

基于高效 LoRA 自适应的语法感知图检索增强生成

浦震宇¹, 车延楠^{2†}, 黄 勃^{1,3,4†}, 倪福川⁵

1. 上海工程技术大学, 电子电气工程学院, 上海 201620;
2. 上海科学技术交流中心, 上海 200126;
3. 安徽志国智能科技有限公司, 安徽 滁州 239000;
4. 华中农业大学, 农村智能技术教育部工程中心, 湖北 武汉 430070;
5. 华中农业大学, 信息学院, 农业农村部智慧养殖技术重点实验室, 湖北 武汉 430070

收稿日期: 2025-10-16 † 通信联系人 E-mail: cheyn@sstec.org.cn(车延楠), huangbosues@sues.edu.cn(黄勃)

基金项目: 上海市协同创新项目(25xtcx00100); 上海市中央引导地方科技发展资金课题(YDZX20253100004004003); 华中农业大学农业智能技术教育部工程研究中心开放课题(ERCITA-KF002)

第一作者: 浦震宇, 男, 硕士生, 现从事人工智能方面的研究。E-mail: 1741580514@qq.com

摘 要: 随着用户查询需求的日益复杂化, 现有检索增强生成(RAG)模型在处理查询时, 往往将问题视为扁平化的词元序列, 忽略了其中关键的语法结构信息, 从而削弱了处理复杂问题的能力。针对这一问题, 本文提出了一种融合语法感知能力的 GraphRAG 框架。该框架的核心创新在于: 引入基于依存关系的语法结构与低秩自适应微调策略(LoRA), 以增强模型在知识密集型问答任务中的表现。具体方法: 在框架的检索与生成阶段引入源于依存句法树的语言学先验特征; 同时, 采用 LoRA 技术对 LLaMA-3.1-8B-Instruct 模型进行高效参数适配与微调。实验在该领域两个常用数据集 QALD-h 和 DBpedia QA 上进行, 结果表明, 所构建框架在各项关键指标上均优于基线模型, 消融实验进一步证实了引入语法结构与参数高效微调所带来的性能提升。

关 键 词: 检索增强生成; 图结构检索; 语法感知; 低秩自适应; 参数高效微调

中图分类号: TP391

文献标志码: A

文章编号: 1671-8836(2026)04-0497-10

Syntax-Aware GraphRAG with Efficient LoRA-Based Adaptation

PU Zhenyu¹, CHE Yannan^{2†}, HUANG Bo^{1,3,4†}, NI Fuchuan⁵

1. School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China;
2. Shanghai Science and Technology Exchange Center, Shanghai 200126, China;
3. Anhui Zhiguo Intelligent Technology Co., Ltd., Chuzhou 239000, Anhui, China;
4. Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Huazhong Agricultural University, Wuhan 430070, Hubei, China;
5. Key Laboratory of Smart Breeding for Ministry of Agriculture and Rural Affairs, College of Informatics, Huazhong Agricultural University, Wuhan 430070, Hubei, China

Abstract: As user queries grow increasingly complex, existing Retrieval-Augmented Generation (RAG) models tend to treat questions as flattened token sequences during query processing, overlooking critical syntactic structure information and consequently degrading their ability to handle complex queries. To address this issue, this paper proposes a syntax-aware GraphRAG framework. The core innovation lies in incorporating dependency-based syntactic structures into the GraphRAG framework while employing Low-Rank Adaptation (LoRA) for efficient parameter adaptation. Specifically, prior linguistic features derived from dependency syntax trees are injected into both the retrieval and generation phases of the framework; meanwhile, LoRA is employed to achieve

引用格式: 浦震宇, 车延楠, 黄勃, 等. 基于高效 LoRA 自适应的语法感知图检索增强生成[J]. 武汉大学学报(理学版), 2026, 72(4): 497-506. DOI: 10.14188/j.1671-8836.2026.0004.

PU Zhenyu, CHE Yannan, HUANG Bo, et al. Syntax-Aware GraphRAG with Efficient LoRA-Based Adaptation [J]. J Wuhan Univ (Nat Sci Ed), 2026, 72(4): 497-506. DOI: 10.14188/j.1671-8836.2026.0004(Ch).

efficient parameter adaptation and fine-tuning of the LLaMA-3.1-8B-Instruct model. Experiments are conducted on two widely-used datasets, QALD-h and DBpedia QA, and the results demonstrate that our proposed framework outperforms baseline models across all key metrics. Ablation studies further confirm the performance gains brought by the incorporation of syntactic structures and parameter-efficient fine-tuning.

Key words: retrieval-augmented generation; graph structure retrieval; syntax-aware; low-rank adaptation; parameter-efficient fine-tuning

0 引言

检索增强生成(Retrieval-Augmented Generation, RAG)通过分离检索与生成模块,使系统能调用外部知识库生成基于实证的语境化响应^[1-2],为处理知识密集型自然语言任务提供了一种有效的方式^[3-4]。随着技术成熟,基于检索或图谱的RAG架构不仅在开放域问答中被广泛应用,更深入到大量垂直领域的复杂任务中^[5-6]。在医疗领域中RAG技术常结合知识图谱被用于临床辅助决策与医学问答^[7-8];同时也被应用于医学特征提取和重症监护图谱的构建中^[9-10]。在工业界,大模型架构加速了智能制造和新材料研发的进程^[11-12]。此外,它在建筑工程管理和高性能计算等场景中同样展现了其潜力^[13-14]。除了特定的法律辅助解析与智能座舱问答系统^[15-16],研究者在扩展应用的同时,也越来越关注系统的人机交互体验评估与抵御隐私推断攻击的安全性^[17-18]。通过结合外部证据,RAG弥补了大语言模型在参数化记忆方面的缺陷,显著提升了生成答案的事实准确性。

传统RAG以文本为中心的设计理念^[19-20],难以充分挖掘知识图谱等结构化输入中的细粒度信息。GraphRAG等前沿技术通过构建以实体为中心的图结构(节点表示实体,边编码语义关系)强化了结构化推理能力^[21-22]。基于长时记忆机制或知识子图的拓展进一步提升了这一能力^[23-24],同时神经符号等图强化技术也常被引入复杂模型中^[25]。但在处理输入查询时仍将其视为扁平序列,忽略了核心的主谓宾结构、修饰与限定关系等关键语法线索^[26-27]。句法结构能够明确揭示问题焦点、谓词-论元关系及逻辑作用域,对实体消歧与答案构建至关重要。现有语法感知方法多通过句法注意力掩码或基于图卷积、网络熵的图编码器整合结构树^[28-29],但往往需要引入额外编码器或大幅修改模型架构,导致计算开销增加。依存树作为一种常用的句法表示,能有效捕捉词语间的语法关系,例如主谓宾结构和修饰关系。这种结构化的信息对于语义的解析、关系的

抽取或是问答系统需理解问题语义关系的任务起到了重要作用。因此本文利用依存树修改输入图的结构并影响生成过程中的注意力的模式。

大语言模型(如LLaMA-3.1-8B)的全参数微调计算成本高昂,且易引发灾难性遗忘或过拟合^[2]。参数高效微调(PEFT)方法通过调整最少量参数实现任务适配,主要包括两类技术路线:一是Adapter方法,在Transformer层中插入小型瓶颈层;二是BitFit方法,仅更新模型偏置项。两者虽能降低内存占用,但在基于推理的问答等复杂任务中表达能力受限。低秩自适应(LoRA)作为一种高效的PEFT技术^[30],将权重更新分解为两个低秩分量,在不牺牲模型容量的前提下显著减少参数量,与现有架构(LLaMA、GPT等)无缝兼容,支持在计算资源受限(如边缘侧或移动端设备)环境下的快速领域适应^[31-32]。因此,本文选择LoRA进行高效微调。

本文的贡献在于提出一种语法感知GraphRAG框架,将低秩自适应(LoRA)与基于依存句法的结构信息整合到检索增强生成流程中,创新如下:

1) 语法感知的图构建与融合。构建基于依存句法树的句法图 $\mathcal{G}_{\text{syntax}}$,并将其与检索得到的知识图 $\mathcal{G}_{\text{retrieved}}$ 对齐融合为联合图 $\mathcal{G}_{\text{joint}}$ 。通过跨图对齐边集Esr实现语法信息与语义信息的协同约束,使模型在推理过程中同时利用结构化语法先验与外部知识证据。

2) 语法注意力偏置与结构感知编码。在Transformer编码器内部引入基于依存关系的注意力偏置机制,通过语法偏置矩阵 $\mathcal{B}_{\text{syntax}}$ 调节词元间注意力分数,使模型优先关注句法关联词元。同时,采用消息传递机制计算语法感知嵌入,将句法结构在上游阶段即整合至词元表示,与注意力偏置形成互补。

3) LoRA参数高效适配。将LoRA嵌入GraphRAG框架的生成器中,仅更新注意力层的低秩矩阵(秩 $r=8$),在不破坏大模型原有能力的基础上实现领域知识的快速迁移。

1 研究方法

1.1 语法感知 GraphRAG 框架概述

RAG 流程中,给定自然语言问题 q ,检索器 R 从大规模语料库 D 中检索出 top- k 个相关段落,随后将这些段落与问题文本进行拼接,并输入生成器 G 。虽然该架构能够基于外部证据生成答案,但它将检索到的段落视为独立的文本块,忽略了它们之间潜在的结构关联和共享实体^[33]。GraphRAG 通过构建以实体为中心的图结构部分缓解了这一问题,但在处理输入查询时仍将其视为扁平的词元序列,未能利用句法结构中的关键语法线索。因此,本文对 GraphRAG 框架进行了扩展,在检索与生成阶段同步注入依存句法树导出的结构化语言学先验,并通过低秩自适应(LoRA)技术实现大语言模型的高效参数适配。本文框架如图 1,包含知识图、句法图和联合图的构建。

1) 知识图构建:将问题向量化后在知识图谱

中检索,生成由实体和关系构成的知识图 $\mathcal{G}_{\text{retrieved}}$;

2) 句法图构建:对问题进行依存句法分析,生成依存树并通过剪枝构建句法图 $\mathcal{G}_{\text{syntax}}$;

3) 联合图构建:将句法图与知识图对齐融合为联合图 $\mathcal{G}_{\text{joint}}$,利用 LoRA 微调后的 LLaMA-3.1-8B-Instruct 模型进行答案生成。

句法图不仅作为独立的知识源参与检索,而且通过语法注意力偏置机制直接影响生成器的注意力分布,使模型在解码过程中优先关注句法关联词元。同时,LoRA 技术仅微调少量注意力权重,即可在不破坏大模型通用能力的前提下,完成领域知识的快速迁移。

在计算效率方面,本文通过图剪枝和子图选择机制控制计算复杂度。对于完整图结构 $\mathcal{G}$,系统基于启发式规则或可学习的评分函数 $s(v_i)$ 提取任务相关子图 $\mathcal{G}' \subseteq \mathcal{G}$,保留与问题最相关的 top- m 个节点。该剪枝操作既确保了最终图结构的简洁性和下游推理的可处理性,又保留了准确生成答案所需的关键连接。

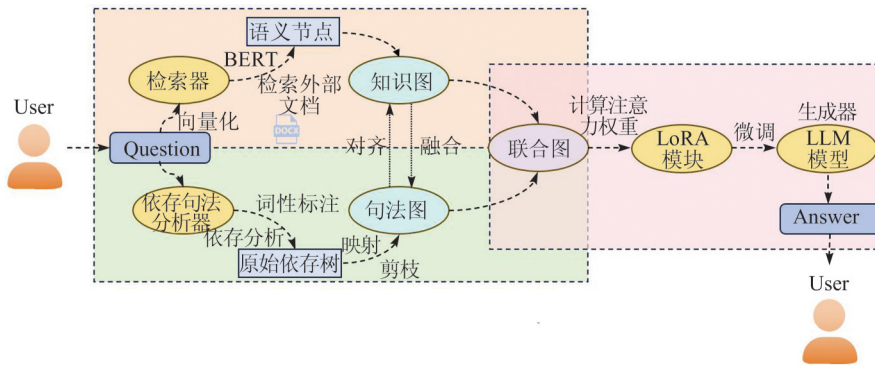


图1 语法感知 GraphRAG 框架示意图

Fig. 1 Diagram of the syntax-aware GraphRAG framework

1.2 基于依存树的语法增强

1.2.1 依存句法分析与句法图构建

将句法结构融入语言模型能为需要细粒度推理的任务提供语法先验。设问题序列 $q = \{q_1, q_2, \dots, q_n\}$, 利用依存句法分析器获得原始依存树 $\mathcal{D} = (\mathcal{N}, \mathcal{R})$, 其中每个节点 $w_i \in \mathcal{N}$ 对应一个词元, 每条边 $(w_i, w_j) \in \mathcal{R}$ 标注有“nsubj”(名词性主语)、“dobj”(直接宾语)、“det”(限定词)或“amod”(状语修饰)等语法关系。如图 2 所示, 这种依存树以紧凑且可解释的方式表征句法关系, 能够有效捕捉论元结构和问题焦点的关键结构线索。

为了将依存结构应用于实际计算, 本文将依存关系转化为句法图 $\mathcal{G}_{\text{syntax}} = (\mathcal{V}, \mathcal{E}_s)$, 其中 $\mathcal{V}_s = \mathcal{N}$ 且 $\mathcal{E}_s = \mathcal{R}$ 。边表示为加权邻接矩阵, 当存在从词元 w_i

到 w_j 的依存边时 $\mathcal{A}_{ij} = 1$, 否则为 0。本文还引入关系标签对应的边类型嵌入 e_{ij} , 这里的 e_{ij} 是对不同依存关系类型进行向量化表示后的嵌入向量, 旨在表征特定的依存关系信息。

如图 3 所示, 左侧的多个 DOCX 文档代表从外部知识库中检索到的相关文本片段。这些文档被输入到一个基于 BERT 架构的编码器中, 该编码器通过序列标注技术识别出文本中的关键实体, 并为其生成富含上下文的 Context Vectors。这些被识别出实体的初始化特征向量由编码器输出的上下文向量经过处理后得到, 进而利用语义节点构建知识图 $\mathcal{G}_{\text{retrieved}}$ 。

1.2.2 图对齐融合策略与注意力偏置机制

图对齐 在图融合阶段, 句法图与检索到的知识图对齐, 形成联合图, 使得语法信息和语义信息

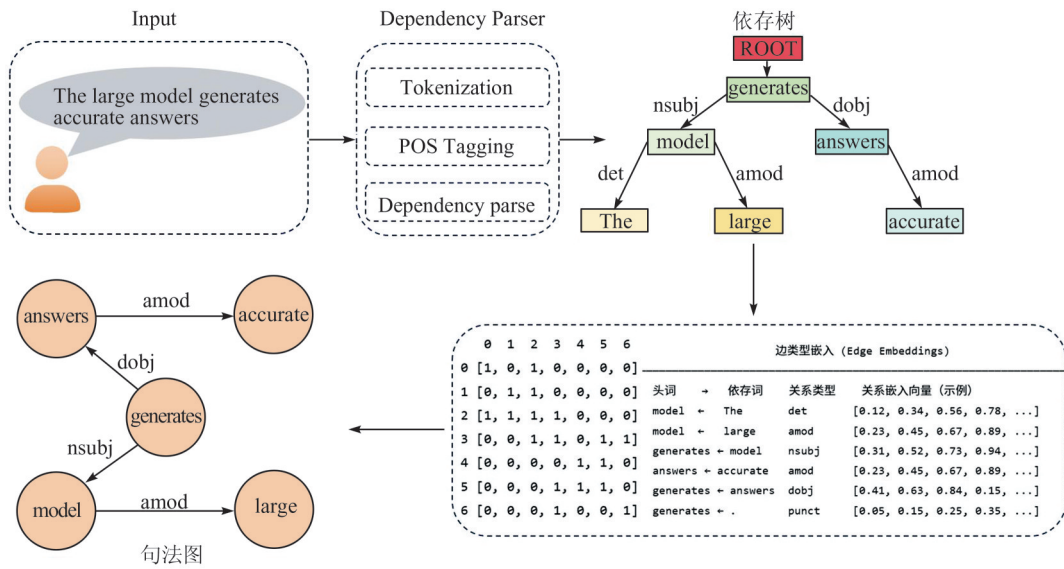


图2 语法感知流程示意图

Fig. 2 Diagram of the syntax-aware pipeline

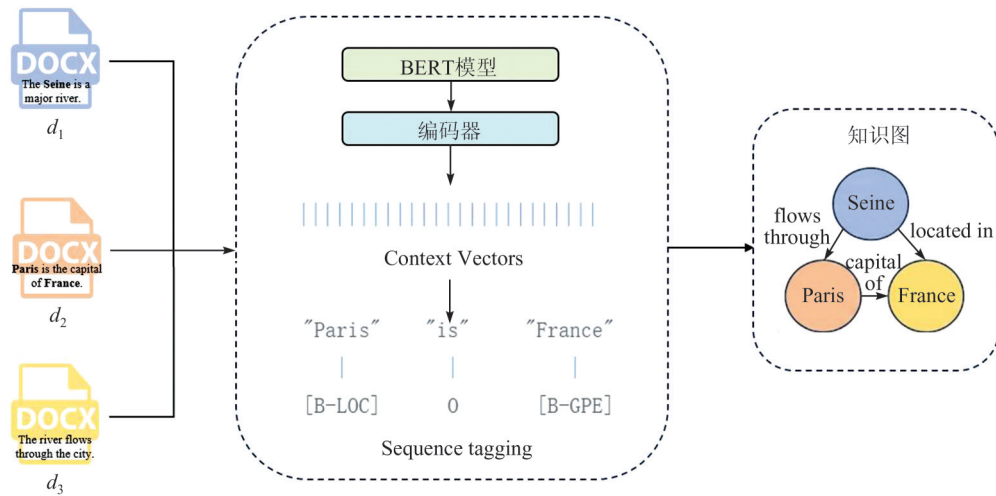


图3 知识图生成流程

Fig. 3 Knowledge Graph Generation Workflow

能够共同约束推理过程。这种融合不仅增强了模型对复杂问题的理解,还提高了检索和生成的准确性。

给定 $\mathcal{G}_{\text{retrieved}} = (\mathcal{V}_r, \mathcal{E}_r)$, 构建联合图 $\mathcal{G}_{\text{joint}} = (\mathcal{V}_s \cup \mathcal{V}_r, \mathcal{E}_s \cup \mathcal{E}_r \cup \mathcal{E}_{sr})$, 其中 $\mathcal{E}_{sr}$ 表示跨图对齐边集。这些对齐边可通过词汇重叠或实体链接函数 $\phi(w_i, e_j)$ 计算得到, 对齐过程依赖于一个预设的对齐阈值 τ , 用于判断两个节点是否足以构成语义匹配, 当 $\phi(w_i, e_j) > \tau$ 时添加边 $(w_i, e_j) \in \mathcal{E}_{sr}$ 。该融合方法构建了一个混合推理空间, 使语法信息和语义信息共同约束推理过程。

$$\phi(w_i, e_j) = \frac{w_i \cdot e_j}{\|w_i\| \cdot \|e_j\|} \quad (1)$$

其中, w_i 和 e_j 均为向量形式, 分别表征相关节点的特征。

联合图的图编码和注意力注入 联合图通过图神经网络(GNN)处理生成节点表征 $H = \text{GNN}(\mathcal{G})$, 并与LLM的隐状态融合。图编码特征通过交叉注意力机制注入解码器:

$$a_t = \text{Attention}(q_t, H) \quad (2)$$

其中, q_t 表示时间步 t 时的解码器查询向量。这种融合机制使生成器能够基于结构化上下文并条件化输出, 从而提升事实准确性和一致性。

语法注意力偏置机制 除图融合外, 本文还实现了语法注意力偏置机制。 A_{ij} 表示为注意力块中词元 w_i 与 w_j 之间的注意力分数, 基于依存树定义

语法偏置矩阵 $S_{ij} \in \{0, 1\}$, 并按以下公式更新注意力分数:

$$A'_{ij} = \frac{(Q_i K_j^T)}{\sqrt{d_k}} + \gamma \cdot S_{ij} \quad (3)$$

其中, γ 为控制语法影响强度的标量超参数。该调整促使模型在上下文编码与解码过程中, 优先关注具有句法关联的词元。

结构感知的词元编码 本文还探索了结构感知的词元编码方法, 通过将句法图嵌入到固定长度的向量中来增强词元表示。对于每个节点 w_i , 采用消息传递机制计算其语法感知嵌入 s_i , 计算公式如下:

$$s_i^{(t+1)} = \sigma \left(\sum_{j \in \mathcal{N}(i)} W_e [h_j^{(t)}; r_{ji}] + b \right) \quad (4)$$

其中, $h_j^{(t)}$ 表示迭代 t 时节点 j 的隐藏状态, r_{ji} 表示从 w_j 到 w_i 的关系嵌入, W_e 为可学习的投影矩阵。最终得到的 s_i 与原始词元嵌入拼接后输入编码器:

$$x_i = [\text{Embed}(w_i); s_i] \quad (5)$$

该方法确保了语法结构在上游阶段就被整合, 与注意力偏置机制形成互补。

边剪枝与动态权重调整 本文通过边剪枝来限制 $\mathcal{G}_{\text{syntax}}$ 的规模。对于每个节点, 按边类型优先级进行保留。 $\tilde{S}_{ij}$ 是经过动态调整后的节点间综合关联度量值。它并非简单的原始关联表示, 而是通过可变的系数对基础关联值进行加权调整得到。边权重 α_{ij} 可以通过语法置信度分数 c_{ij} 进行动态调整, 这样可以使模型既能学习含噪声的解析结果, 又能有效利用语法线索。

$$\tilde{S}_{ij} = \alpha_{ij} \cdot c_{ij} \quad (6)$$

1.3 基于 LoRA 的参数高效微调

在语法感知 GraphRAG 框架中, 生成器需要对融合后的联合图进行深度推理与答案生成。若对 LLaMA-3.1-8B 进行全参数微调, 不仅计算成本高昂, 且易引发灾难性遗忘。因此, 本文采用低秩自适应 (LoRA) 技术实现高效参数适配。

LoRA 基本原理 给定预训练权重矩阵 $W_0 \in \mathbb{R}^{d \times k}$, LoRA 通过添加低秩更新 ΔW 修正前向计算过程, 其数学表达为:

$$W = W_0 + \Delta W = W_0 + AB \quad (7)$$

其中, $A \in \mathbb{R}^{d \times r}$ 和 $B \in \mathbb{R}^{r \times k}$ 为可训练矩阵, 且满足 $r \ll \min(d, k)$ 。这种分解方式既能显著节省参数量, 又能保持模型的表达能力。

输入向量 x 的维度为 d , 如图 4 所示 x 同时流经上下两个分支: 上路分支为冻结的预训练路径, x 与被冻结的“Pretrained weights”矩阵相乘, 保留模型

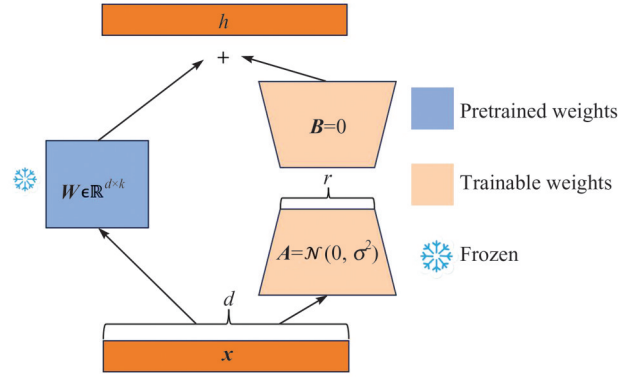


图4 基于 LoRA 的注意力权重调整

Fig. 4 LoRA-based attention weight adaptation

的通用知识且参数在微调过程中不发生改变; 下路分为可训练的 LoRA 旁路通过矩阵 A (投影至低维隐空间) 和矩阵 B (投影回原始空间) 实现低秩变换。空间的维度 r 是 LoRA 的“秩”, 决定了旁路的信息瓶颈宽度。最后, 两条路径的输出在线性层后直接相加, 得到最终的输出 h 。在反向传播时, 梯度仅流经旁路中的 A 和 B 矩阵, 而主路的 W 则保持不变。

在注意力层的实现 $X \in \mathbb{R}^{n \times d}$ 为隐藏状态序列, W_Q 表示原始的查询, 查询投影重新定义为:

$$Q = X(W_Q + A_Q B_Q) \quad (8)$$

其中, $A_Q \in \mathbb{R}^{d \times r}$, $B_Q \in \mathbb{R}^{r \times d}$, 注意力权重的计算公式为:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (9)$$

该方案仅需微调少量新增参数, 完整保留了 LLaMA-3.1-8B 的泛用能力。

训练目标与语法偏置的协同 LoRA 的训练过程通过标准梯度下降法优化低秩矩阵 A 和 B 。训练目标与原始语言建模损失函数保持一致: 给定真实答案序列 $y = \{y_1, \dots, y_T\}$ 和模型预测 $\hat{y}_t = G(q, R(q); \theta_{\text{LoRA}})$, 其损失函数定义为:

$$\mathcal{L}_{\text{LM}} = - \sum_{t=1}^T \log p(\hat{y}_t | \hat{y}_{<t}, q, R(q)) \quad (10)$$

其中, θ_{LoRA} 是 LoRA 特有的参数。检索模块和冻结的基础参数 θ_0 在训练过程中不会更新。

在语法感知 GraphRAG 中, LoRA 与语法偏置协同。 $\mathcal{B}_{\text{syntax}} \in \mathbb{R}^{n \times n}$ 为语法偏置矩阵, 当词元 i 与 j 之间存在依存边时 $\mathcal{B}_{ij} = 1$ 。本文将此偏置注入注意力得分的计算过程的公式为:

$$\tilde{a}_{ij} = \frac{(Q_i K_j^T)}{\sqrt{d_k}} + \gamma \cdot \mathcal{B}_{ij} \quad (11)$$

$$\alpha_{ij} = \frac{\exp(\tilde{\alpha}_{ij})}{\sum_k \exp(\tilde{\alpha}_{ik})} \quad (12)$$

其中, γ 为可训练标量或固定超参数。LoRA 负责学习如何在这种句法约束下进行有效的参数适配, 而无需修改基础模型权重。

秩数选择与效率分析 通过调整低秩矩阵的秩 r 探究模型容量与效率的平衡关系^[26]。实验数据显示, 当 $r \in \{4, 8, 16\}$ 时, 模型在 QALD-h 和 DBpedia QA 数据集上均能保持优异性能。 $P_{\text{LoRA}} = 2dr$ 表示每个投影层新增的可训练参数量, 相较于全参数微调所需的 $P_{\text{full}} = d^2$ 参数, 可获得如下相对缩减比率:

$$\text{Reduction} = \frac{P_{\text{LoRA}}}{P_{\text{full}}} = \frac{2r}{d} \quad (13)$$

效率的提升对于在资源受限环境中适配大模型, 以及尝试语法融合、基于图的编码等多种结构增强策略至关重要。

2 实验

2.1 数据集及参数设置

为了检验本文所提出的语法感知的 GraphRAG 架构在知识密集型问答场景中的实际表现能力。我们选用两个数据集 QALD-h 和 DBpedia QA。

在本文的框架中, 检索组件采用混合检索策略(稀疏检索与密集检索相结合)以最大化召回率和段落相关性。稀疏检索方面, 本文在经过预处理的英文 DBpedia 语料库文本段落上实现了 BM25 (Best Matching 25) 检索器, 查询 q 经过分词处理后通过倒排索引匹配, 获得按词汇相似度排序的候选集 $\mathcal{D}_{\text{BM25}} = \{d_1, d_2, \dots, d_k\}$, 最终检索集合 $\mathcal{D}_{\text{final}}$ 通过合并去重 $\mathcal{D}_{\text{BM25}} \cup \mathcal{D}_{\text{dense}}$ 获得。本文在实验中设 $k=10$ 。密集检索方面, 本文采用双编码器架构, 其中查询编码器 f_q 和段落编码器 f_d 分别生成向量表示 $\mathbf{v}_q, \mathbf{v}_d \in \mathbb{R}^d$, 基于余弦相似度检索 top- k 段落:

$$\text{sim}(q, d) = \frac{f_q^T(q) f_d(d)}{\|f_q(q)\| \cdot \|f_d(d)\|} \quad (14)$$

使用 LoRA 框架进行微调, 设定秩 $r=8$, 针对每个多头注意力块中的查询和值投影层。每层新增可训练参数量为 $2dr$ (其中 $d=4096, r=8$), 总计引入约 130 万个参数, 相当于全模型参数数量的 0.19%。训练配置如下: 批量大小 16, 采用 AdamW 优化器(初始学习率 $\eta = 5 \times 10^{-5}, \beta_1 = 0.9, \beta_2 = 0.999$), 包含 500 步的线性预热, 最大序列长度设为 512 个词元。

为评估语法增强的效果, 在微调过程中注入基

于依存关系的注意力偏置。给定依存树 $\mathcal{D}$, 本文计算偏置矩阵 $\mathcal{B}_{\text{syntax}} \in \mathbb{R}^{n \times n}$, 当存在从词元 w_i 到 w_j 的依存边时 $\mathcal{B}_{ij}=1$ 。

通过消融实验控制 $\gamma \in \{0, 1, 3\}$ 以调节语法结构的影响强度。在所有实验运行中, 解码器均同时对检索上下文和句法图嵌入进行交叉注意力计算。

与本文方法进行对比的基线模型包括: 1) 基于 T5-large 的经典 RAG 流程(仅检索并拼接 top- k 文本片段, 缺乏图关系推理能力); 2) 无语法组件的标准 GraphRAG(采用实体中心图推理的基础版本); 3) 未做结构适配的 LLaMA-3.1-8B-Instruct 全参数微调模型。所有模型均在统一的稀疏+稠密混合检索配置下进行评估^[34]。

采用多维度指标进行评估: 精确匹配 (Exact Match, EM) 用于衡量预测答案与标准答案的完全一致率; 基于知识三元组或答案字符串计算精确率 (Precision)、召回率 (Recall) 和 $F1$; 采用 BLEU 和 ROUGE-L 评估生成答案的流畅性和 n -gram 重叠度; Hits@ k ($k \in \{1, 3, 5\}$) 和平均倒数排名 (MRR) 量化检索质量。最终得分为各指标未加权平均值。评估指标定义如表 1 所示。

表 1 评估指标

Table 1 Evaluation Metrics

指标	定义
EM	字符串匹配准确率
$F1$	词元/实体重叠度
BLEU-4	预测结果与参考结果之间的重叠度
ROUGE-L	最长公共子序列
MRR	正确检索项的平均倒数排名
Hits@1/@3/@5	top- k 检索的准确率

2.2 实验结果

2.2.1 主体问答性能表现

从表 2 可以看出, 本文的语法感知 GraphRAG 框架在多项评估指标上均优于所有基线模型。在 QALD-h 数据集上, 该模型的精确匹配率 (EM) 达到 54.3%, 相较标准 GraphRAG (47.2%) 和 T5-RAG (45.8%) 分别提升 7.1 和 8.5 个百分点; $F1$ 达到 73.4%, 较 GraphRAG+LoRA (71.0%) 提升 2.4 个百分点。在 DBpedia QA 数据集上, 性能提升趋势一致但幅度稍缓, 这很可能源于该数据集以实体为中心的查询语句语法相对简单。尽管如此, BLEU-4 和 ROUGE-L 评分仍反映出更优的语言流畅性和内容匹配精度。此外, 该框架在 Hits@ k 和 MRR 指标上亦维持较高水准, 表明句法信息在检索对齐与聚焦过程中的有效性。

表2显示,在 QALD-h 数据集中,GraphRAG+LoRA 方案 $F1$ 为 71.0%,已超过 GraphRAG+Full FT 方案(68.5%)。这表明 LoRA 在低秩约束下保持模型表达能力的有效性。

基于依存句法的语法增强进一步放大了 LoRA 带来的性能优势。当启用语法注意力机制时,尤其

针对包含嵌套从句、长程依赖或共指关系的复杂问题,效果更明显。以 QALD-h 数据集为例:与 GraphRAG+LoRA 相比,GraphRAG+LoRA+Syntax Bias 方案的 $F1$ 指标提升 2.4 个百分点,ROUGE-L 提升 1.9 个百分点。表明句法结构能够为检索对齐和生成真实度提供价值。

表2 各模型在 QALD-h 与 DBpedia QA 上的整体性能对比

Table 2 Overall performance comparison of various models on QALD-h and DBpedia QA

Model	Dataset	EM	$F1$	BLEU-4	ROUGE-L	MRR	Hits@1	Hits@3	Hits@5
T5-RAG (full FT)	QALD-h	45.8	63.9	21.7	44.5	52.3	41.1	58.0	65.3
GraphRAG (vanilla)	QALD-h	47.2	65.8	22.9	46.1	54.7	43.3	60.2	67.9
GraphRAG+Full FT	QALD-h	50.3	68.5	24.2	48.6	57.1	46.7	63.9	70.8
GraphRAG+LoRA	QALD-h	52.9	71.0	26.5	50.2	59.8	49.5	66.2	72.4
GraphRAG+LoRA+Syntax Bias	QALD-h	54.3	73.4	27.3	52.1	61.4	51.3	68.7	74.1
T5-RAG (full FT)	DBpedia QA	56.4	62.7	28.1	45.9	60.2	53.8	65.6	71.4
GraphRAG (vanilla)	DBpedia QA	57.9	64.5	29.0	47.2	61.8	55.4	67.9	73.2
GraphRAG+Full FT	DBpedia QA	60.3	67.1	30.4	49.8	63.7	58.1	69.4	75.1
GraphRAG+LoRA	DBpedia QA	61.1	68.2	32.0	51.0	65.3	59.7	70.6	76.5
GraphRAG+LoRA+Syntax Bias	DBpedia QA	62.4	69.5	33.7	52.6	66.9	61.2	72.1	77.3

按问题类型对 QALD-h 测试集进行进一步分析,将其划分为事实型、布尔型和组合型三类。如表3所示,语法感知 GraphRAG 在组合型查询上展现出较大的相对优势,EM 指标较基础模型(GraphRAG+Full FT)提升 8.2 个百分点。这一结果说明依存树能有效帮助模型保持多子句输入的结构对应关系。

表3 不同问题类型下不同模型的 EM 指标对比

Table 3 Comparison of EM scores across different models and question types

Model	事实型	布尔型	组合型
GraphRAG (vanilla)	52.4	61.2	37.6
GraphRAG+Full FT	55.9	63.8	42.1
GraphRAG+LoRA	57.1	65.0	45.8
GraphRAG+LoRA+Syntax Bias	59.2	67.1	50.3

综合上述结果,融合了 LoRA 策略的语法感知 GraphRAG 架构,在准确率、文本流畅度以及检索对齐度等多个维度上,均全面超越了基线模型,且这种性能优势并未因数据集特征、问题类型或评估标准的变化而发生波动,表明该方法在复杂场景下具有鲁棒性。

2.2.2 消融实验

为深入理解语法感知建模与参数高效调优的

独立贡献,本文通过系列消融实验对模型各组件进行隔离分析。

1) 语法结构消融。在保留 LoRA 微调的前提下移除语法结构,结果显示(表2)模型性能显著下降。这一现象表明依存句法结构能为检索与生成阶段的协同对齐提供关键引导,其作用远超浅层词共现特征或预训练权重本身。

2) LoRA 秩数分析。在保持其他参数不变的情况下,测试了 $r \in \{4, 8, 16, 32\}$ 的配置方案。如表4所示当 $r \geq 8$ 时模型性能进入平台期,继续增大秩数边际收益显著递减。

当 $r=8$ 时的表现已可与全参数微调媲美,而所需训练参数仅为全量微调的极小部分。该结果不仅验证了主实验选用 $r=8$ 的合理性,还凸显了 LoRA 在低秩约束下保持模型表达力的独特优势。

3) 依存边类型消融。对比三种策略:① 使用

表4 LoRA 秩数对可训练参数量与问答性能的影响

Table 4 Impact of LoRA rank on the number of trainable parameters and QA performance

LoRA Rank(r)	Trainable Params/ 10^6	QALD-h		DBpedia	
		$F1/\%$	EM/ $\%$	$F1/\%$	EM/ $\%$
4	0.66	68.1	50.2	66.2	60.3
8	1.31	71.0	52.9	68.2	61.1
16	2.63	71.5	53.2	68.7	61.6
32	5.25	71.7	53.5	68.9	61.8

完整依存标签(如“nsubj”“dobj”等);② 仅保留无类型句法链接(二值树结构);③ 完全禁用句法信息。如表 5 所示,完整标签配置性能最优。

表 5 不同句法策略对问答性能的消融实验结果
Table 5 Ablation study on the effect of different syntactic strategies on QA performance

策略	QALD-h		DBpedia	
	F1	EM	F1	EM
完全禁用句法信息	71.0	52.9	68.2	61.1
无类型句法链接	72.1	53.7	68.8	61.8
完整依存标签	73.4	54.3	69.5	62.4

4) 语法注意力头覆盖率分析。如表 6 所示,总的来看,模型的性能随语法整合范围的扩大而提升,但覆盖率在 75% 之前提升更快。覆盖率在 25% 和 50% 时,语法影响注意力头有限,会导致注意力分布碎片化,降低检索结果与生成序列的连贯性;而当覆盖率超过 75% 继续扩展至 100% 时,边际收益有限且计算开销显著增加。因此,采用 75% 覆盖率可在准确率与结构覆盖率之间实现平衡。

表 6 语法注意力头覆盖率对模型性能的影响
Table 6 Impact of syntactic attention head coverage on model performance

Syntax head coverage	QALD-h		DBpedia	
	F1	EM	F1	EM
25% Heads	70.5	52.7	67.8	60.7
50% Heads	71.6	53.4	68.4	61.3
75% Heads	72.9	54.1	69.2	62.0
100% Heads	73.4	54.3	69.5	62.4

5) 鲁棒性测试。本文对模型进行了对抗性扰动实验:输入问题保持语义等价但改变句法结构(如改为被动语态、调整从句顺序)。语法增强版本的 F1 仅下降 1.1 个百分点,而非语法版本则下降 3.7 个百分点。这表明模型通过显式编码句法结构,获得了对语言形式变化的不变性特征,在面临改写或噪声输入时表现更加稳健。

综合消融实验验证了各模块的必要性:语法感知建模在所有指标和语言现象中均带来明确增益;LoRA 秩数分析证实低秩自适应能实现效率与性能的最优平衡;细粒度语法类型与广泛头部覆盖则是充分提取依存结构价值的关键要素。这些从实证角度支撑了模型架构的核心设计决策。

2.3 案例分析

为补充量化实验结果,本文通过一系列案例分

析展示语法感知建模对检索与生成的双重提升效果。研究聚焦包含嵌套从句、修饰语歧义或多实体指代的结构复杂查询^[26],所有案例均选自 QALD-h 测试集,通过对比有无语法增强的模型内部行为可视化,发现依存结构的引入能产生更聚焦的注意力分布和更高保真的答案内容,尤其在语言歧义或长程依赖场景下表现突出。

案例 1 多重约束查询。对于复杂查询“Which river that flows through Paris is longer than 500 kilometers”,基线模型受到“longer”一词的强干扰,召回的文档是“The Loire is France’s longest river”,因此错误输出“Loire”。本文框架则通过依存句法分析精准锁定了包含“The Seine flows through Paris…”的关键语料,给出了正确答案“Seine”。该案例表明,当面对包含多重逻辑约束的查询时,句法信息能够更高效地穿透表层语义的迷惑,提升检索与生成的双重准确性。

案例 2 布尔型查询。对于查询“Are the capital of Germany and Paris in the same time zone”,基础模型仅检索到关于柏林和巴黎的通用文本,无法给出明确结论。本文的框架通过“located”和“time zone”之间的依存关系引导检索器与解码器,获取更具时间针对性的地理事实,最终生成正确答案“Yes”。

案例 3 多重约束实体匹配。通过对比基线模型和本文提出的方法,展示了在实体识别和匹配准确性上的提升。对于查询“哪位科学家发现了青霉素并在 1945 年获得了诺贝尔奖”,基线模型错误地匹配了 Howard Florey(青霉素临床应用的推动者),而本文的方法通过依存结构正确识别出“发现青霉素”与“1945 年获诺贝尔奖”两个约束条件,给出正确答案 Alexander Fleming。这一结果验证了本文方法在处理多约束查询时的实体识别与匹配能力。

案例 4 并列结构查询。对于问题“List two European cities that are both capitals and have large international airports”,基础模型因注意力稀释仅正确生成一个城市,而本文框架通过句法分组对每个并列成分保持独立注意力聚焦,最终以高置信度输出“Paris and Berlin”。生成结果严格遵循依存树的结构约束,显著提升了列举答案的连贯性与完整性。

案例 5 介词依附查询。对于“Which museum in New York was founded in 1870”,无语法引导的模型会检索纽约所有博物馆信息,最终无法生成有效答案,而本文框架通过“museum”与“founded”的依存对齐,将检索范围精准限定至大都会艺术博物馆的相关事实,并正确输出答案。

上述案例充分验证了句法信息对知识检索与响应生成的引导价值^[19]。通过依存结构线索与语义内容的对齐,模型生成的答案不仅准确,而且结构合理、符合用户意图,在长句、复合句或从句密集型问题中表现尤为显著。

3 结 语

本文提出了一种语法感知 GraphRAG 框架,将低秩适配(LoRA)与基于依存句法的结构信息整合到检索增强生成流程中。通过 LoRA 对模型进行微调,并注入从依存句法树中导出的句法信息,本文的方法在数据集 QALD-h 与 DBpedia QA 进行了测试,多个指标得到了提升,而且需训练参数明显减少。消融实验表明,语法引导显著提升了检索精度与生成的连贯性,尤其在结构复杂查询上表现突出。未来工作将进一步扩展该框架,以支持更为复杂的问题查询,并且可以研究垂直领域,从而服务于更多的现实场景之中。

参考文献:

- [1] XIONG H Y, BIAN J, LI Y C, *et al.* When search engine services meet large language models: Visions and challenges [J]. *IEEE Transactions on Services Computing*, 2024, **17** (6): 4558-4577. DOI:10.1109/TSC.2024.3451185.
- [2] SU W H, TANG Y C, AI Q Y, *et al.* Parametric retrieval augmented generation[EB/OL]. 2025: arXiv: 2501.15915. <https://arxiv.org/abs/2501.15915>.
- [3] POLIAKOV M, SHVAI N. Multi-meta-RAG: Improving RAG for Multi-hop queries using database filtering with LLM-extracted metadata [C]//Information and Communication Technologies in Education, Research, and Industrial Applications. Cham: Springer, 2025: 334-342. DOI:10.1007/978-3-031-81372-6_25.
- [4] 任海玉, 刘建平, 王健, 等. 基于大语言模型的智能问答系统研究综述[J]. 计算机工程与应用, 2025, **61**(7): 1-24. DOI:10.3778/j.issn.1002-8331.2409-0300.
REN H Y, LIU J P, WANG J, *et al.* Research on intelligent question answering system based on large language model [J]. *Computer Engineering and Applications*, 2025, **61** (7): 1-24. DOI: 10.3778/j.issn.1002-8331.2409-0300 (Ch).
- [5] 成志宇, 陈星霖, 王菁, 等. 一种基于知识图谱的检索增强生成情报问答技术[J]. 计算机科学, 2025, **52**(1): 87-93. DOI:10.11896/jsjx.240900064.
CHENG Z Y, CHEN X L, WANG J, *et al.* Retrieval-augmented generative intelligence question answering technology based on knowledge graph[J]. *Computer Science*, 2025, **52** (1): 87-93. DOI: 10.11896/jsjx.240900064 (Ch).
- [6] BARRON R C, GRANTCHAROV V, WANNA S, *et al.* Domain-specific retrieval-augmented generation using vector stores, knowledge graphs, and tensor factorization [C]//2024 International Conference on Machine Learning and Applications (ICMLA). New York: IEEE Press, 2025: 1669-1676. DOI: 10.1109/ICMLA61862.2024.00258.
- [7] ZHAO X J, LIU S Y, YANG S Y, *et al.* MedRAG: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot [DB/OL]. [2025-10-10]. <https://dl.acm.org/doi/abs/10.1145/3696410.3714782>.
- [8] JIANG X K, FANG Y, QIU R H, *et al.* TC-RAG: Turing-complete RAG's case study on medical LLM systems[EB/OL]. 2024: arXiv: 2408.09199. <https://arxiv.org/abs/2408.09199>.
- [9] PIYA F L, BEHESHTI R. Advancing feature extraction in healthcare through the integration of knowledge graphs and large language models[DB/OL].[2025-10-10]. <https://ojs.aaai.org/index.php/AAAI/article/view/35224>.
- [10] YANG H, LI J X, ZHANG C, *et al.* Large language model-driven knowledge graph construction in sepsis care using multicenter clinical databases: Development and usability study[J]. *Journal of Medical Internet Research*, 2025, **27**: e65537. DOI:10.2196/65537.
- [11] FAN H L, LIU C S, JANVISLOO N E, *et al.* MaViLa: Unlocking new potentials in smart manufacturing through vision language models [J]. *Journal of Manufacturing Systems*, 2025, **80**: 258-271. DOI:10.1016/j.jmsy.2025.02.017.
- [12] WEI C Y, SHI Y T, MU W B, *et al.* Large language models assisted materials development: Case of predictive analytics for oxygen evolution reaction catalysts of (oxy)hydroxides[J]. *ACS Sustainable Chemistry & Engineering*, 2025, **13**(14): 5368-5380. DOI:10.1021/acssuschemeng.5c00798.
- [13] WU C K, DING W J, JIN Q S, *et al.* Retrieval augmented generation-driven information retrieval and question answering in construction management [J]. *Advanced Engineering Informatics*, 2025, **65**: 103158. DOI:10.1016/j.aei.2025.103158.
- [14] YIN J Q, HINES J, HERRON E, *et al.* ChatHPC: Empowering HPC users with large language models [J]. *The Journal of Supercomputing*, 2024, **81**(1): 194. DOI:10.1007/s11227-024-06637-1.
- [15] HINDI M, MOHAMMED L, MAAZ O, *et al.* Enhancing the precision and interpretability of retrieval-augmented

- generation (RAG) in legal technology: A survey[J]. *IEEE Access*, 2025, **13**: 46171–46189. DOI:10.1109/ACCESS.2025.3550145.
- [16] DONG H M, WANG W B, SUN Z J, *et al.* Knowledge graph construction for intelligent cockpits based on large language models[J]. *Scientific Reports*, 2025, **15**: 7635. DOI:10.1038/s41598-025-92002-y.
- [17] ZHENG Q X, CHEN M R, SHARMA P, *et al.* EvAlignUX: Advancing UX evaluation through LLM-supported metrics exploration [DB/OL]. [2025-10-10]. <https://dl.acm.org/doi/full/10.1145/3706598.3714045>.
- [18] LIU M R, ZHANG S X, LONG C. Mask-based membership inference attacks for retrieval-augmented generation [DB/OL]. [2025-10-10]. <https://dl.acm.org/doi/abs/10.1145/3696410.3714771>.
- [19] PENG B C, ZHU Y, LIU Y C, *et al.* Graph retrieval-augmented generation: A survey[EB/OL]. 2024: arXiv: 2408.08921. <https://arxiv.org/abs/2408.08921>.
- [20] PROCKO T T, OCHOA O. Graph retrieval-augmented generation for large language models: A survey [C]// 2024 Conference on AI, Science, Engineering, and Technology (AIXSET). New York: IEEE Press, 2024: 166–169. DOI:10.1109/AIXSET62544.2024.00030.
- [21] EDGE D, TRINH H, CHENG N, *et al.* From local to global: A graph RAG approach to query-focused summarization[EB/OL]. 2024: arXiv: 2404.16130. <https://arxiv.org/abs/2404.16130>.
- [22] HE X X, TIAN Y J, SUN Y F, *et al.* G-Retriever: Retrieval-augmented generation for textual graph understanding and question answering[J]. *Advances in Neural Information Processing Systems*, 2024, **37**: 132876–132907.
- [23] GU Y, GUTIÉRREZ B, SHU Y H, *et al.* HippoRAG: Neurobiologically inspired long-term memory for large language models[DB/OL]. [2025-10-10]. https://proceedings.neurips.cc/paper_files/paper/2024/hash/6ddc001d07ca4f319af96a3024f6dbd1-Abstract-Conference.html.
- [24] ZENG Z F, CHENG Q, HU X C, *et al.* KoSEL: Knowledge subgraph enhanced large language model for medical question answering[J]. *Knowledge-Based Systems*, 2025, **309**: 112837. DOI:10.1016/j.knosys.2024.112837.
- [25] DE GIORGIS S, GANGEMI A, RUSSO A. Neurosymbolic graph enrichment for grounded world models[J]. *Information Processing & Management*, 2025, **62** (4): 104127. DOI:10.1016/j.ipm.2025.104127.
- [26] HO C T, REN H X, KHAILANY B. VerilogCoder: Autonomous verilog coding agents with graph-based planning and abstract syntax tree (AST)-based waveform tracing tool[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, **39** (1): 300–307. DOI:10.1609/aaai.v39i1.32007.
- [27] ZHANG Y Y, YANG R Z, XU X Q, *et al.* TELE-Class: Taxonomy enrichment and LLM-enhanced hierarchical text classification with minimal supervision [C]// Proceedings of the ACM on Web Conference 2025. New York: ACM, 2025: 2032–2042. DOI:10.1145/3696410.3714940.
- [28] 罗妹秋, 张春霞, 彭成, 等. 基于解析图嵌入和加权图卷积网络的知识图谱补全[J]. *中国科学: 信息科学*, 2022, **52**(11): 2037–2057.
- LUO M Q, ZHANG C X, PENG C, *et al.* Knowledge graph completion based on parsing graph embedding and a weighted graph convolutional network[J]. *Scientia Sinica (Informationis)*, 2022, **52**(11): 2037–2057 (Ch).
- [29] XIAN Y T, LI P, PENG H, *et al.* Community detection in large-scale complex networks *via* structural entropy game[C]//Proceedings of the ACM on Web Conference 2025. New York: ACM, 2025: 3930–3941. DOI:10.1145/3696410.3714837.
- [30] HU E J, SHEN Y L, WALLIS P, *et al.* LoRA: Low-rank adaptation of large language models[EB/OL]. 2021: arXiv: 2106.09685. <https://arxiv.org/abs/2106.09685>.
- [31] ZHENG Y, CHEN Y H, QIAN B, *et al.* A review on edge large language models: Design, execution, and applications[J]. *ACM Computing Surveys*, 2025, **57** (8): 1–35. DOI:10.1145/3719664.
- [32] CHEN D H, LIU Y H, ZHOU M Y, *et al.* LLM for mobile: An initial roadmap[J]. *ACM Transactions on Software Engineering and Methodology*, 2025, **34** (5): 1–29. DOI:10.1145/3708528.
- [33] SARMAH B, MEHTA D, HALL B, *et al.* HybridRAG: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction [C]// Proceedings of the 5th ACM International Conference on AI in Finance. New York: ACM, 2024: 608–616. DOI:10.1145/3677052.3698671.
- [34] WAN Y W, CHEN Z Y, LIU Y, *et al.* Empowering LLMs by hybrid retrieval-augmented generation for domain-centric Q&A in smart manufacturing[J]. *Advanced Engineering Informatics*, 2025, **65**: 103212. DOI:10.1016/j.aei.2025.103212.

□