。在每一轮中，模型会：

1：推理：生成关于当前进展以及下一步所需信息的“思考”。**2：选择工具：**从动态生成的可用工具列表中选择一项操作。只有当智能体的搜索次数尚未达到上限_时，才可使用搜索_数据库工具。**3：执行并观察：**执行所选工具（例如，执行数据库搜索），并将结果（“观察结果”）返回给模型。**4：更新预算：**更新令牌使用量，并重新计算下一轮的剩余预算。

该循环会持续进行，直到模型调用 ready_to_answer 工具，或者预算耗尽。

可选的预先规划 启用后，智能体会将问题拆解为循序渐进的检索规划，并考虑可用的搜索和令牌预算，从而以更有条理的方式处理多跳问题。

可选反思 此机制会定期提示模型回顾进展并进行策略调整，对于较小的模型识别无效搜索路径并优化预算使用尤其有效。

这些规划和反思策略的实现遵循了近期智能体搜索领域前沿研究的模式（Xia 等人，2025；Yao 等人，2023；Asai 等人，2024），并进行了调整，使其能够在明确的预算约束下运行。

3.5. 信息检索组件

为分析组件调优（研究问题2），我们为智能体配备了三种不同的信息检索（IR）策略，可针对每项实验进行配置。

- **带短语增强的 BM25：**我们的基线检索器采用标准 BM25 算法，并结合查询解析器对两个词和三个词的短语进行增强。这种方法优先检索对查询概念覆盖密集的文档，其灵感来自覆盖密度（Clarke 等人，2000）。
- **混合搜索（HS）：**该方法将 BM25 检索器与稠密向量搜索相结合，兼顾词汇匹配与语义匹配。
- **重排序（RR）：**我们使用上下文重排序模型重新评分，并选出最相关的前五个结果。此方法可应用于 BM25 或混合搜索之上。

我们的信息检索后端基于 ParadeDB（Noël，2024），这是一款 PostgreSQL扩展。借此，我们可以针对文档构建 BM25 与向量搜索相结合的混合搜索引擎。

检索与重排序流程。 每个搜索请求都会从数据集专用索引返回块级候选项。在BM25模式下，系统返回排名前五的BM25文本块。在混合模式（**HS**）中，词法检索与稠密检索相结合，并将排名前五的文本块返回至智能体上下文。在混合 + 重排序模式（**HS+RR**）中，我们首先检索100个混合候选项，然后对它们重新评分

使用交叉编码器重排序器，并保留排名前五的文本块用于模型上下文。

我们特意将最终上下文窗口限制在返回的五个文本块以内，以确保改进的主要来源是搜索策略，而不是不加区分地扩展上下文。在重排序条件下，包含 100 个候选项的候选池足够大，能够体现出有意义的排序差异，同时运行成本仍然足够低，可以遍历整个消融实验网格。

向量嵌入

我们使用了 `BGE-M3` （Chen 等人，2024）模型，在混合检索期间对文档块和搜索查询进行向量嵌入。`BGE-M3` 因其在多语言基准中的广泛有效性而被选用，相关内容详见（Chen 等人， 2024）。向量相似度采用标准的余弦相似度指标进行测量。在我们的结果中，我们将混合检索器搜索记为**HS**。

重排序

我们测试了在检索阶段对结果进行重排序的影响。我们使用了 `bge-reranker-v2-m3` 来执行此任务，因为其在常见基准测试中表现出较强的竞争力（Li 等人， 2023）。在涉及重排序器的测试案例中，我们通过混合检索检索 100 个文档块，使用重排序模型计算这些文档块的分数，并返回排名前五的文档块。在我们的结果中，我们将重排序记为**RR**。

3.6. 评估

我们在六个大语言模型和三个基准测试上评估 BCAS，以衡量准确率如何随搜索深度和补全词元数量变化。我们还通过消融实验评估各种检索组件对无约束 BCAS 搜索准确率的影响。

大语言模型选择

我们评估了不同规模类别中的六个模型，以分析不同模型能力下的表现（RQ1）：**o4-mini**(OpenAI，2025），**DeepSeek V3（0324 版本）**(DeepSeek，2024），**GPT-4.1-mini** (OpenAI，2024），**Gemma 3 27B** (Team 等人，2024），**通义千问 3 14B**(Cloud，2024），以及**LLaMA3.18B**(Dubey 等人，2024）。这些模型的能力、成本和架构各不相同，并且都能熟练地在 BCAS 任务中使用工具。

评分

我们使用 GPT-4o-mini 作为二元正确性评判器，依据此前证据：大语言模型评判器能够较好地追踪问答场景中的人工答案评分（Zheng 等人，2023）。评判器会接收问题、参考答案和模型答案，并根据一套受约束的评分准则输出正确/错误标签。据我们所知，GPT-4o-mini 与接受评估的六个答题模型不属于同一模型系列。除了文献支持外，我们还对 200 个 TriviaQA 样本、200 个“困难”HotpotQA 样本和 200 个 2WikiMultihopQA 样本进行了独立人工审查。我们确认，TriviaQA 的 200/200 个样本、HotpotQA 的 198/200 个样本，以及 2WikiMultihopQA 的 197/200 个样本均得到了正确的大语言模型评分。五处不一致均为评判器给出的假阴性，因此报告的准确率略显保守。我们仍将二元评判标签视为代理指标，并在第 7 节讨论这一局限。

3.7. 数据集

我们在三个不同的问答基准上进行了规模扩展评估：

1：TriviaQA (Joshi 等人，2017）是广泛使用的开放域问答基准，要求从特定文档中（通常是单个文档）查找具体事实。我们使用未经筛选的测试集，其中包含超过 50 万份文档。**2：HotpotQA** (Yang 等人，2018）侧重于多跳推理，问题要求综合多份文档中的信息来得出答案。

3：2WikiMultihopQA (Ho 等人，2020）是三个基准中难度最高的，要求通过多跳推理处理更复杂的实体关系。

选择这些数据集旨在涵盖从单事实检索到复杂多步推理的各种难度。每个样本都包含一个问题、一个答案和相关文档。

每个数据集的文档都被划分到不同的集合中，测试期间的检索仅限于相应的集合。对于所有情况，我们确保每个模型都在相同的样本集上进行评估，并通过在评估前对数据集进行哈希来验证这一点。由于评估期间偶尔会发生 API 故障，最终样本数在各数据集间为 467 至 537 个，且每个数据集内的样本数保持不变。每次运行都会记录每个问题的元数据，包括搜索次数_、输入令牌_、输出令牌_、提前停止_和评分器判定，并附带成本账本，方便从业者根据不同价格表重新计算美元指标。

3.8. 消融研究

我们对 BCA 进行了消融研究，以衡量可选功能和调整搜索

工具对检索准确率的影响。我们使用 HotpotQA 中的 467 个测试样本（随代码一同发布）对全部六种模型进行了测试。我们的基线配置将 max_total_tokens 设为 16000，将 max_searches 设为“无限制”，搜索端点仅设为 BM25，禁用预先规划和反思，并启用上下文 +搜索提示。我们展示的所有功能调整均以其对基线配置准确率的净影响来衡量。我们衡量以下因素的影响：

1. 预先规划 2. 反思（无预先规划）3. 预先规划和反思 4. 混合 BM25 + 向量搜索（n=5）。5. 混合 BM25 + 向量搜索（n=100），并对结果进行重排序（选取前 5 个）。

如局限性部分所述，我们仅在 HotpotQA 上报告全因子消融，因为多跳综合推理最能体现各组件之间的交互作用。我们并不认为消融增益的确切排序在不同数据集上都普遍适用。

3.9. 搜索扩展研究

为了解迭代搜索能力如何影响性能（RQ3），我们系统地调整了 max_搜索次数参数，设置了四种配置：1、2、3 次搜索，以及不限次数的搜索。所有实验均采用混合搜索，每次查询返回 5 条结果，max_total_令牌 = 16,000，并禁用了预先规划和反思。这个标准化基线确保观察到的性能差异归因于搜索预算变化，而非混杂因素。此外，我们还评估了两种采用不限次数搜索的增强配置：启用预先规划，以及同时启用预先规划和反思。这样，每种模型—数据集组合共有六种配置，使我们能够追踪核心搜索扩展效应及其与策略性推理组件的交互作用。

3.10. 上下文扩展研究

为考察令牌预算约束如何影响智能体行为（RQ3），我们使用混合搜索并将搜索次数设为无限，在五种上下文上限下评估性能：500、1K、2K、4K 和 16K 个令牌，以确保性能变化反映的是上下文预算效应，而非搜索限制。由于 o4-mini 的推理模型架构会将内部推理过程计入生成令牌，因此人为设置过低的令牌上限可能会导致回答在推理过程中途被截断，而非体现出对预算的感知行为，所以我们将其排除在外。适用时，我们也纳入了增强推理配置，这些配置与搜索扩展研究中的相同条件相对应。因此，各上下文上限之间的比较不包括 o4-mini，以避免将预算策略与模型内部的推理令牌化相混淆。

4. 结果

我们发现，不同模型和数据集呈现出三种一致的规律：（i）准确率在搜索次数达到约三次之前稳定提升；（ii）在各个组件中，结合重排序的混合检索带来的收益最为稳定；（iii）增加 token 生成预算，主要有助于处理信息综合要求较高、难度较大的多跳问题。图2 展示了扩展趋势；图3 展示了消融研究结果；表1 报告了扩展指标，表2 详细说明了各组件的消融实验，并回应了 RQ1 至 RQ3。

4.1. 不同模型容量下的性能（RQ1）

迭代搜索缩小了模型容量差距：允许进行更多搜索时，较小模型的得分可接近甚至超过较大模型仅进行一次搜索时的得分。例如，在 HotpotQA 上，通义千问 Qwen 3 14B 在不限搜索次数并进行规划的情况下，得分为 75.33%，超过了 o4-mini 单次搜索的 70.17 %，但未超过 o4-mini 两次搜索时的得分（86.05%；表1）。在 TriviaQA 和 2WikiMultihopQA 上也呈现出类似规律：与单步搜索基线相比，多步搜索提高了较小模型的准确率（图2）。

4.2. 预算感知的组件调优（研究问题2）

在 HotpotQA 上，在混合搜索的基础上进行重排序，在各模型中带来了最大的平均提升（+9.29 个百分点）；仅使用混合搜索也有帮助（+6.36 个百分点），而规划和反思使较小模型提升了 4 到 12 个百分点，对 o4-mini 的影响则有限（表2）。我们将这些效果大小视为 HotpotQA 特有的任务效果，不认为其确切排序会原样迁移到 TriviaQA 或 2WikiMultihopQA。

4.3. 准确率-预算权衡（RQ3）

增加搜索步数后，准确率会持续提升，直至搜索次数约为三次，之后收益趋于平稳；上下文扩展对 TriviaQA 的影响有限，将令牌数从 4k 增至 16k 时，HotpotQA 的准确率明显提升，而对 2WikiMultihopQA 的提升则微乎其微（图2）。各配置的完整结果见表1。

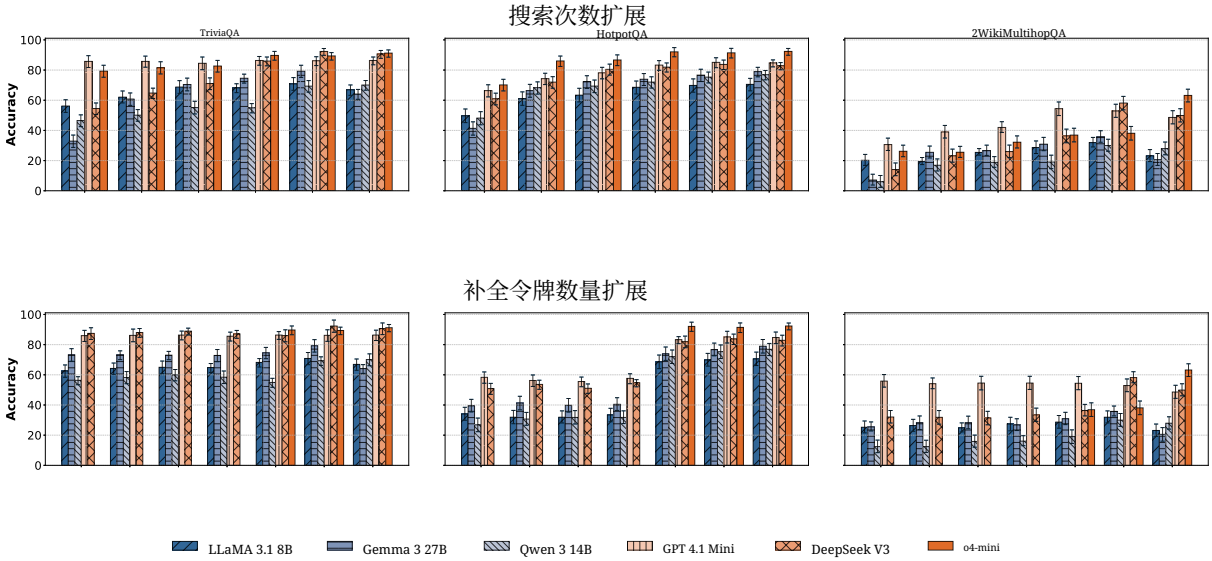

图2：TriviaQA、HotpotQA 和 2WikiMulti-hopQA 上的搜索扩展和上下文扩展表现。“规划”列展示了添加预先规划和反思（RF）策略的影响。此图提供了分析准确率-预算权衡（RQ3）的主要数据。置信区间采用 95% 威尔逊得分区间（Wilson，1927）。

模型	生成令牌				搜索次数（16K 令牌上限）				Plan	规划+RF
	500	1K	2K	4K	1	2	3	无限制		
TriviaQA										
o4-mini	-	-	-	-	79.33	81.56	82.68	89.72	89.39	**91.25**
DeepSeek V3	87.45	88.01	88.95	87.08	54.51	64.78	71.07	85.98	**92.34**	90.65
GPT-4.1-mini	85.96	86.14	86.33	85.58	85.74	85.74	84.49	**86.36**	86.17	86.36
Gemma 3 27B	73.22	73.22	73.03	72.85	32.91	60.80	70.44	74.58	**79.25**	64.11
Qwen 3 14B	56.37	58.24	59.93	58.43	46.54	50.10	55.14	54.95	69.35	**70.09**
LLaMA 3.1 8B	62.55	64.23	64.98	64.79	56.18	62.05	68.76	68.22	**70.84**	66.92
HotpotQA										
o4-mini	-	-	-	-	70.17	86.05	86.70	**92.92**	91.43	92.36
DeepSeek V3	50.86	53.65	51.07	54.72	60.95	72.02	80.48	77.94	**83.74**	82.87
GPT-4.1-mini	58.33	56.25	55.56	57.64	66.38	74.30	78.16	80.47	**85.10**	84.73
Gemma 3 27B	39.40	41.33	39.83	40.47	41.38	66.38	72.20	68.52	76.66	**78.96**
Qwen 3 14B	26.82	30.69	31.76	31.76	48.18	68.31	69.38	64.45	75.33	**76.91**
LLaMA 3.1 8B	34.05	31.91	31.91	33.40	49.68	61.03	63.38	65.31	69.83	**70.52**
2WikiMultihopQA										
o4-mini	-	-	-	-	26.18	25.54	32.19	36.91	38.12	**63.20**
DeepSeek V3	27.68	26.18	26.39	29.18	14.07	23.16	25.97	36.36	**58.15**	49.89
GPT-4.1-mini	**55.84**	54.11	54.55	54.55	30.62	38.97	41.97	54.39	52.89	48.61
Gemma 3 27B	25.48	28.27	28.27	26.77	7.13	25.49	26.57	30.89	**35.76**	20.56
Qwen 3 14B	12.42	12.42	11.56	13.28	6.00	16.92	18.84	19.06	**29.98**	28.66
LLaMA 3.1 8B	25.27	26.34	24.84	27.41	20.22	19.35	25.43	28.70	**31.91**	23.13

表1：所有模型在所有数据集及 BCAS预算配置下的完整准确率数据（%）。每行中的最高值以粗体标示。请注意，TriviaQA 的表现整体上几乎没有变化。

模型	Plan	RF	规划+RF	HS	HS+RR
LLaMA 3.1 8B	+4.52	+6.33	+5.21	+5.73	+8.80
Gemma 3 27B	+8.14	+7.37	+10.43	+8.93	+13.01
Qwen 3 14B	+10.87	+5.19	+12.45	+12.36	+17.86
GPT-4.1-mini	+4.63	+3.89	+4.26	+4.07	+5.93
DeepSeek V3	+5.79	+4.55	+4.92	+6.23	+9.02
o4-mini	+0.17	+0.92	+1.10	+0.83	+1.10
平均值	+5.69	+4.71	+6.40	+6.36	+9.29

表2：检索方法组件对 HotpotQA 准确率（%）的影响。本分析为预算约束下的组件调优提供依据（研究问题2）。

5. 讨论

结果表明，在我们的网格中，实用的预算策略是优先增加搜索次数（通常最多增加到三次），通过混合检索与重排序来提升证据质量，并主要针对多跳推理任务提高令牌生成上限。这样的排序几乎涵盖了观察到的所有收益，同时限制了计算资源。

这一框架带来的一个结果是，本文对配置决策的支持比对“存在普遍最优智能体”这一论断的支持更为直接。所有结果都以相同的提示框架、检索器栈和预算规则为前提。这使得这些比较有助于决定如何使用固定预算，但也意味着应将本文视为一项关于受控条件下设计杠杆作用的研究，而非异构系统排行榜。

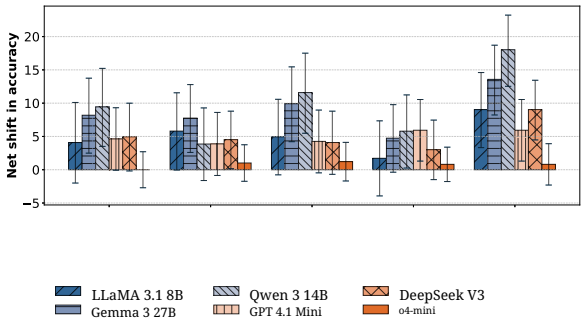

图3：HotpotQA 上 BCAS 特征的消融研究：我们使用 BM25 检索器，衡量不同特征相较于 BCAS 基线对检索准确率的净影响。95% 置信区间采用 Newcombe 方法 9 计算（Newcombe，1998年）。

5.1. 因数据集而异的预算杠杆效应

扩展生成令牌预算所带来的最不寻常影响出现在 HotpotQA 上。TriviaQA 和 2WikiMultihopQA 的准确率基本持平，而 HotpotQA 在预算从 4k 增至 16k 令牌时，准确率则大幅提升。我们推测，不同数据集所考查的困难类型有所不同：2WikiMultihopQA 似乎主要受检索限制，因为找到正确证据并不容易；而 HotpotQA 对检索的依赖似乎较低，表现更受限于将多条证据综合成最终回答所需的推理能力。这可以解释为何增加迭代搜索次数时，2WikiMultihopQA 的准确率提升比 HotpotQA 更明显；而 HotpotQA 则在令牌数量增加时取得显著收益，这与 2WikiMultihopQA 不同。

这些模式还揭示了搜索深度与上下文分配之间的不对称关系。所有模型和数据集的准确率都会随着搜索步数增加而单调上升，但搜索次数超过三次后，边际收益递减。然而，上下文扩展表明，当模型能够执行多次搜索时，受限的生成令牌数可能优于不受限的上下文。这一现象在 2WikiMultihopQA 上最为明显：GPT-4.1-mini 使用 500 个生成令牌时准确率达到 55.84%，而使用 16K 个令牌进行单次搜索时准确率仅为 30.62%。同样，DeepSeek V3 使用 500 个令牌时准确率达到 27.68%，而单次搜索时仅为 14.07%。在令牌约束严格的情况下，模型会采取简洁的行为方式，高效调用工具，从而在预算内执行多次搜索。相反，不受限制的单次搜索会促使模型生成冗长的回答，耗尽全部令牌预算，却无法额外收集信息。

在我们的实验中，将预算用于增加搜索深度，往往比将同样的预算用于延长单次生成更有效。对于从业者而言，适度受限的令牌预算（500 至 2K）结合多次搜索机会，可能比宽裕的单次搜索配置带来更高的准确率。

5.2. o4-mini 异常

在我们对不同模型规模的评估中，o4-mini 呈现出最令人费解的行为模式。尽管它总体表现名列前茅，却对大多数系统增强措施几乎没有反应。在 HotpotQA 上，所有增强策略带来的提升都不到 1.10 个百分点，与其他模型提升 4 至 18 个百分点的表现形成鲜明对比。

这种对增强措施的抵制，或许源于 o4-mini 是一种内置思维链处理能力的推理模型。因此，外部规划组件往往是多余的。一个例外出现在 2WikiMultihopQA 上：将规划与反思结合后，性能提升了 25 个百分点（从 38.12% 升至 63.20%）。这表明，虽然基础的预先规划可能是多余的，但反思能够为搜索过程中调整策略创造机会，从而补充 o4-mini 的内部推理过程。

5.3. 组件层级与模型多样性

消融结果表明，各组件的有效性因模型而异，差异十分显著。结合混合搜索的重排序带来了最稳定的收益（对大多数模型为6至18个百分点），使其成为跨模型类别最可靠的增强手段。然而，改进幅度因模型而显著不同，较小的模型往往比更大的模型更能从检索增强中获益。

预先规划策略呈现出类似模式：较小的模型（LLaMA 3.1 8B、Qwen 3 14B）从规划组件中获得4至12个百分点的提升，而较大的模型改进更为有限。这表明，对于固有能力更强的模型，外部脚手架的价值会降低。将预先规划与反思相结合会略微降低部分模型的性能，说明自适应重新规划可能适得其反；对常规模型而言，遵循初始规划似乎比在搜索过程中持续修订策略更为有效。

5.4. 部署建议

基于我们的评估趋势，我们建议在预算分配中优先考虑迭代搜索深度，其次是组件选择，最后是令牌生成限制。对于成本敏感型部署，

LLaMA 3.1 8B（所评估的最小模型）进行三次搜索并使用 BM25 后，性能已经超过更大型系统的单次搜索基线，但 o4-mini 和 GPT-4.1-mini 除外。加入预先规划和反思后，它能够达到或超过所有仅限一次搜索的模型，同时成本仍大幅降低。

为实现最高准确率，o4-mini 在两种规划策略下于 2WikiMultihopQA 上均达到 63.20%，这是观察到的最高性能，但这需要针对复杂任务进行专门配置。缺乏原生思维链能力的模型，从混合检索和重排序中获得的收益大于从复杂推理框架中获得的收益，因此增强措施应与模型自身的能力特征相匹配。

实际部署时，首先应该考虑的问题通常是“我们能负担多少次搜索？”，而不是“应该购买多大的补全窗口？”。在我们的网格实验中，最明显的提升来自增加一次有针对性的检索步骤，或改进对检索到的证据的排序。只有在检索到的材料已经包含所需信息、而模型仍需要空间来综合这些信息时，较长的生成结果才会带来主要帮助。

6. 结论

我们提出了 BCAS，这是一套预算感知评估工具，用于在明确的搜索和令牌限制下研究智能体式检索。我们使用六个大语言模型，在 TriviaQA、HotpotQA 和 2WikiMultihopQA 上量化了搜索深度、检索配置和补全预算之间的权衡，以及它们对准确率和成本的影响。

在我们的实验中，有三个一致的规律。第一，在大多数场景中，增加搜索深度可提升准确率，但达到较低上限后，收益会逐渐递减；搜索次数约为三次后，提升尤为有限。第二，在 HotpotQA 消融实验网格中，混合词法+稠密检索与轻量级重排序带来的平均准确率提升最大。第三，在综合需求较高的 HotpotQA 场景中，较大的补全预算最为重要；而在检索难度较低的场景中，性能更快趋于饱和。

这些结果为预算受限的部署提供了切实可行的配置顺序：首先增加搜索深度，然后通过检索和重排序提升证据质量，最后在综合需求足以证明额外成本合理时提高补全预算。通过明确工具和令牌预算，并记录使用情况及结果，BCAS 支持在现实约束下开展可复现的准确率—成本分析。

6.1. 未来方向

未来工作应增加直接的非智能体基线，并将组件消融实验扩展到 Hot- 之外potQA，并将评估扩展到开放网络和多模态环境。第二个方向是更精细的运行时报告，包括生产环境并发限制下的排队和吞吐量行为。

7. 局限性

我们的研究聚焦于静态基准语料库上的事实问答基准，而非开放网络检索。所报告的趋势未必能直接适用于知识快速漂移的领域、多语言语料库、多模态证据或长时程交互任务。

这些基准测试也侧重于简短且有已知参考答案的最终答案。这有助于进行受控比较，但未涵盖那些成功取决于引文忠实度、长篇综合、交互式澄清或用户特定约束的场景。因此，适用于基准测试问答的策略，在生产工作流中可能需要采用不同的停止规则或检索广度。

完整的组件消融仅在 HotpotQA 上报告。我们并不声称组件收益的确切排序在不同数据集上普遍适用；TriviaQA 和 2WikiMultihopQA 可能会改变检索质量、搜索深度和补全预算的相对价值。

对于每种方法，我们在不同模型系列中使用同一个提示模板，而不针对每个模型进行提示调优。此举有助于控制比较测量，但可能无法充分优化某些模型，因此，跨模型比较应被理解为模型对共享框架的敏感性，而非各模型的最佳表现。

我们通过大规模二元大语言模型裁判来衡量答案正确性。尽管人工审查表明两者高度一致，二元评分仍可能无法捕捉部分正确性以及推理质量方面的细微差异。

成本和延迟估算取决于服务提供商定价、速率限制和运行时设置。因此，我们将令牌使用量和搜索使用量作为稳定的预算代理指标，与成本估算一并报告，并将跨服务提供商的成本比较视为近似值。

本版本不包含纯粹的非智能体单轮先检索后阅读基线。因此，所呈现的比较仅用于分析受预算约束的智能体策略之间的差异，无法直接估算多轮智能体能力相较于单轮检索所带来的增量价值。

我们也未报告与经强化学习训练的开放网络搜索智能体的正面对比结果。这些系统的训练机制和检索环境各不相同，因此直接比较需要统一工具、语料库和模型训练条件。

8. 数据与代码可用性

本研究使用的代码、评估工具配置、提示词和分析脚本均可在 BCAS 仓库中获取：https://github.com/kmccleary3301/BCAS_检索增强生成。我们还提供了数据集划分哈希值和评估输出，以便在相同预算设置下复现文中报告的图表。

9. 参考答案

Anthropic。2024。介绍 Claude 的研究模式。https://www.anthropic.com/research。访问日期：2025-07-01。

Akari Asai、Zeqiu Wu、Yizhong Wang、Avirup Sil 和 Hannaneh Hajishirzi。2024。Self-RAG：自反思检索增强生成。发表于国际学习表征会议（*ICLR*）。

Jianlv Chen、Shitao Xiao、Peitian Zhang、Kun Luo、Defu Lian 和 Zheng Liu。2024。BGE M3-Embedding：通过自知识蒸馏实现多语言、多功能、多粒度文本嵌入。

Charles L. A. Clarke、Gordon V. Cormack 和 Elizabeth A. Tudhope。2000。一至三个词项查询的相关性排序。信息处理与管理，36(2)：291–311。

阿里云。2024。通义千问技术报告。访问日期：2025-01-15。

DeepSeek。2024。DeepSeek-V3 技术报告。访问日期：2025-01-15。

Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Amy Yang、Angela Fan 等。2024。Llama 3 模型家族。*arXiv* 预印本 *arXiv:2407.21783*。

Gemini Team。2025。Gemini 2.5：以先进推理、多模态能力、长上下文和新一代智能体能力突破前沿。技术报告，GoogleDeepMind。访问日期：2025-07-01。

Xanh Ho、Anh-Khoa Duong Nguyen、Saku Sug- awara 和 Akiko Aizawa。2020。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，西班牙巴塞罗那（线上）。国际计算语言学委员会。

Gautier Izacard 和 Edouard Grave。2021。利用段落检索与生成模型进行开放域问答。载于欧洲计算语言学协会第*16*届会议论文集：主会场，第874–880页。P. Jiang、J. Lin、L. Cao、R. Tian、S. Kang、Z. Wang、J. Sun 和 J. Han。2025。Deepretrieval：通过强化学习，利用大型语言模型攻破真实搜索引擎和检索器。*arXiv* 预印本 *arXiv:2503.00223*。

Zhengbao Jiang、Frank F Xu、Luyu Gao、Zhiqing Sun、Qian Liu、Jane Dwivedi-Yu、Yiming Yang、Jamie Callan 和 Graham Neubig。2023。主动检索增强生成。载于*2023*年自然语言处理实证方法会议（*EMNLP*）论文集。

B. Jin、H. Zeng、Z. Yue、J. Yoon、S. Arik、D. Wang、H. Zamani 和 J. Han。2025。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*。

Mandar Joshi、Eunsol Choi、Daniel S. Weld 和 Luke Zettlemoyer。2017。TriviaQA：一个用于阅读理解的大规模远程监督挑战数据集。载于第 *55* 届计算语言学协会年会论文集，加拿大温哥华。计算语言学协会。

Patrick Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich Küttler、Mike Lewis、Wen-tau Yih、Tim Rocktäschel、Sebastian Riedel 和 Douwe Kiela。2020。面向知识密集型自然语言处理任务的检索增强生成。载于神经信息处理系统进展，第 33 卷，第 9459–9474 页。

Chaofan Li、Zheng Liu、Shitao Xiao 和 Yingxia Shao。2023。让大语言模型成为更好的稠密检索基础。

Robert G. Newcombe。1998。基于配对数据改进二项比例之差的置信区间。医学统计学，17(22):2635–2650。

Philippe Noël。2024。ParadeDB。https://github.com/paradedb/paradedb。访问日期：2024-7-01。

OpenAI。2023。函数调用及其他 API 更新。https://platform.openai.com/docs/guides/function-calling。访问日期：2025-07-01。

OpenAI。2024。在 API 中推出 GPT-4.1。访问日期：2025-07-04。

OpenAI。2025。OpenAI o3 系统卡。访问日期：2025-06-23。

Ofir Press、Muru Zhang、Sewon Min、Ludwig Schmidt、Noah A Smith 和 Mike Lewis。2023。衡量并缩小语言模型中的组合性差距。发表于计算语言学协会年会 *EMNLP 2023* 论文集。

A. Singh 等人。2024。智能体式检索增强生成：智能体检索增强生成综述。*arXiv* 预印本 *arXiv:2501.09136*。

Gemma 团队、Morgane Riviere、Shreya Pathak、Pier Giuseppe Sessa、Cassidy Hardin、Surya Bhupatiraju、Léonard Hussenot、Thomas Mesnard、Bobak Shahriari、Alexandre Ramé 等人。2024。Gemma 2：以实用规模改进开放语言模型。

Edwin B Wilson。1927。概率推断、继承律与统计推断。《美国统计协会杂志》，22(158):209–212。

Yuan Xia、Jingbo Zhou、Zhenhui Shi、Jun Chen 和 Haifeng Huang。2025。通过自我推理改进检索增强语言模型。*AAAI* 人工智能会议论文集，39(24):25534–25542。

Shi-Qi Yan、Jia-Chen Gu、Yun Zhu 和 Zhen-Hua Ling。2024。CRAG：纠错式检索增强生成。*arXiv* 预印本 *arXiv:2401.15884*。

Zhilin Yang、Peng Qi、Saizheng Zhang、Yoshua Bengio、William W. Cohen、Ruslan Salakhutdi- nov 和 Christopher D. Manning。2018。Hot- potQA：一个用于多样化、可解释的多跳问题解答的数据集。载于自然语言处理经验方法会议（*EMNLP*）。

Shunyu Yao、Jeffrey Zhao、Dian Yu、Nan Du、Izhak Shafran、Karthik Narasimhan 和 Yuan Cao。2023。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*）。

Lianmin Zheng、Wei-Lin Chiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、Eric P. Xing、Hao Zhang、Joseph E. Gonzalez 和 Ion Stoica。2023。使用 MT-Bench 和 Chatbot Arena 评估大语言模型评审。