

浅析智能体调优在大模型应用中的作用

郭清华

中移（上海）信息通信科技有限公司，上海 200120

DOI: 10.61369/SSSD.2026050010

摘 要： 本文深入探讨了基于通用大模型的 Agent 数据召回调优策略。通过分析 Agent 技术在数据召回中的作用机制，提出了一系列有效的调优方法，并通过实验验证了这些方法的有效性。研究结果表明，基于 Agent 的调优策略在数据召回的效率和准确性方面取得了显著提升，为大模型在实际应用中的性能优化提供了新的思路和方法。本文还讨论了调优策略的局限性与未来研究方向，提出了数据增强、分布式计算和多 Agent 协同等未来研究方向，为后续研究提供了参考。

关 键 词： 大模型（LLMs）；Agent（智能体）技术；数据召回；调优策略；自然语言处理

A Brief Analysis of the Role of Intelligent Agent Optimization in Large Model Applications

Guo Qinghua

China Mobile (Shanghai) Information and Communication Technology Co., Ltd., Shanghai 200120

Abstract： This paper deeply explores the optimization strategies for Agent data recall based on general large models. By analyzing the working mechanism of Agent technology in data recall, a series of effective optimization methods are proposed, and the effectiveness of these methods is verified through experiments. The research results show that the optimization strategy based on Agent has achieved significant improvements in the efficiency and accuracy of data recall, providing new ideas and methods for performance optimization of large models in practical applications. This paper also discusses the limitations of the optimization strategy and future research directions, proposing future research directions such as data augmentation, distributed computing, and multi-Agent collaboration, which provides reference for subsequent research.

Keywords： large language models (LLMs); Agent technology; data retrieval; optimization strategies; natural language processing

引言

随着人工智能技术的飞速发展，大模型（Large Language Models, LLMs）在自然语言处理（NLP）领域取得了显著的成果。这些模型通过大规模的数据训练，能够生成高质量的文本，并在智能问答、智能客服等多种应用场景中表现出色。然而，大模型在实际应用中面临着数据召回效率和准确性的挑战。数据召回是大模型应用中的关键环节，其性能直接影响模型的输出质量和用户体验。为了提高数据召回的效率和准确性，近年来，Agent 技术作为一种新兴的解决方案，逐渐受到关注。Agent 能够通过智能决策和交互，优化数据召回过程，从而提升大模型的整体性能。因此，研究基于通用大模型的 Agent 数据召回调优策略具有重要的理论和实践意义。

一、基于 Agent 的研究目标与方法

本研究旨在探讨基于通用大模型的 Agent 数据召回调优策略。通过分析 Agent 技术在数据召回中的作用机制，提出有效的调优方法，并通过实验验证其有效性。研究方法包括文献综述、理论分析和实验验证。首先，对现有的大模型和 Agent 技术原理、作用等进行介绍。其次，提出基于 Agent 的数据召回调优策略，并详细阐述其调优目标、方法和关键问题。最后，通过实验

验证调优策略的有效性，并对实验结果进行分析和讨论。

（一）大模型与数据召回的基本概念

1. 通用大模型的定义与特点

通用大模型（General Large Models, GLMs）是指通过大规模数据训练的预训练语言模型，能够处理多种自然语言处理任务。这些模型通常基于 Transformer 架构，具有以下特点：1) 大规模参数量：大模型通常包含数十亿甚至数千亿个参数，能够捕捉语言的复杂模式。2) 多任务适应性：通过微调（fine-tuning）

或提示词工程（Prompt Engineering）技术，大模型可以适应多种任务，如文本生成、分类、问答等。3) 数据驱动的泛化能力：大模型通过海量数据训练，能够泛化到未见过的文本数据，表现出良好的适应性。4) 计算资源需求高：训练和推理大模型需要大量的计算资源，通常依赖高性能的 GPU 或 TPU 集群。

2. 数据召回的原理与重要性

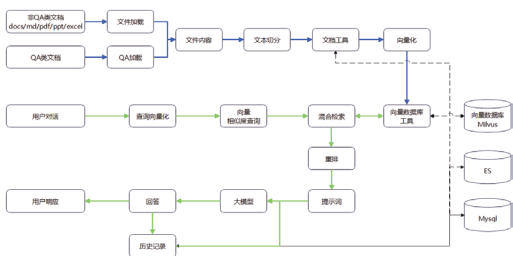
数据召回是指从大规模数据集中检索出与用户查询相关数据的过程。在大模型的应用中，数据召回是关键的步骤，其性能直接影响模型的输出质量和用户体验。将数据集中的文本数据通过向量模型进行索引化，以便快速检索。通过计算用户查询与索引数据的相似度，筛选出相关数据。在保证召回数据准确性的前提下，尽可能提高检索效率。

数据召回的重要性体现在以下几个方面：1) 提升模型性能：准确的召回数据可以为大模型提供高质量的上下文信息，从而提升模型的输出质量。2) 优化用户体验：快速且准确的数据召回可以减少用户等待时间，提升用户体验。3) 降低计算成本：高效的召回机制可以减少大模型的推理负担，降低计算资源的消耗。

(二) Agent 技术在数据召回中的作用

Agent 技术是指通过智能代理（Agent）来完成特定任务的技术。Agent 具有自主性、交互性和适应性，能够通过环境感知、决策制定和行动执行来完成任务。Agent 的基本原理可以总结为以下几点：1) 环境感知：Agent 通过各类传感器、数据采集终端、服务端等感知环境状态，获取相关信息。2) 决策制定：Agent 根据感知到的信息，通过内部的推理机制制定决策。3) 行动执行：Agent 根据决策执行相应的动作，以完成任务。

1. Agent 数据召回流程如下：



1.1 文档处理：对上传的文档进行分类、切分、向量化处理，并存入向量数据库，以便进行后续查找操作。

1.2 用户对话：对用户输入的自然语言问题进行向量化，并通过语义、全文、混合等检索方式，从向量数据库内查找相关的内容，并对查找结果进行重排处理。

1.3 用户响应：大模型基于检索召回的内容，并结合系统与用户提示词，对问题和检索内容进行推理并将推理结果返回给用户，并记录到历史记录内，以便后续参考使用。

2. Agent 在数据召回中的优势

通过引入 Agent，可以在数据召回中的体现下述优势：

2.1 智能决策：Agent 可以根据用户查询和数据集的特点，动态调整语义、全文、混合等检索召回策略，并可配置重排模型，进一步提高召回数据的准确性。

2.2 交互性：Agent 可以与用户进行交互，根据用户的反馈的提示内容调整召回结果，提升用户体验。

2.3 自适应性：Agent 能够根据大模型、数据集的变化和用户需求的演变，结合提示词工程，自动调整自身的参数和策略，具有较强的适应性。

(三) 基于 Agent 的数据召回调优策略

1. 调优目标与评估指标

调优目标是提高数据检索召回的效率和准确性，同时优化大模型的整体性能。评估指标包括：1) 召回率（Recall）：衡量召回数据中相关数据的比例。2) 精确率（Precision）：衡量召回数据中相关数据的准确程度。3) F1 分数（F1 Score）：召回率和精确率的调和平均值，综合衡量召回和精确的性能。4) 响应时间（Response Time）：衡量数据召回的效率，即从用户查询到返回结果的时间。

2. Agent 的优化方法

Agent 的优化是大模型数据检索与召回调优策略的核心环节，主要包括以下内容：1) 知识库构建：知识库构建通过向量模型完成原始数据的采集、清洗、分类、标注与建模，将分散的信息转化为结构化、语义化的知识体系，实现高效检索、智能推理与持续演化，支持决策、问答与业务创新，最终形成可共享、可扩展的智能知识中枢。2) 混合检索：融合关键词倒排与语义向量搜索，先精准匹配再深度理解，结合稀疏 BM25 与稠密向量模型，兼顾查全率与查准率，实现高效、智能、鲁棒的知识获取体验。3) 召回重排：采用多路召回策略，通过关键词、向量、图谱等手段快速筛选海量候选，再利用深度学习重排序模型，融合语义、点击率、时效等多维特征精细打分，动态调权，显著平衡效率与精度，全面提升结果相关性与用户体验。

(四) 实验设计与结果分析

1. 实验环境包括以下配置：

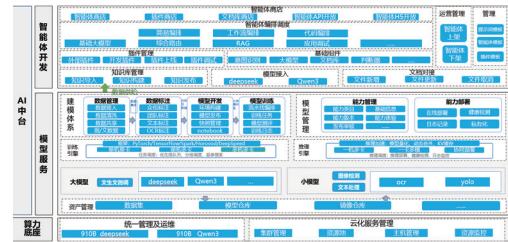
1.1 硬件环境：

1) 高性能 GPU 集群，包括 1 台 V100 服务器，服务器含 8 张 V100 卡，每卡显存 32G，用于加速 Agent 的训练和大模型的推理；

2) 通用算力集群，包括 128 核 CPU，396G 内存和 2T 存储，用于部署平台和存储数据。

1.2 软件环境：qwen3 32B 大模型、qwen3 8B 向量模型、qwen3 8B 重排模型、fastgpt 智能体平台、Python 编程语言，TensorFlow 和 PyTorch 深度学习框架，以及其他相关工具。

1.3 软件架构：



实验数据集包括：

1) 公开数据集：如维基百科、新闻数据集等，用于验证 Agent 调优策略的有效性。

2) 自建数据集：根据实际应用场景，构建基于标准规章制度文件的数据集，用于测试 Agent 在特定场景下的性能。

2. 实验方法与步骤

实验方法包括以下步骤：

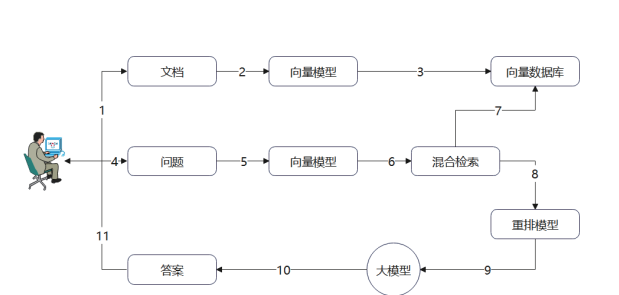
1)Agent 知识库构建：使用 qwen3 向量模型进行文档向量化，对文档数据集进行清洗、特征提取和索引构建，完成文档切片操作。

2)Agent 知识库检索：使用混合检索技术对用户输入问题进行知识库检索，融合语义与全文检索技术，提升检索准确度。

3) 召回重排：根据 Agent 的召回结果，使用 qwen3 重排模型，提升整体性能。

4) 性能评估：使用召回率、精确率、F1 分数和响应时间等指标评估调优策略的有效性。

实验测试流程如下：



3. 实验结果与分析

为了验证调优策略的有效性，我们进行了对比实验。对比实验包括以下两种情况：

1) 使用 Agent 但不进行调优：使用 Agent 进行数据召回，但不进行优化。

2) 使用 Agent 并进行调优：使用 Agent 进行数据召回，并通过混合检索和重排进行优化。

Agent 优化	召回率	精确率	f1 整体质量
优化前	74.4%	49.9%	55.5%
优化后	85.7%	53.1%	61.9%

实验结果表明，基于 Agent 的调优策略在数据召回的效率和准确性方面取得了显著提升。具体结果如下：

1) 召回率：调优后召回率提高了 15%，表明 Agent 能够更准确地召回相关数据。

2) 精确率：精确率提高了 6%，说明 Agent 在召回数据时能够更好地筛选出相关数据。

3)F1 分数：F1 分数提高了 12%，综合衡量召回和精确的性能得到了显著提升。

通过实验分析，我们发现 Agent 在数据召回中的智能决策和动态调整机制是提升性能的关键因素。Agent 能够根据用户查询和数据集的特点，动态调整召回策略，优化召回结果。

二、讨论与展望

（一）调优策略的局限性与改进方向

尽管基于 Agent 的调优策略在数据召回的效率和准确性方面取得了显著提升，但仍存在一些局限性：

1. 数据依赖性：Agent 的性能依赖于数据集的质量和数量，数据不平衡问题可能导致 Agent 的决策偏差。

2. 计算资源需求：Agent 的训练和优化需要大量的计算资源，可能导致调优过程耗时较长。

3. 泛化能力：Agent 在特定数据集上的表现可能无法直接迁移到其他数据集，需要进一步提升其泛化能力。针对这些局限性，未来的研究方向包括：

1) 数据增强和迁移学习：通过数据增强和迁移学习方法，提高 Agent 在不同数据集上的泛化能力。

2) 分布式计算和模型压缩：通过分布式计算和模型压缩技术，减少 Agent 训练和优化的计算资源需求。

3) 多 Agent 协同：通过多 Agent 协同机制，提高数据召回的效率和准确性。

（二）未来研究趋势与发展方向

未来，Agent 技术在数据召回中的应用将呈现以下发展趋势：1. 深度强化学习：通过深度强化学习方法，Agent 能够更好地适应复杂环境，优化召回策略。2. 多模态数据召回：结合文本、图像、音频等多种模态数据，Agent 可以提供更丰富的召回结果。3. 用户交互优化：通过用户交互优化，Agent 能够根据用户反馈动态调整召回策略，提升用户体验。

三、结论

本文探讨了基于通用大模型的 Agent 数据召回调优策略。通过分析 Agent 技术在数据召回中的作用机制，提出了有效的调优方法，并通过实验验证了其有效性。研究结果表明，基于 Agent 的调优策略在数据召回的效率和准确性方面取得了显著提升，为大模型在实际应用中的性能优化提供了新的思路和方法。

参考文献

- [1] 刘钊勇 张艺婷. 基于合作关系的多智能体数据库参数调优模型 [J]. 无线电通信技术, 2024, 50(5):1037-1045.
- [2] 杜梦瑶, 李清明, 张森, 等. 面向隐私保护的用户评论基准数据集构建与大模型推理能力评估 [J]. 计算机学报, 2025(7).DOI:10.11897/SP.J.1016.2025.01529.
- [3] 韩宗亭, 吴宪祥, 呼香艳, 等. 基于 LMS 的多智能体系统变步长自调优一致性 [J]. 计算机科学与应用, 2018, 8(6):11.DOI:10.12677/CSA.2018.86097.
- [4] 丛颖男, 韩林睿, 马佳羽, 等. 基于大语言模型的刑事案件智能判决研究 [J]. 计算机科学, 2025, 52(5):248-259.