基于大语言模型的用户目标导向小众词汇主题挖掘方法

曾祥[1]，王晔[2]，杨博捷[1]，周斌[1]，王志超[3]，贾焰[1]
（1. 国防科技大学计算机学院，湖南 长沙 410073；2. 哈尔滨工业大学（深圳）计算机科学与技术学院，广东 深圳 518055；
3. 湖南四方天箭信息科技有限公司，湖南 长沙 410006）

摘要：针对小众主题词挖掘领域黑话语义隐晦、新兴词语料稀缺及语义漂移显著等问题，提出了SeedLLM框架。该方法通过跨上下文一致性约束对黑话术语进行全局推理以刻画其深层语义，并结合检索增强生成补充新兴词外部语义，同时设计新兴词评估模型与在线强化学习机制以抑制过度检索，实现对语义演化的动态适应。多领域数据集实验结果表明，所提方法在小众主题词识别与主题聚合任务上优于现有方法，具有良好的泛化能力与应用价值。

关键词：目标主题挖掘；大语言模型；强化学习；检索增强生成
中图分类号：TP391.1
文献标志码：A
doi：10.11959/j.issn.1000–436x.TXXB260132

User-targeted rare word topic mining method based on large language models

Zeng Xiang[1], Wang Ye[2], Yang Bojie[1], Zhou Bin[1], Wang Zhichao[3], Jia Yan[1]
1. College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China
2. School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Shenzhen 518055, China
3. Hunan Seefore Information Technology Co., Ltd., Changsha 410006, China

Abstract: To address challenges in niche topic word mining, including implicit slang semantics, scarce corpora for emerging terms, and significant semantic drift, the SeedLLM framework was proposed. It introduced cross-context consistency constraints to perform global reasoning over slang terms and capture their deep semantics. Retrieval-augmented generation was further employed to supplement external semantics for emerging words. Meanwhile, an emerging-word evaluation model and an online reinforcement learning mechanism were designed to suppress excessive retrieval and dynamically adapt to semantic evolution. The experimental results on multiple cross-domain datasets demonstrate that the proposed method outperforms existing methods in niche topic word identification and topic aggregation, showing strong generalization and practical value.

Key words: target topic mining, large language model, reinforcement learning, retrieval-augmented generation

0 引言

随着社交媒体与在线平台的快速发展，舆论信息呈现出高噪声、强动态与语义多样化特征，使目标议题的准确识别与语义结构刻画面临挑战。因此，从海量非结构化文本中挖掘与特定目标相关的主题信息成为网络文本分析的重要研究问题。传统主题挖掘方法[1-3]如基于潜在狄利克雷分配的主题模型（latent Dirichlet allocation, LDA）、BERTopic多采用完全无监督建模，难以直接聚焦用户关注的特定议题。为此，研究者提出了面向用户目标的主

收稿日期：2026–03–20；修回日期：2026–05–20
通信作者：王晔，wangye2020@hit.edu.cn
基金项目：国家重点研发计划基金资助项目（No.2023YFC3306103）
Foundation Item：The National Key Research and Development Program of China (No.2023YFC3306103)

题挖掘方法[4-5]，通过引入用户提供的种子词对目标语义空间进行约束，从而在大规模语料库中挖掘与目标议题相关的语义主题。具体而言，该类方法通常以同一语义维度的一组种子词作为参考，如在“俱乐部”语义维度下给定种子词集合{皇马，巴萨，拜仁}，其中“皇马”作为目标锚点，“巴萨”和“拜仁”作为对比种子，通过构建二者之间的语义区分关系，挖掘与目标种子词高度相关且与其他种子词明显区分的判别性主题词，从而刻画目标议题的语义特征，并为议题治理中的定向识别与线索聚合提供支撑[6-8]。

现有种子引导主题挖掘方法[5,9]通常将主题词发现建模为“语义表征加相似性匹配”问题：先通过 skip-gram 或 BERT 学习词汇表征，再以种子词为锚点，在语义空间中选取相似词作为同一主题。该范式隐含假设词汇语义可由语料库内部信息充分表征，从而将主题词发现简化为既有表示空间中的静态语义匹配过程，而不需要在建模过程中动态获取额外信息。这一范式在高频词场景下通常有效，但在小众主题词挖掘中存在明显局限。小众主题词（如领域黑话、新兴词）普遍具有低频、语义隐晦及动态演化等特征。领域黑话单一语境信息有限，语义依赖跨语境理解。新兴词的语义在早期阶段尚未稳定，且随事件演化快速变化，语料库内部可获得的语义证据有限，难以支撑可靠建模，因而需要通过外部信息进行语义信息补充。但现有方法仍属于静态语义建模，缺乏动态补充外部信息的能力，因而难以刻画低频与演化语义，限制了挖掘性能。

为缓解新兴词内部语义信息不足的问题，基于大语言模型（large language model, LLM）的检索增强生成（retrieval-augmented generation, RAG）方法通过从外部语料库检索补充信息，为动态语义建模提供了一种可行路径。现有 RAG 方法通过引入检索触发机制（如基于置信度[10-11]、忠实性[12]或辅助判别模型[13]）决定何时进行信息获取，从而提升对复杂语义的处理能力。由于其面向语义稳定的问答任务，语料库语义分布在推理阶段保持不变，固定规则或预训练模型已能满足检索需求。然而，在新兴词占比较高的小众语义场景中，词汇语义随事件持续演化，触发决策本质上依赖当前语义的不确定性状态，静态触发策略易导致检索不足或检索过度，进而影响对新兴词语义的建模效果。

为应对领域黑话与新兴词带来的小众主题词挖掘难题，本文提出了面向用户目标的 SeedLLM 框架。以“语义证据获取与决策”为统一主线，针对小众主题词语义在单一上下文中信息不足且跨语境不一致的问题，引入跨上下文一致性约束以实现跨文本内部语义聚合。针对新兴词语料稀缺的问题，引入检索增强机制以补充外部语义证据。本文的核心创新在于将“是否检索”从固定触发机制重构为可学习的自适应决策过程。不同于现有方法依赖静态规则或预训练判别模型进行检索触发，本文构建新词判别模型，并以语义增益为奖励信号，通过在线强化学习对触发策略持续优化，使检索决策能够随新兴词语义演化动态调整，从而在建模机制上与现有静态触发范式形成本质区别，有效避免检索不足与过度检索。在获得高置信语义后，模型基于种子词对候选术语进行主题指派，挖掘与用户目标对齐的主题。

本文研究工作的主要贡献如下。

（1）提出了一种面向用户目标的小众主题词挖掘框架。针对种子引导式主题发现中小众词挖掘困难的问题，系统梳理中文语境下小众词识别与语义刻画的关键挑战，并提出相应的解决思路。

（2）提出了面向小众词的差异化语义建模方法：针对领域黑话，引入跨上下文一致性约束进行全局联合推理；针对新兴词，引入检索增强机制补充语义证据。进一步，将检索触发建模为可学习的自适应决策过程，通过新兴词判别模型与在线强化学习实现检索策略的动态优化，从而自适应语义演化并抑制过度检索。

（3）开展跨领域实验验证本文方法的有效性。在 4 个跨领域数据集上基于入侵检测任务（intrusion detection task, IDT）、主题内部多样性（intra-topic diversity, IntraTD）、主题词平均准确率（mean accuracy,MACC）等指标及下游任务进行综合评估，并在中医文献与足球社区等中文语料场景中进一步验证所提方法的适配性。实验结果表明，所提方法在小众主题词识别与语义刻画方面具有良好的泛化能力。

1　相关工作

1.1　面向用户目标的主题挖掘

目前，针对面向用户目标的主题挖掘，现有研

究大致可以分为两类：目标主题建模[4]（targeted topic modeling, TTM）与种子引导式主题发现[14]（seed-guided topic discovery, SGTD）。在目标主题建模方面，SeedLDA[15]是较早的经典方法，通过在LDA模型中融入种子词，对主题与词汇分布施加约束，使模型倾向于生成与种子词语义相关的主题，从而实现用户兴趣导向的主题挖掘。随后，一些研究在此基础上进一步改进主题模型结构。例如，TTM[4]通过在概率主题模型中引入种子词信息，对主题–词分布进行偏置，以强化用户关注主题的生成能力。GTM[16]针对短文本语料中普遍存在的数据稀疏问题，通过选择性过采样策略缓解稀疏性，从而提升目标主题挖掘效果。Foto[9]则将种子词视为伪文档，通过构建与真实文档的关联关系来辅助相关主题及文档的提取与可视化。在种子引导式主题发现方面，CatE[17]通过结合词嵌入相似度与分布特异性，检索具有类别代表性的主题词。SeedTopicMine[14]在CatE的基础上，通过融合多种术语嵌入信息来挖掘具有判别性的主题术语，在通用数据集上取得了优异性能。DiskTM-LK[5]利用大语言模型提取术语并生成目标主题的种子词，将其注入词汇实值多项分布以构建先验知识矩阵，并在模型学习过程中通过主题–词分布的主题层约束实现与预定义关键词对齐，从而挖掘用户感兴趣的主题。

1.2 基于大语言模型的主题发现方法

近年来，随着LLM的快速发展，越来越多研究开始尝试利用其强大的语义理解与生成能力进行主题发现。TopicGPT[18]利用LLM的自然语言生成与语义理解能力，通过提示工程引导模型挖掘潜在主题，其挖掘的主题与人类分类结果具有更高的契合度。在此基础上，一些研究进一步探索LLM在主题建模中的结构化应用。DeTiME[19]提出了一种基于扩散机制的主题建模框架，通过结合编码器–解码器结构的LLM来增强主题生成过程，从而提升主题发现的稳定性与表达能力。PVTM[20]则关注短文本场景，利用LLM进行上下文扩展，并结合前缀调优的变分自编码器，在语义信息稀疏的短文本中挖掘更有意义和相关的主题。此外，Doi等[21]利用LLM的预训练上下文学习能力进行主题建模，并提出并行与顺序提示两种策略，以缓解LLM输入长度限制的问题，从而在主题一致性和多样性方面相较传统主题模型取得更好的表现。

总体来看，现有方法多关注主流主题的语义模式，对用户关注主题中的黑话与新兴词等小众词汇在识别与语义刻画方面仍显不足，尤其在结合种子词实现小众词语义定位与演化建模方面缺乏系统方法。

2 问题定义

遵循Meng等[17]的定义，面向用户目标的主题挖掘是指给定语料库D和一组种子词seeds = $(\text{seed}_1,\text{seed}_2,\cdots,\text{seed}_N)$，为每个种子$s_i$找一组术语集合$T_i$以形成连贯的主题。其中，种子词一般为主题名，可为单字词或短语。集合T_i中的术语在语义上应当接近种子词s_i代表的主题，同时远离其他种子词$s_j(j \neq i)$。换言之，每个种子词s_i表示一个语义主题，任务的核心是从语料库D中挖掘属于种子词s_i的区分性术语集合T_i。

3 方案设计

本文提出的小众主题挖掘模型SeedLLM的整体框架如图1所示。本文从“语义证据获取与决策”的统一视角对小众主题词挖掘过程进行建模，将目标主题挖掘过程视为一个由证据获取、语义推理与决策控制共同驱动的动态过程。在该视角下，模型由4个相互耦合的模块构成。

（1）术语提取模块作为语义建模的起点，从语料库中识别候选术语，为语义推理提供基本单元。

（2）跨上下文语义一致性推理模块通过聚合术语在不同上下文中的语义证据，在跨上下文一致性约束下对小众词进行联合推理，从而完成对语料库内部语义信息的充分挖掘与整合。

（3）新兴词检索增强推理模块进一步扩展语义证据来源，将“是否检索外部信息”建模为一个可学习的决策过程：模型基于当前语义状态判断证据是否不足，并在必要时触发外部检索，同时通过在线强化学习以“语义增益”为反馈持续优化决策策略，从而实现对语义演化的自适应调节。

（4）种子引导主题分配模块在获得高置信度语义后，将术语语义映射至用户关注的主题空间，完成主题指派。

下面将详细介绍这4个核心模块。

3.1 术语提取

为了有效提取术语的全面表征，本文利用大语言模型的指令遵循能力和少样本学习能力，通过提

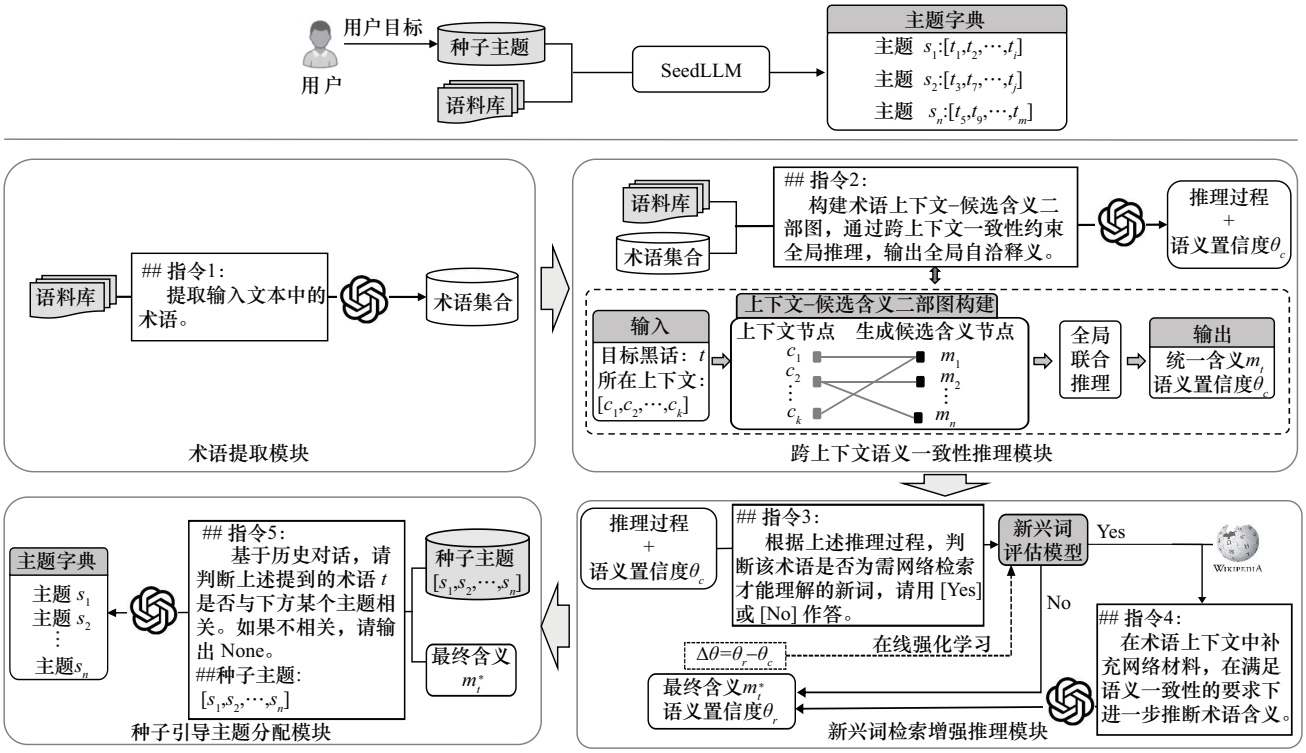

图1　SeedLLM 整体框架

供相关语料库的术语提取例子，提示大语言模型从语料库中提取所有术语，使模型能够准确提取罕见词的精确形式。此模块使用的提示词如图2所示。

Prompt 1: 术语提取

##指令：提取下面输入文本中所包含的所有术语

##输入文本：{ Input_text }

##样例：

输入：{“这些伪球迷，还在挺赫伊森”}

输出：[“伪球迷”，“赫伊森”]

图2　术语提取提示词

具体而言，本文首先使用提示词Prompt 1提示LLM从输入文本中提取其所含术语。然后，本文遍历整个语料库以获取完整的术语集合 $T=\{T_1,T_2,\cdots,T_n\}$。最后，本文对每个术语执行反向匹配以识别其在文档中的上下文，得到相应的上下文集合 $C=\{C_1,C_2,\cdots,C_n\}$。

3.2　跨上下文语义一致性推理

完成术语提取后，为支撑后续基于术语语义的主题分配任务，需进一步获取每个术语更为完整的语义内涵。小众术语往往散布于大量碎片化的上下文中，单一上下文难以提供充分的语义信息，导致现有方法对术语含义的理解存在片面性与不完整性。为此，本文提出了一种面向小众术语的跨上下文一致性语义推理思维链方法，通过构建跨上下文一致性约束的思维链，驱动模型主动关联并聚合多个上下文中的语义信息，从而形成对术语含义更为完整和自洽的综合理解。具体思维链设计如图3所示。

Prompt 2: 小众术语跨上下文一致性语义推理思维链

##指令：请遵循下面思维链，为目标黑话术语构建上下文-候选含义二部图，利用跨上下文一致性约束进行全局联合推理，输出一组在所有语境下都自洽无矛盾的统一含义。

1. 切分语境：收集含黑话的多条语境，标出黑话句与共现词。
2. 抽取语境锚点信号，提取指示、评价词汇并划分强弱锚点。
3. 生成术语候选义集：给出术语可能的后选含义$m_1,\cdots,m_n$。
4. 建二部图：左侧语境节点$\{c_i\}$，右侧术语候选含义节点$\{m_j\}$。
5. 设一致性约束：要求含义对象一致、语域一致、覆盖强锚点。
6. 全局联合校验筛选：剔除锚点覆盖不足、语域不符的语义。
7. 输出统一释义：明确核心内涵并阐释其对各语境的适配逻辑。

##目标黑话：[term t_i]

##黑话术语所在上下文:[c_1,c_2,⋯, c_n]

样例：example$_1$

图3　小众术语跨上下文一致性语义推理思维链

首先输入选取术语 t_i 及其上下文集合，然后让大语言模型根据集合 $C=\{C_1,C_2,\cdots,C_n\}$ 中上下文语境给出候选语义集 $M=\{M_1,M_2,\cdots,M_n\}$，随后构

造对应术语上下文–候选含义二部图G_i。结合此二部图的上下文一致性约束推理出黑话术语的全局深层次语义m_i^c及其对应大语言模型的输出置信度k_i^c。形式化为：

$$m_i^c,k_i^c = \mathrm{LLM}(t_i,c_i,\mathrm{Prompt\,2}) \tag{1}$$

3.3 新兴词检索增强推理

针对语料库中出现的新兴词，现有语料往往缺乏必要的背景知识与上下文线索，所以术语语义推断需要借助域外检索补充信息以完善语义表征与推理依据。然而，若对所有术语不加区分地执行域外检索，容易引入与任务无关的噪声，干扰语义推理与下游主题分配效果。因此，有必要训练能够识别哪些术语属于新兴词并需要外部信息支撑的新兴词判别模型；同时，该模型应能够随语义演化在线自适应更新其判别准则。

对此，本文提出了基于在线强化学习的新兴词检索增强推理框架。受Self-RAG[13]工作启发，本文训练了一个生成反思词元的评估模型，用于判断术语是否遇到了新术语并需要外部检索。若判定术语t_i是新兴词，则进一步检索最新互联网信息以补充LLM对该术语的理解。具体检索增强过程如下：首先使用术语t_i作为关键词，并使用检索器（Bing等）获取其对应的外部信息r_i。随后，本文使用新兴词检索增强推理提示词Prompt 3补充LLM对术语的理解，如图4所示。通过检索增强获得的语义表示m_i^r及其对应大语言模型的输出置信度k_i^r分别为：

$$\begin{aligned} r_i &= \mathrm{Retriever}(t_i) \\ m_i^r,k_i^r &= \mathrm{LLM}(c_i,r_i,\mathrm{Prompt\,3}) \end{aligned} \tag{2}$$

Prompt 3: 检索增强的新兴词语义补充推理

##指令：你将获得一个术语上下文和其对应补充互联网相关资料。你知道这两段文本中提到的术语吗?

##术语: [t_i]
##上下文: [$c_1,c_2,\cdots,c_n$]
##网络资料: [$w_1,w_2,\cdots,w_m$]

##样例:

输入:
指令：你知道上下文和网络资料中提到的“赫伊森”吗？
上下文:
[“赫伊森自己眼睛也不大吧”，“赫伊森也配歧视亚洲人”]
网络资料:
[“迪恩·赫伊森（Dean Huijsen），男，拥有荷兰和西班牙双重国籍，西班牙足球运动员，司职中后卫，现效力于皇家马德里足球俱乐部…”]

输出：是的，赫伊森指的是西班牙中后卫迪恩·赫伊森，现效力于皇家马德里。

图4 检索增强的新兴词语义补充推理提示词

为了训练新兴词评估模型，本文从黑话术语跨上下文一致性约束的全局推理模块中随机采样出5k个输入输出对(Q_c,A_c)，并输入提示词Prompt 4提示大语言模型使其做评估工作，新兴词判别器提示词如图5所示。该指令后接少量示例I和输入输出对(Q_c,A_c)，以预测反思词元z：$p(z|I,Q_c,\overline{A_c})$。本文采用最新GPT-4（具体版本号为GPT-4-0613，后续均用此版本）的预测结果作为监督信号。最终在收集训练数据D_{critic}后，本文使用Qwen2.5-Instruct-7B初始化评估模型M，并在D_{critic}上使用标准条件语言建模目标进行训练，最大化似然函数为：

$$\max_M \mathbb{E}_{((Q_c,A_c),z)\sim D_{\mathrm{critic}}} \ln\left[p_C(z|Q_c,A_c)\right] \tag{3}$$

Prompt 4: 新兴词判别器

指令：你将获得一个术语及其上下文，以及大语言模型对其含义的推理过程。请判断该术语是否属于需要通过网络检索来更好理解的新词。回答[是]或[否]，并提供解释。

术语: [t_i]
上下文: [$c_1,c_2,\cdots,c_n$]
大模型推理过程: [Q_c,A_c]

##样例:
输入:
指令：你知道上下文中提及的术语吗？
术语：赫伊森
上下文:“[赫伊森自己眼睛也不大吧”，“赫伊森也配歧视亚洲人”]
大语言模型对其含义的推理过程：文中提及赫伊森可能是一个人名，并歧视亚洲人。
输出:
需要检索？：[是]
解释：结果非常模糊，仅提供了一个粗略的猜测，没有具体说明赫伊森是谁。因此，需要检索更多信息才能准确理解该术语。

图5 新兴词判别器提示词

然而，在特定领域术语集合上训练的评估模型难以应对不断涌现的新兴词带来的语义演化挑战，因此本文引入了在线强化学习相关技术。在线强化学习技术可以使模型具备持续学习能力，即当模型对新兴词执行语义定位时，将根据检索增强带来的LLM语义置信度变化情况实时调整评估模型。

在强化学习的背景下，评估模型优化被形式化为马尔可夫决策过程5元组$<S,A,P,R,\gamma>$。其中，状态空间S代表由词汇表与序列长度限制的有限集合；动作空间A等同于词汇表，即模型生成的反射标记（“Yes”或“No”）；转移概率P是由判别模型C_θ（策略网络）决定的；奖励函数R用于计算检索增强前后大语言模型的语义置信度差异，反映语义理解的提升程度；折扣因子γ用于控制未来奖励的权重，平衡即时奖励与长期奖励。

具体地，经过预热后的评估模型 C_θ 是初始策略模型 π_0。在每一步 t，动作 a_t 是基于对当前状态的观察生成下一个词元 x^t，即 $s_t=[x,\hat{x}_{<t}]$，当生成被序列结束词元停止时结束该流程。在检索增强后，模型通过评估检索增强带来的 LLM 语义置信度来计算奖励。具体而言，奖励值通过计算检索增强前后语义响应置信度水平的差异来确定。

本文采用近端策略优化（proximal policy optimization, PPO）[22]，遵循 Ramamurthy 等[23]的研究，奖励函数 R 期望的最大化形式为：

$$\max_\theta \mathbb{E}_{(\hat{x}\sim p_\theta(\cdot|x)}[R(x,\hat{x})] \tag{4}$$

优化问题可以进一步表示为：

$$\begin{aligned}&\max_\theta \mathbb{E}_{(s_t,a_t)\sim\pi_\theta}[\min\{r_{t,\theta}F^{\theta_{\text{old}}}(s_t,a_t);\\&\text{clip}(r_{t,\theta},1-\varepsilon,1+\varepsilon)F^{\theta_{\text{old}}}(s_t,a_t)\}]\end{aligned} \tag{5}$$

其中，$r_{t,\theta}$ 定义为：

$$r_{t,\theta}=\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \tag{6}$$

其中，θ_{old} 表示用于采样的临时固定策略参数，θ 被更新。F 表示优势函数，基于价值网络 V_ϕ 的估计形成。价值网络从策略网络 π_0 初始化。公式遵循广义优势估计[24]（generalized advantage estimation, GAE）。优势函数表示为：

$$\begin{aligned}&\delta_t=R(s_t,a_t)+\gamma V_\phi(s_{t+1})-V_\phi(s_t)\\&F_t^\theta(s_t,a_t)=\sum_{l=0}^{\infty}\lambda^l\delta_{t+l}\end{aligned} \tag{7}$$

其中，λ 表示偏差–权衡参数，γ 表示折扣因子。

奖励函数 R 反映了检索增强带来的大语言模型术语理解的语义置信度增益。受大语言模型中置信度评估[10]的启发，本文使用 LLM 输出的 token 置信度来表示术语理解的准确性。关于置信度变化与语义增益之间的相关性分析，详见第 4.6 节。因此，假设术语 t_i 检索增强带来的语义置信度增益为 Δk_i，则 $\Delta k_i=k_i^r-k_i^c$。其中，k_i^c 表示检索增强前推断术语 t_i 含义的语义置信度，即图 1 中跨上下文一致性推理过程中输出的置信度；k_i^r 表示检索增强后的术语语义置信度。此外，添加 KL（Kullback-Leibler）散度正则化[25]以防止模型过度偏离初始化状态。

$$R(s_t,a_t)=\Delta k-\beta\text{KL}(\pi_\theta\|\pi_0) \tag{8}$$

最终损失函数由策略损失 $\mathcal{L}_\theta$ 和价值损失 $\mathcal{L}_\phi$ 组成。

$$\begin{aligned}&\mathcal{L}_\theta=-\frac{1}{|\mathcal{S}|T}\sum_{\tau\in\mathcal{S}}\sum_{t=0}^{T}\min(r_{t,\theta}F_t^{\theta_{\text{old}}},\text{clip}\,F_t^{\theta_{\text{old}}})\\&\mathcal{L}_\phi=\frac{1}{|\mathcal{S}|T}\sum_{\tau\in\mathcal{S}}\sum_{t=0}^{T}(V_\phi(s_t)-R_t)^2\\&\mathcal{L}_{\text{PPO}}=\mathcal{L}_\theta+\lambda_v\mathcal{L}_\phi\end{aligned} \tag{9}$$

其中，S 表示遇到的新兴词集合，T 表示步数。

3.4　种子引导主题分配

在利用多轮历史对话对术语语义信息进行精确理解后，进一步通过提示词 Prompt 5 引导大语言模型判断该术语最可能归属的种子主题，并生成与该主题相关的解释说明，以支持后续对各个种子主题下术语相关性的筛选与排序，从而选取最相关的前 K 个术语。术语的主题分配提示词如图 6 所示。

Prompt 5: 种子引导主题分配

##指令：根据历史对话，请判断上述提到的术语是否属于以下某个主题。如果是，输出该术语对应主题的解释；如果不是，输出0。

##术语: [t_i]
##主题: [s_1,s_2,…,s_n]

##样例:
输入
指令：根据历史对话，请判断上述提到的“赫伊森”是否与以下某个主题相关。如果是，输出该术语对应主题的解释；如果不是，输出0。
主题：
[“曼彻斯特联足球俱乐部”，“皇家马德里足球俱乐部”,…]

输出：
主题：“皇家马德里足球俱乐部”
解释：赫伊森指的是迪恩·赫伊森，皇家马德里的中后卫。

图 6　主题分配提示词

完成上述分配后，即可获得每个种子主题 s_i 对应的术语集合 T_i。

$$\begin{aligned}&x_i=\begin{cases}m_i^c, & z=\text{“No”}\\ m_i^r, & z=\text{“Yes”}\end{cases}\\&(e_i,\text{topic}_i)=\text{LLM}(x_i,\text{Prompt 5})\end{aligned} \tag{10}$$

其中，topic_i 表示术语 t_i 的种子主题，e_i 表示术语 t_i 在该主题下的语义解释。

随后，本文对集合 T_i 中各术语的解释文本进行 Sentence-BERT[26]编码，得到语义向量表示，并与种子主题 s_i 的向量表示计算余弦相似度，最终选取得分最高的前 K 个术语。

$$\begin{aligned}&h_j=\text{SBERT}(e_j),\quad t_j\in T_i,\\&\boldsymbol{H}_{T_i}=\{h_j|t_j\in T_i\}\end{aligned} \tag{11}$$

$$\text{Top}K(T_i)=\underset{t_j\in T_i}{\text{argtop}K}[\cos(h_j,\text{SBERT}(s_i))] \tag{12}$$

其中，$\cos(\cdot,\cdot)$表示余弦相似度函数，$\text{argtop}K[\cdot]$表示选取相似度最高的K个术语。

4 实验与分析

4.1 实验设置

4.1.1 数据集

为全面评估SeedLLM在不同领域与术语类型上的性能，本文采用3类共4个多领域数据集：目标主题挖掘经典纽约时报新闻数据集（NYT）[17]与餐厅评论数据集（Yelp）[17]、专业领域神农中医药数据集（TCM）[27]，以及富含新兴词的哔哩哔哩足球社区评论数据集，覆盖中英文双语与多样化应用场景。

（1）经典英文数据集。为便于与现有目标主题挖掘方法（如CatE、SeedTopicMine）进行横向比较，本文采用其常用的基准数据集进行实验评估。具体选取两个经典英文数据集：NYT和Yelp。其中，NYT数据集包含31 997篇新闻报道，种子集涵盖10个国家类别；Yelp数据集包含29 280条餐厅评论，种子集涵盖8种食品类别，属于典型的短文本场景，可用于验证模型在社交媒体类短文本主题发现任务中的性能。

（2）专业领域数据集。本文选取TCM作为专业领域的代表，中医领域是一个包含大量专业术语和小众词汇的专业领域，其中的专业化词语显然属于黑话小众词的定义，这一特性使中医领域成为验证模型跨上下文推理黑话术语能力的理想场景。此数据集源于ShenNong-TCM-Dataset/EB神农中医药[27]数据集，包含医生–患者对话数据，以及中草药、方剂以及疾病名称等，共有32 252篇中医相关文档，根据种子集的不同分为两个子任务：药类子任务包含6种中药类别检测，简称TCM-M；病类子任务包含5种疾病类别，简称TCM-D。

（3）新兴词场景数据集。本文选取中文足球社区评论数据作为新兴词场景的代表。足球球迷社区是网络流行语和新兴术语快速涌现的典型环境，大量由球迷自发创造的昵称与隐喻表达具有明显的时效性、语境依赖性和非规范性，因此适合作为检验模型适应新兴词带来语义演变能力的实验场景。该数据集源于哔哩哔哩平台，通过检索相关种子主题词爬取2025年1月至2026年3月的视频评论区及二级评论，并进行数据清洗与语句长度限制（50~150字）等预处理。最终按月份划分，每月约1 000条，共构建包含15 000条评论的数据集。本文以7个足球俱乐部作为种子词。该数据集包含大量新兴术语与社区黑话，能够有效评估模型在动态语义环境下的主题挖掘能力，简称Soccer数据集。

以上所有数据集均采用6:4的训练和测试划分，详细统计信息如表1所示。

4.1.2 对比方法

本文将所有对比方法分为以下4类，以全面覆盖目标主题挖掘领域的代表性方法，使实验结果更具说服力。

（1）关键词检索模型。KeyBERT[28]是一种基于BERT嵌入的关键词检索方法。该方法首先使用BERT对文档进行编码，然后计算每个候选词与文档的语义相似度，选择语义相似度最高的词作为关键词。虽然KeyBERT并非专门为目标主题挖掘设计，但其基于语义相似度的关键词选取机制与目标主题挖掘任务有一定关联，因此作为基线进行比较。即将种子词视为一个文档，计算不同词与种子之间的语义相似度。

（2）传统监督概率主题模型。SeedLDA[15]是一种监督型概率主题建模方法。该方法通过引导每个主题倾向于生成更多种子词，同时引导每个文档优先选择与文中出现的种子相关的主题，从而通过对LDA的生成过程进行改进来挖掘用户感兴趣的主题。FoTo[9]是一种面向短文本的目标主题建模方法。该方法面向用户感兴趣的关键词集合实现聚焦主题

表1 4个领域数据集的详细统计信息

数据集	种子主题/个	文档数量/篇	种子集合
NYT	10	31 997	China, United States, Iraq, United Kingdom, Japan, Canada, France, Italy, Russia, Germany
Yelp	8	29 280	steak, seafood, pizza, dessert, salad, noodles, sushi, burger
中医药类TCM-M	6	32 252	解表药、清热药、温里药、祛湿药、理血药、补益药
中医病类TCM-D	5	32 252	外感病、内伤病、痰饮病、瘀血病、虚损病
Soccer	7	15 000	曼联、曼城、利物浦、阿森纳、切尔西、皇马、巴萨

分析，针对短文本数据稀疏、传统模型全局均等建模导致目标主题不精准的问题，基于文档–关键词距离与文档对排序关系来实现主题概率模型学习与文档低维语义嵌入联合优化，强化目标主题的挖掘，在学习全局主题的同时精准捕获用户关注的相关话题。

（3）种子引导主题发现模型。CatE[17]是一种种子引导的嵌入学习方法。该方法首先基于种子词学习类别感知的词嵌入，然后通过最大化术语与目标类别的相似度同时最小化与其他类别的相似度来优化嵌入空间。SeedTopicMine[14]在 CatE 的基础上进行了改进，整合了多种术语嵌入（包括静态词嵌入和预训练嵌入），通过多视角嵌入融合来提高术语表示的丰富性。该方法在多个数据集上超越了 CatE 方法，成为种子引导主题发现领域的代表性方法。

（4）基于大语言模型的主题挖掘。DiskTM-LK[5]利用大语言模型提取术语并生成目标主题的种子词，将其注入词汇的实值多项分布构建先验知识矩阵，用于刻画词语与主题的语义相似性。在模型学习过程中，对主题–词分布施加主题层约束，使提取的主题与预定义关键词对齐，从而挖掘用户感兴趣的主题。

4.1.3　实现细节

在术语提取部分，对于所有基线方法，本文首先从语料库中提取候选术语。具体步骤如下：使用 AutoPhrase 算法从 NYT 和 Yelp 数据集中提取高质量短语。对于 TCM 和足球社区数据集 Soccer，本文采用基于大语言模型的术语提取方法。

在新兴词评估模型预热部分，本文从 NYT 数据集的测试集中提取了 5 000 个词，初始预热模型为 Qwen2.5-Instruct-7B。并通过使用 GPT-4 对术语推理过程的标注结果来预热新兴词模型。然后，该模型直接应用于另外 3 个数据集。其中模型温度参数设为 1，最大输出 token 数设为 200。对于 GPT-4 未遵循指定输出格式或序列（即与预期类别名称不匹配）的实例，本文将其丢弃。最终，本文收集了 4 966 个实例。预热后的 Critic 模型直接应用于其他 3 个数据集，不需要额外的领域特定微调。本文使用标准的条件语言建模目标进行训练，最大化正确反射 token（“Yes”或“No”）的似然。训练参数如下：学习率为 2×10^{-5}；批次大小为 8；训练轮数为 3；最大序列长度为 512；优化器为 AdamW；权重衰减为 0.01。

在检索增强部分，检索结果数量为前 10 条。检索时间范围为过去一年（对于足球社区数据集 Soccer，限制为过去 3 个月以保证时效性）。

4.1.4　评估指标

本文分别采用 3 种不同维度的指标对实验结果进行评估，具体实现如下。

（1）主题连贯性指标 IDT。该指标主要衡量模型在同一主题下挖掘出的术语在语义层面是否具有较高的相关性和一致性。由于本文的研究重点是挖掘特定主题中的小众词，因此像 NMPI[29]这类基于词频共现的传统主题一致性度量方法并不适用于包含大量小众词的目标主题挖掘任务。为此，本文参考了 Stammbach[30]等利用大语言模型评估主题一致性的相关方法，通过设计主题 IDT 来评估挖掘主题词之间的语义连贯性。具体而言，本文从某一主题对应的 10 个最相关术语中随机选取 5 个术语，并加入 1 个来自其他种子主题的术语作为入侵词，共同构成待检测列表。若大语言模型能够更容易地识别出入侵词，则说明该主题的语义连贯性越好。其中大语言模型用到的 Prompt 6 如图 7 所示。已知种子集合为 seeds，则 IDT 指标的具体算法为：

$$f(w_k)=\mathrm{LLM}(L_i,\mathrm{Prompt}_{\mathrm{IDT}})$$

$$\mathrm{IDT}=\frac{1}{|\mathrm{seeds}|}\sum_{i=1}^{|\mathrm{seeds}|}\frac{1}{|\mathcal{W}_i|}\sum_{k=1}^{|\mathcal{W}_i|}\mathbb{I}(w_k=f(w_k)) \quad (13)$$

其中，$\mathbb{I}(w_k=f(w_k))$ 为指示函数，表示大语言模型正确识别出入侵词 w_k。

Prompt 6: 主题词入侵检测任务

指令：**给定一组属于同一主题的相关术语及其对应解释。请对这些术语进行语义分析，判断其中哪一个是“入侵词”，即与该主题语义关联度最低、与其他术语语义一致性最弱的词语，并给出理由。**

术语：[$<t_1$: $e_1>$, $<t_2$: $e_2>$, …,$<t_6$: $e_6>$]

样例:

输入
指令：请从下面术语中选择一个与其他术语最不相关的词语并给出理由。

术语:
[“皇马”：皇家马德里俱乐部,
“赫伊森”：皇家马德里中卫，…
“亚马尔”：“巴塞罗那前锋”]

入侵词：亚马尔

理由：其他术语都是与皇马相关的术语，而亚马尔是巴萨相关术语。

图 7　主题连贯性指标计算提示词

（2）主题内部多样性指标IntraTD。该指标用于衡量同一主题内部术语的多样性。传统主题多样性指标（如Diversity[31]）通常通过统计不同主题之间的独特词汇进行评估，但在种子引导的主题发现任务中，每个术语仅属于一个主题，不会在不同主题间重复。所以CatE和SeedTopicMine在实验结果中该指标通常为1，故传统指标并不适用于该任务场景。

为了解决上述问题，本文提出了IntraTD指标，使用GPT-4基于每个模型挖掘的Top-10主题术语构建主题分类树（具体的Prompt如图8所示），然后计算构建的树中一级节点的数量（去除无关节点分支），以表示主题结构的语义多样性，因为它直接反映了种子主题类别的广度。

Prompt 7：主题分类树构建

指令：针对给定的主题词s_i，构建一个“种子主题词s_i”术语主题树，该主题树必须满足以下要求：

1. 设定种子主题词为唯一根节点，统摄全部相关词汇与概念
2. 按照语义层级关联对节点进行分类：上层节点为大类，下层节点为子类/实例。关系类型包括 但不限于：整体-部分、类别-实例、属性-特征、派生-关联。
3. 无关术语直接归入一级分类“无关节点”之下。

术语：[$<t_1: e_1>, <t_2: e_2>, \cdots, <t_6: e_6>$]

种子主题：[s_i]

样例：

指令：根据以下种子主题词和 K 个术语，构建“皇马主题树”：
主题术语：[“皇马”，“赫伊森”，“伯纳乌”，“hala啃”，…]
种子主题：[皇家马德里]
输出：BEGIN_TREE
　　[0：皇家马德里-相关概念（根节点）
　　1：人员（父节点ID=0）
　　2：赫伊森（父节点ID=1）
　　3：球迷文化（父节点ID=0）
　　4：hala啃（父节点ID=3）....
　　K：无关节点（父节点ID=0）
　　K+1：亚马尔（父节点ID=K）]
END_TREE

图8　IntraTD指标提示词

最终结果通过对所有种子的一级节点数量取平均值获得。

$$\text{IntraTD}=\frac{1}{|\text{seeds}|}\sum_{i=1}^{|\text{seeds}|}N_{\text{first-level},s_i} \tag{14}$$

其中，$N_{\text{first-level,seed}_i}$表示种子主题$\text{seed}_i$对应主题树中一级节点的数量。

（3）主题准确性指标MACC。该指标用于评估挖掘得到的术语与对应种子主题之间的语义匹配程度，通过衡量术语与种子主题的相关性，反映模型在术语分配过程中的准确性。

$$\text{MACC@}K=\frac{1}{|\text{seeds}|}\sum_{i=1}^{|\text{seeds}|}\frac{1}{K}\sum_{j=1}^{K}\mathbb{I}(t_{i,j}\in\text{seed}_i) \tag{15}$$

其中，$t_{i,j}$表示主题T_i中第j个最相关的术语，$\mathbb{I}(t_{i,j}\in\text{seed}_i)$表示指示函数，用于判断$t_{i,j}$是否属于$\text{seed}_i$，$K$设置为20，即评估挖掘出的前20个术语。

由于该评估依赖人工判断，本文聘请了5名标注员进行独立评估。标注员均为相关领域的研究生或具有相关背景的专业人士。标注流程如下：①向标注员展示种子词和候选术语；②标注员判断该术语是否与种子词语义相关（是/否）；③对于边界案例，标注员可提供“不确定”判断，该判断不计入最终统计。报告的MACC@K分数反映了5名标注员的平均结果。标注员间一致性通过Fleiss’ kappa进行验证，NYT、Yelp、TCM-M、TCM-D和Soccer的Fleiss’ kappa分数分别为0.894、0.870、0.802、0.765和0.711，表明标注结果具有高度一致性。

4.2　实验结果

本节主要分析SeedLLM与现有模型的性能。各个模型在4个数据集上的IDT、IntraTD和MACC分数的对比结果如表2所示（其中下划线为次优性能，加粗为最优性能）。首先，本文分析两个经典数据集上模型性能的表现，Keybert作为非种子引导的关键词检索方法，在所有数据集上均表现最差，IDT和MACC分数均低于0.3。这表明简单的语义相似度计算难以有效处理种子引导的主题发现任务。CatE和SeedTopicMine作为专门的种子引导方法，性能显著优于Keybert。除此之外，简单使用LLM提取术语而不去探究其语义的DiskTM-LK方法并不能完全优于传统基于嵌入的方法。相反，本文提出的SeedLLM通过挖掘词汇的深层含义，在经典数据集上均获得最佳性能。SeedLLM在NYT数据集上达到0.893的MACC分数，比SeedTopicMine提升约5.7%。其次，本文观察中医数据集和足球社区数据集的表现，在这类充满大量黑话术语的场景下，采用大语言模型提取术语的方法（如DiskTM-LK、SeedLLM）明显优于传统基于嵌入的方法。在此类场景下，SeedLLM的优势极大地展现出来了，在MACC指标上相比次优模型分别在中医数据集和足球社区数据集上提升了约19%和16%。综上所述，SeedLLM模型在3个指标上均达到了当前最优（state-of-the-art, SOTA）性能。除

表 2　各个模型在 4 个数据集上的性能对比

模型	NYT			Yelp			TCM						Soccer		
	国家类			食物类			中药类（TCM-M）			疾病类（TCM-D）			俱乐部类		
	IDT	IntraTD	MACC	IDT	IntraTD	MACC	IDT	IntraTD	MACC	IDT	IntraTD	MACC	IDT	IntraTD	MACC
KeyBERT	0.236	1.633	0.207	0.221	1.451	0.281	0.226	1.753	0.201	0.211	1.513	0.192	0.234	1.217	0.198
SeedLDA	0.325	1.987	0.312	0.331	1.892	0.365	0.308	2.015	0.297	0.289	1.956	0.278	0.317	1.762	0.273
FoTo	0.412	2.235	0.428	0.425	2.187	0.493	0.402	2.618	0.526	0.514	2.544	0.379	0.462	2.415	0.491
CatE	0.553	2.511	0.748	0.548	2.418	0.681	0.361	2.114	0.471	0.435	2.279	0.434	0.425	2.156	0.410
SeedTopicMine	0.639	2.696	0.836	0.657	2.752	0.791	0.417	2.452	0.497	0.317	2.452	0.347	0.414	2.389	0.447
DiskTM-LK	0.669	3.119	0.782	0.612	3.258	0.717	0.535	2.734	0.542	0.508	2.664	0.442	0.511	2.930	0.518
SeedLLM	**0.735**	**3.467**	**0.893**	**0.742**	**3.729**	**0.856**	**0.642**	**3.279**	**0.703**	**0.627**	**3.132**	**0.663**	**0.655**	**3.415**	**0.684**

了指标评估外，本文还设计了一个弱监督分类任务来评估主题词的区分性。具体做法是利用挖掘得到的主题词作为弱监督信号对文本进行标注，并微调 BERT 模型，通过最终的分类性能来衡量主题词的区分能力。若分类效果越好，说明挖掘得到的主题词越具有区分能力。具体而言，对于每个模型挖掘得到的多个目标主题，本文选取每个主题排名前 20 的主题词作为弱监督信号，用于构建多类别文本分类任务的数据集。首先，从原始语料中随机采样 2 000 条文本，并根据文本中是否包含各主题关键词集合中的词汇进行自动标注：若文本包含某一主题关键词集合中的词汇，则标记为该主题样本；若同时匹配多个主题，则分配给关键词匹配数量最多的主题；若未匹配任何主题关键词，则标记为负样本，从而构建弱监督训练集。随后，利用该训练集对 BERT 模型进行微调以学习不同主题间的语义差异。测试阶段从语料中随机抽取 500 条文本并由人工标注真实主题构建测试集，通过计算 Macro-F1 评估模型主题词对不同目标主题文本的区分能力。Macro-F1 越高，说明主题关键词的代表性与判别能力越强。最终实验结果如表 3 所示。

表 3　基于主题关键词的弱监督分类结果（Macro-F1）

模型	NYT	Yelp	TCM-M	TCM-D	Soccer
KeyBERT	0.186	0.173	0.179	0.168	0.182
SeedLDA	0.223	0.238	0.214	0.229	0.241
FoTo	0.368	0.392	0.351	0.307	0.384
CatE	0.427	0.439	0.468	0.301	0.386
SeedTopicMine	**0.611**	0.532	0.412	0.344	0.331
DiskTM-LK	0.489	0.472	0.453	0.419	0.445
SeedLLM	0.602	**0.619**	**0.586**	**0.573**	**0.558**

表 3 结果表明，SeedLLM 挖掘的词汇在弱监督分类任务中的大多数数据集上均取得了最优性能。这一结果验证了其挖掘出的术语具有较强的主题区分能力，能够有效提升在下游弱监督分类任务中的表现，体现出良好的应用潜力与实际价值。

4.3　消融实验

为了验证模型各个部分的有效性，本文设计以下变体在 TCM-M、TCM-D 和 Soccer 数据集上进行消融实验。

（1）W/O Context&RAG。不考虑术语所在的多个上下文，直接让大语言模型从单个文本中提取术语及其解释，也不进行域外检索，该变体模拟了直接使用大语言模型进行零样本主题发现的情况，用于验证上下文信息和检索增强的必要性。

（2）W/O RAG。黑话术语跨上下文语义一致性推理模块，但消除检索增强模块。所有术语均仅基于上下文进行语义理解，不进行额外的网络检索。该变体用于验证检索增强对理解新兴术语的必要性。

（3）W/O Critic Model。消除新兴词评估模型，对所有术语统一应用检索增强策略。该变体用于验证新兴词评估模型在识别新兴术语方面的有效性，以及不加选择地进行检索可能带来的负面影响。

消融实验结果如图 9~图 11 所示。完整模型 SeedLLM 在所有指标上均显著优于各消融版本，验证了模型设计的有效性。

实验结果表明，W/O Context&RAG 变体在所有数据集上的表现最差，其 IDT 得分为 0.382~0.497，MACC 得分为 0.356~0.533。相比之下，W/O RAG 变体通过引入上下文信息后性能显著提升，MACC 指标提升 2%~50%。这一结果表明，跨上下

文语义一致性推理对于准确理解黑话术语语义具有关键作用。若缺乏多上下文约束而直接进行术语提取，大语言模型容易产生幻觉或形成片面理解，从而错误包含无关词语或遗漏稀有术语。W/O Critic Model变体的较好表现进一步验证了检索增强在理解新兴术语方面的重要性。例如，在足球社区数据集中，“滕甲军”等新兴术语往往依赖最新的网络语境才能被准确理解，仅依赖上下文信息难以捕获其最新语义。然而，如图9所示，该变体不加区分地检索域外信息，导致主题内部多样性指标IntraTD出现下降。这是因为无差别引入的外部语义在一定程度上替代了原本具有丰富语境特征的域内黑话含义。例如，“蓝月亮”在足球社区数据集中通常指代曼城，但外部检索结果可能将其解释为洗衣液品牌，从而削弱术语的社区语义关联并降低语义多样性。

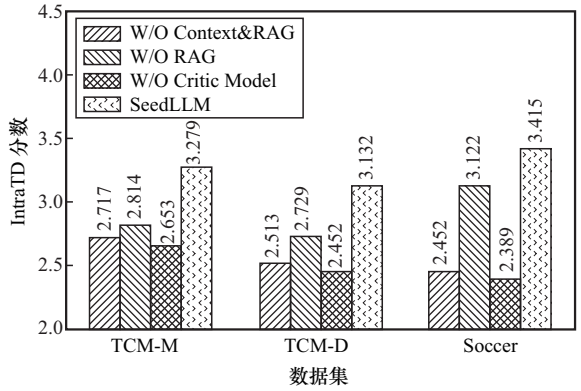

图9　不同变体在多个数据集上的IntraTD性能对比

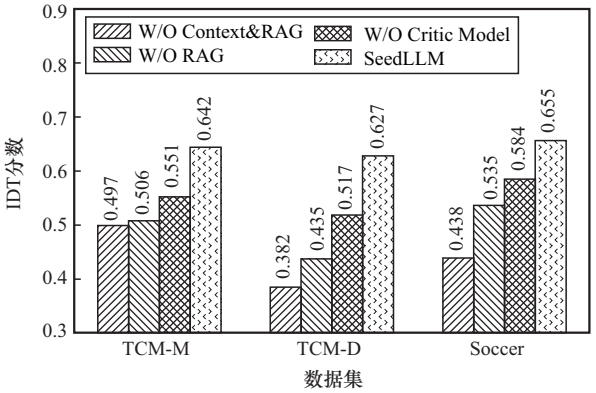

图10　不同变体在多个数据集上的IDT性能对比

此外，可以观察到W/O Critic Model变体的整体性能仍低于完整的SeedLLM模型，表明无选择的检索机制容易引入噪声。例如，对于一些小众术语（如中医领域的经典药材名称），检索结果可能包含大量无关信息，从而干扰模型对术语本义的理解；同时，检索过程本身也会带来额外的计算开销与推理时延。相比之下，完整的SeedLLM模型通过Critic模型动态识别真正需要外部检索的新兴术语，在保持性能的同时提升了整体效率。实验结果显示，相较于W/O Critic Model，SeedLLM在MACC指标上进一步提升5%~7%。

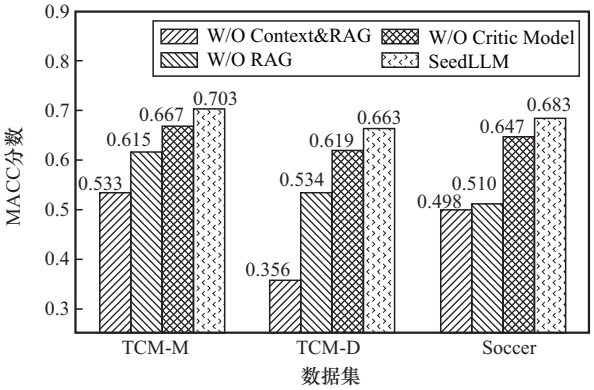

图11　不同变体在多个数据集上的MACC性能对比

4.4　参数分析

4.2节验证了识别新兴词从而触发检索的重要性，接下来本节将验证在线强化学习调整新兴词评估模型的策略是否能适应新兴词涌入带来的动态语义演变问题。本文通过模拟新兴词不断涌入的场景，截取Soccer数据集2025年6月—2026年3月这10个月的数据，并分为10个批次，每个批次1 000条数据逐次输入，模拟随着时间的推移不断有新文本加入的场景。本文对比的4个变体分别如下。

（1）Other-Warm CM。使用NYT数据集进行热身的新兴词评估模型，后续不进行在线强化学习微调。

（2）Same-Warm CM。使用Soccer数据集2025年1—5月的数据进行热身的新兴词评估模型，后续不进行在线强化学习微调。

（3）GPT-4 CM。直接采用GPT-4作为新兴词评估模型。

（4）Same-Warm RL。使用Soccer数据集2025年1—5月的数据进行热身的新兴词评估模型，且后续采用在线强化学习微调。

本文重点评估MACC指标性能的变化，最终对比结果如图12所示。

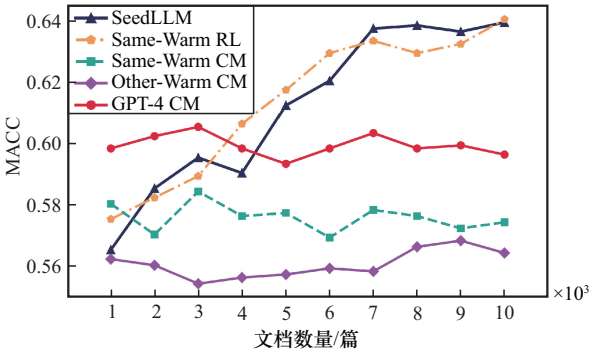

图 12　不同新兴词评估模型面对语义演变 MACC 性能对比

从图 12 可以得出以下结论。

（1）同领域热身的优势与固定模型的局限性。Same-Warm CM 采用同领域数据热身，性能显著优于跨领域热身的 Other-Warm CM，说明领域匹配对评估模型初始性能具有重要影响。GPT-4 CM 凭借其更丰富的知识储量和更大的参数规模，整体性能优于上述两类小模型。然而，这类参数固定的模型在新兴词持续涌现、语义不断演变的场景下，尽管初期性能表现优于 SeedLLM，但随着时间推移，其性能提升愈发有限，难以实现动态适应。

（2）SeedLLM 的持续自适应能力。相比之下，SeedLLM 在强化学习机制的驱动下，能够依据检索增强带来的效果反馈持续更新评估标准，从而动态适应语义演变。其性能随批次递增而稳步提升，并最终超越包括 GPT-4 CM 在内的所有固定参数模型。这表明以 GPT-4 CM 为代表的静态判别模型难以随新兴词的引入进行动态调整，在持续演变的语义环境中存在明显的适应性瓶颈。SeedLLM 则展现出在网络舆情场景中应对新兴语料持续涌入的良好动态适应能力与实际应用价值。

（3）热身数据分布对强化学习收敛的影响。值得注意的是，采用同领域数据热身并结合在线强化学习的 Same-Warm RL，其性能轨迹随训练推进逐渐向 SeedLLM 靠拢。这一现象侧面表明，无论初始热身阶段采用何种数据分布，经过在线强化学习机制的持续调整，模型最终均能收敛并适应新兴数据的语义分布，体现了强化学习机制在跨分布泛化方面的鲁棒性。

4.5　样例分析

为了定性考察 SeedLLM 各功能单元对目标主题挖掘的边际贡献，本文以“皇马”为固定主题种子开展 SeedLLM 模型四阶段对比样例分析实验。第一阶段直接在完成术语提取后进行主题分配；第二阶段在第一阶段基础上增加黑话术语跨上下文语义一致性推理模块后再进行主题分配；第三阶段是在第二阶段基础上增加无差别检索后，进行主题分配的结果；第四阶段是完整流程结果。对比结果如图 13 所示。从第一阶段可以看出，直接使用大语言模型提取术语并解释存在显著幻觉，其将“哈兰德”（曼城球员）、“萨卡”（阿森纳球员）等错误分配至皇马主题，主题相关率仅 50%。引入黑话术语跨上下文语义一致性推理模块后，模型通过综合术语所在的多个上下文信息，获取其在领域内的深层次语义，从而将主题相关率提升至 70%，但“调教”等抽象表述仍因缺乏显式指征而残留。引入检索增强模块后，上下文语义不足的术语信息得以补全，成功召回大量新兴实体名词（如球员名等），但其过度检索导致挖掘到的术语偏向于具体实体，甚至引入噪声，将原始黑话术语信息覆盖，从而其挖掘的主题内部多样性下降。相比之下，SeedLLM 采用新兴词评估模型的选择性检索，且能实时调整评估标准的策略，挖掘了大量黑话术语和新兴词。从渐进式结果来看，本文方法具有显著的幻觉抑制、歧义消除与挖掘新兴术语的能力，因此它有效缓解了传统 LLM 在开放域主题挖掘中的语义漂移与知识陈旧问题。

4.6　置信度提升与语义增益的一致性分析

为了探究检索增强前后大语言模型置信度的提升是否能够反映真实的语义增益，本文对置信度增益与语义增益之间的一致性进行了系统分析。首先，在 Soccer 数据集中，收集 SeedLLM 模型中触发 RAG 后置信度提升的术语共计 1 720 个（即被模型判定为新词的样本）。在此基础上，进一步调用高性能大语言模型（GPT-4 与 Claude）对上述新词进行语义理解能力评估，筛选出模型能够准确理解其含义的新词子集，并依据预设提示词（Prompt 3）为每个新词自动生成对应释义。随后经由人工审核对生成结果进行质量把控，最终保留模型输出置信度（Softmax 概率）高于 0.9 的条目，共获得 200 个高质量新词语义释义标注，作为后续真实语义增益计算的参考基准。对于每个新词，两个大语言模型分别生成 3 次释义，共得到 6 个候选语义表示。随后，采用 Sentence-BERT 对其进行编码，并取平均值得到参考语义表示 h_g，与此同时，本文将

皇马:（**第一阶段**）
1. “皇马”：皇家马德里足球俱乐部是西班牙足球甲级联赛的一支豪门球队，以其辉煌历史成绩著称。
2. “弗洛伦蒂诺”：弗洛伦蒂诺·佩雷斯是皇家马德里俱乐部主席，推动了“银河战舰”时代的建设。
3. “银河战舰”：银河战舰是球迷对皇家马德里的昵称，指球队汇聚众多世界级球星。
4. “伯纳乌”：伯纳乌是皇家马德里的主场球场，常被用来象征皇马的主场氛围和历史。
5. “哈兰德”：哈兰德是皇家马德里的潜在前锋之一，以其强大的身体和进球能力著称。【× 曼城球员】
6. “佩德里”：佩德里是皇家马德里的年轻中场核心，被认为是球队未来的组织者。【× 巴塞罗那球员】
7. “萨卡”：萨卡是皇家马德里的边锋球员，以速度和突破能力闻名。【× 阿森纳球员】
8. “德布劳内”：凯文·德布劳内是皇家马德里的中场核心，负责球队进攻组织。【× 曼城球员】
9. “安帅”：安切洛蒂是皇家马德里主教练，带领球队赢得多项冠军。
10. “乔治娜”：乔治娜是皇家马德里球员本泽马的妻子，经常出现在伯纳乌看台。【× 皇马C罗的妻子】

皇马:（**第二阶段**）
1. “皇马”：皇家马德里足球俱乐部是西班牙足球甲级联赛的一支豪门球队，以其辉煌历史成绩著称。
2. “弗洛伦蒂诺”：弗洛伦蒂诺·佩雷斯是皇家马德里俱乐部主席，常被球迷称为“老佛爷”。
3. “银河战舰”：银河战舰是对皇家马德里的经典称呼，形容球队拥有大量世界级球星。
4. “伯纳乌”：伯纳乌是皇家马德里的主场球场，常被用来指皇马的主场氛围。
5. “国家德比”：国家德比通常指西班牙两大豪门皇家马德里和巴塞罗那之间的比赛。【× 非区分性词汇】
6. “魔笛”：魔笛是卢卡·莫德里奇的球迷昵称，因其组织能力和技术细腻而闻名。
7. “水爷”：水爷是对皇马传奇后卫拉莫斯的昵称，源于其名字谐音和场上领袖气质。
8. “老佛爷”：老佛爷是球迷对皇家马德里俱乐部主席弗洛伦蒂诺的昵称。
9. “银河”：银河常被用来指皇马的“银河战舰”阵容。【× 非区分性词汇】
10. “某俱乐部”：某俱乐部是球迷讨论足球时使用的一种模糊指代。【× 无关词汇】

皇马:（**第三阶段**）
1. “皇马”：皇家马德里足球俱乐部是西班牙足球甲级联赛的一支豪门球队，以其辉煌历史成绩著称。
2. “弗洛伦蒂诺”：弗洛伦蒂诺·佩雷斯是皇家马德里俱乐部主席，长期主导俱乐部运营。
3. “伯纳乌”：伯纳乌是皇家马德里的主场球场，这里常被用来比喻皇马的主场优势。
4. “库尔图瓦”：库尔图瓦是比利时门将，皇家马德里的主力门将。
5. “赫伊森”：赫伊森是西班牙足球运动员，皇家马德里的后卫球员，发表歧视亚裔言论。
6. “卡瓦哈尔”：卡瓦哈尔是西班牙足球运动员，皇家马德里的右后卫。
7. “巴尔韦德”：巴尔韦德是乌拉圭足球运动员，皇家马德里的中场球员。
8. “国家德比”：国家德比通常指皇家马德里和巴塞罗那之间的比赛。【× 非区分性词汇】
9. “调教”：调教是指教练通过训练和战术指导提升球队某些球员的表现。【× 无关词汇】
10. “某俱乐部”：某俱乐部是球迷讨论足球时使用的一种模糊指代。【× 无关词汇】

皇马:（**第四阶段**）
1. “皇马”：皇家马德里足球俱乐部是西班牙足球甲级联赛的一支球队，以其辉煌历史和欧冠成绩著称。
2. “弗洛伦蒂诺”：弗洛伦蒂诺·佩雷斯是皇家马德里俱乐部主席，长期主导俱乐部运营。
3. “伯纳乌”：伯纳乌是皇家马德里的主场球场，这里常被用来比喻皇马的主场优势。
4. “库尔图瓦”：库尔图瓦是比利时足球运动员，皇家马德里的主力门将。
5. “赫伊森”：阿拉巴是西班牙足球运动员，皇家马德里的后卫球员，发表歧视言论争议。
6. “牙线口罩”：皇马球员赫伊森辱华的关键词，皇马并未就此道歉。
7. “Hala哨”：Hala哨是球迷对裁判偏向皇马的调侃，源自皇马口号 “Hala Madrid”，在争议判罚后被使用。
8. “渣团”：皇马球迷对自己球队的昵称。
9. “远东镶黄旗”：远东镶黄旗是互联网中极端皇马球迷的戏称，讽刺其在网络讨论中过度维护皇马。
10. “某俱乐部”：某俱乐部是球迷讨论足球时使用的一种模糊指代。【× 无关词汇】

图 13　“皇马”主题下，SeedLLM 在 4 个不同阶段的挖掘效果对比

SeedLLM 在检索之前生成的词语义编码为 h_{before}，检索之后生成的语义编码为 h_{after}。计算真实语义增益 h_{Δ}，具体表达式为：

$$h_{\Delta} = \cos(h_g, h_{\text{after}}) - \cos(h_g, h_{\text{before}}) \tag{16}$$

为了避免编码器带来的噪声，本文选取显著阈值 δ =0.03，并将其分为 3 个区间统计其数量，结果如图 14 所示。

从图 14 可以看出，在检索增强使模型置信度提升的样本中，有 88.5% 的新词语义获得显著增益，8.5% 的样本语义变化不显著，仅有 3.0% 出现语义负向增益（即置信度提升但语义质量下降的情况）。这表明在绝大多数情况下，检索增强带来的置信度提升能够有效反映语义质量的提升。同时，少量不一致现象可归因于模型在引入外部信息时产生的“过度自信”。综上所述，置信度增益与语义增益在整体上表现出较强的一致性，说明置信度提升可以作为衡量检索增强语义收益的一个有效信号。

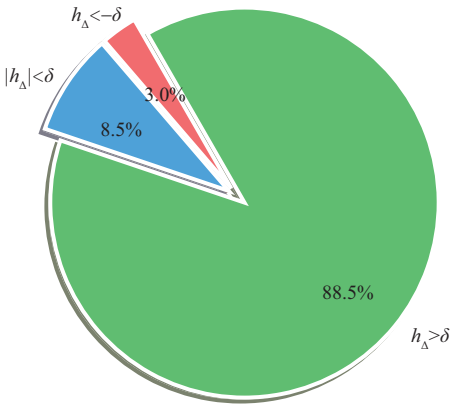

图 14　置信度提升的调节下真实语义增益变化比例分析

4.7　大语言模型指标与人类评估相关分析

由于本文引入多项大语言模型自动评估指标，本节进一步分析大语言模型评价结果与人工标注之间的相关性。参照现有工作[30,32]中大语言模型评估与人类评估一致性分析的方法，本文采用 Spearman[33]相关系数定量衡量两者在排序层面的一致性程度。

（1）实验设置。假设某数据集中共有 N 个种子主题，主实验共使用 M=7 种主题模型（包括本文提出的 SeedLLM）对每个种子分别进行主题挖掘，每个主题由其前 10 个词表示。因此，对于每个种子主题 $c(c=1,2,\cdots,N)$ 和每个模型 $m(m=1,2,\cdots,M)$，可得到一个对应的主题结果。随后，本文分别由人类标注者和 3 种大语言模型（GPT-4、豆包、DeepSeek）对所有主题进行 TD 与 IntraTD 评估。对于人类评估，本文招募 5 位标注者对每个主题进行独立打分，取其平均值作为该主题的人类评分。为保证标注质量，标注者在正式标注前经过统一培训，并在小批量样本上进行校准；标注过程采用盲评方式，标注者不知晓主题来自哪种模型。对于大语言模型评估，为消除大语言模型输出的随机性影响，每个主题在每种大语言模型下以 temperature = 0.7 独立评估 5 次，取平均分作为最终得分。

（2）形式化定义。记第 c 个种子主题下第 m 个模型生成的主题，在人类评估下的平均得分为 $h_{c,m}$，在某一大语言模型评估下的平均得分为 $s_{c,m}$。

评分标准化。由于不同种子主题的绝对评分尺度存在差异（如某些种子本身较难，导致所有模型在该种子下的得分普遍偏低），直接合并不同种子下的评分可能引入种子难度带来的混淆因素。为此，本文对每个种子 c 下的评分分别进行种子内 Z-score 标准化。

$$\tilde{h}_{c,m}=\frac{h_{c,m}-\mu_c^h}{\sigma_c^h},\quad \tilde{s}_{c,m}=\frac{s_{c,m}-\mu_c^s}{\sigma_c^s} \tag{17}$$

其中，$\mu_c^h=\frac{1}{M}\sum_{m=1}^{M}h_{c,m}$，$\sigma_c^h$ 为对应标准差，μ_c^h,σ_c^s 同理。标准化后，每个种子下的评分均值为 0、标准差为 1，消除了种子间绝对分数尺度的差异，使跨种子的评分具有可比性。

相关系数计算。完成标准化后，将该数据集下所有种子的标准化评分合并，形成长度为 $n=NM$ 的评分向量。

$$\begin{aligned}\tilde{\boldsymbol{h}}&=[\tilde{h}_{1,1},\cdots,\tilde{h}_{1,M},\tilde{h}_{2,1},\cdots,\tilde{h}_{N,M}]\in\mathbb{R}^n\\ \tilde{\boldsymbol{s}}&=[\tilde{s}_{1,1},\cdots,\tilde{s}_{1,M},\tilde{s}_{2,1},\cdots,\tilde{s}_{N,M}]\in\mathbb{R}^n\end{aligned} \tag{18}$$

在此基础上，计算 Spearman 相关系数为：

$$\rho=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)} \tag{19}$$

其中，d_i 为第 i 个样本在两组评分向量中的秩差。本文各数据集的种子主题数量不尽相同，NYT 含 10 个种子主题 (n = 70)、Yelp 含 8 个 (n = 56)、Soccer 含 7 个 (n = 49)、TCM-M 含 6 个 (n = 42)、TCM-D 含 5 个 (n = 35)。在上述样本量下，n = 35 时显著性临界值 $\rho_{\text{critical}}\approx 0.334$，$n$ = 7 时 $\rho_{\text{critical}}\approx 0.235$，所有相关系数均大幅超过对应临界值，所有相关系数均通过双尾 t 检验，在 $p<0.05$ 水平下效果显著。TD 的相关性分析结果如表 4 所示，IntraTD 的相关性分析结果如表 5 所示。

表 4　大语言模型 TD 指标与人类评估的 Spearman 相关性分析结果

模型	NYT	Yelp	TCM-M	TCM-D	Soccer
GPT-4	**0.63**	0.57	**0.61**	**0.58**	**0.62**
豆包	0.52	**0.63**	0.59	0.54	0.57
DeepSeek	0.56	0.54	0.55	0.57	0.58

表 5　大语言模型 IntraTD 指标与人类评估的 Spearman 相关性分析结果

模型	NYT	Yelp	TCM-M	TCM-D	Soccer
GPT-4	**0.68**	**0.72**	**0.66**	0.63	**0.62**
豆包	0.62	0.65	0.61	**0.69**	0.61
DeepSeek	0.64	0.69	0.59	0.67	0.54

由表 4 和表 5 可以看出，在 TD 与 IntraTD 指标下，GPT-4、豆包及 DeepSeek 这 3 种大语言模型的评估结果与人类评估之间的 Spearman 相关系数均高于 0.5，依据 Schober 等[34]对相关系数强度的划分标准，整体达到中等及以上相关水平。其中，GPT-4 的整体表现最为突出，在绝大多数数据集上的相关系数均超过 0.6，部分 IntraTD 结果接近或超过 0.7，进一步表明其评估结果与人类判断之间具有较强的一致性与稳定性。综上所述，本文采用的

大语言模型自动评估指标能够有效反映人类判断，可作为人工标注的有效替代。

5 结束语

本文围绕面向用户目标的小众主题词挖掘问题，提出了一种融合大语言模型与检索增强机制的SeedLLM框架。该方法通过引入跨上下文语义一致性约束，对专业领域黑话术语进行全局联合推理，从而缓解单一语境信息不足带来的语义歧义问题；同时，针对主题新兴词语料稀缺、语义证据不足的特点，引入检索增强生成策略，并结合新兴词判别门控与在线强化学习机制，实现对检索触发策略的动态优化，以提升语义定位的准确性与稳定性。在获得术语语义后，进一步利用用户提供的种子词进行主题指派与聚合，从而挖掘与用户关注目标高度对齐的小众主题词集合。多领域实验结果表明，本文方法在小众主题词识别与语义刻画方面优于多种基线方法，在IDT、IntraTD、MACC指标及弱监督分类任务中均取得更好表现。在中医文献与足球社区等中文语料场景中的实验进一步验证了其有效性与适用性。结果表明，SeedLLM框架能够在复杂文本环境中有效挖掘具有代表性和区分性的目标主题词，为舆情分析与议题监测提供技术支持。

参考文献：

[1] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[2] Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure[PP]. V1. (2022-05-11) [2026-03-20]. arXiv: arXiv.2203.05794.

[3] 黄发良, 冯时, 王大玲, 等. 基于多特征融合的微博主题情感挖掘[J]. 计算机学报, 2017, 40(4): 872-888.
Huang F L, Feng S, Wang D L, et al. Mining topic sentiment in microblogging based on multi-feature fusion[J]. Chinese Journal of Computers, 2017, 40(4): 872-888.

[4] Wang S, Chen Z Y, Fei G L, et al. Targeted topic modeling for focused analysis[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 1235-1244.

[5] Wang R, Wang Y N, Li Z A, et al. Distributed keyword-guided topic model with lexical knowledge supervision[J]. ACM Transactions on Knowledge Discovery from Data, 2025, 19(6): 1-19.

[6] Rakesh V, Ding W C, Ahuja A, et al. A sparse topic model for extracting aspect-specific summaries from online reviews[C]//Proceedings of the 2018 World Wide Web Conference. New York: ACM Press, 2018: 1573-1582.

[7] Kargupta P, Komarlu T, Yoon S, et al. MEGClass: extremely weakly supervised text classification via mutually-enhancing text granularities[C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg: ACL Press, 2023: 10543-10558.

[8] 王金芳, 郭渊博, 邓淼磊, 等. 面向威胁情报的小样本文档级事件抽取方法[J]. 通信学报, 2025, 46(12): 234-248.
Wang J F, Guo Y B, Deng M L, et al. Few-shot document-level event extraction method for threat intelligence[J]. Journal on Communications, 2025, 46(12): 234-248.

[9] Kumar S, Le T. FoTo: targeted visual topic modeling for focused analysis of short texts[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Piscataway: IEEE Press, 2024: 7406-7416.

[10] Jiang Z B, Xu F, Gao L Y, et al. Active retrieval augmented generation[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press, 2023: 7969-7992.

[11] Su W H, Tang Y C, Ai Q Y, et al. DRAGIN: dynamic retrieval augmented generation based on the real-time information needs of large language models[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL Press, 2024: 12991-13013.

[12] Liu H S, Zhang H, Guo Z J, et al. CtrlA: adaptive retrieval-augmented generation via inherent control[C]//Proceedings of the Findings of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2025: 12592-12618.

[13] Asai A, Wu Z Q, Wang Y Z, et al. Self-RAG: learning to retrieve, generate, and critique through self-reflection[C]//International Conference on Learning Representations. Vancouver: ICLR, 2024: 9112-9141.

[14] Zhang Y, Zhang Y Y, Michalski M, et al. Effective seed-guided topic discovery by integrating multiple types of contexts[C]//Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2023: 429-437.

[15] Jagarlamudi J, Daumé III H, Udupa R. Incorporating lexical priors into topic models[C]//Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2012: 204-213.

[16] Churchill R, Singh L, Ryan R, et al. A guided topic-noise model for short texts[C]//Proceedings of the ACM Web Conference 2022. New York: ACM Press, 2022: 2870-2878.

[17] Meng Y, Huang J X, Wang G Y, et al. Discriminative topic mining via category-name guided text embedding[C]//Proceedings of the Web Conference 2020. New York: ACM Press, 2020: 2121-2132.

[18] Pham C, Hoyle A, Sun S M, et al. TopicGPT: a prompt-based topic modeling framework[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Stroudsburg: ACL Press, 2024: 2956-2984.

[19] Xu W J, Hu W X, Wu F Y, et al. DeTiME: diffusion-enhanced topic modeling using encoder-decoder based LLM[C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg: ACL Press, 2023: 9040-9057.

[20] Akash P S, Chang K C. Enhancing short-text topic modeling with LLM-driven context expansion and prefix-tuned VAEs[C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024. Stroudsburg: ACL Press, 2024: 15635-15646.

[21] Doi T, Isonuma M, Yanaka H. Topic modeling for short texts with

large language models[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop). Stroudsburg: ACL Press, 2024: 21-33.

[22] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[PP]. V2. (2017-07-20)[2026-03-20]. arXiv: arXiv.1707.06347.

[23] Ramamurthy R, Ammanabrolu P, Brantley K, et al. Is reinforcement learning (not) for natural language processing: benchmarks, baselines, and building blocks for natural language policy optimization[PP]. V3. (2022-10-03)[2026-03-20]. arXiv: arXiv.2210.01241.

[24] Schulman J, Moritz P, Levine S, et al. High-dimensional continuous control using generalized advantage estimation[PP]. V6. (2018-10-20) [2026-03-20]. arXiv: arXiv.1506.02438.

[25] Ziegler D M, Stiennon N, Wu J, et al. Fine-tuning language models from human preferences[PP]. V2. (2020-01-08) [2026-03-20]. arXiv: arXiv.1909.08593.

[26] Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg: ACL Press, 2019: 3980-3990.

[27] Yue W J, Wang X L, Zhu W, et al. TCMBench: a comprehensive benchmark for evaluating large language models in traditional Chinese medicine[PP]. V1. (2024-06-03) [2026-03-20]. arXiv: arXiv. 2406.01126.

[28] Issa B, Jasser M B, Chua H N, et al. A comparative study on embedding models for keyword extraction using KeyBERT method[C]//Proceedings of the 2023 IEEE 13th International Conference on System Engineering and Technology (ICSET). Piscataway: IEEE Press, 2023: 40-45.

[29] Lau J H, Newman D, Baldwin T. Machine reading tea leaves: automatically evaluating topic coherence and topic model quality[C]//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2014: 530-539.

[30] Stammbach D, Zouhar V, Hoyle A, et al. Revisiting automated topic model evaluation with large language models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press, 2023: 9348-9357.

[31] Dieng A B, Ruiz F J R, Blei D M. Topic modeling in embedding spaces[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 439-453.

[32] Fu J L, Ng S K, Jiang Z B, et al. GPTScore: evaluate as you desire[C]// Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Stroudsburg: ACL Press, 2024: 6556-6576.

[33] Zar J H. Spearman rank correlation[M]//Armitage P, Colton T. Encyclopedia of biostatistics. Chichester: John Wiley & Sons, Ltd, 2005.

[34] Schober P, Boer C, Schwarte L A. Correlation coefficients: appropriate use and interpretation[J]. Anesthesia & Analgesia, 2018, 126(5): 1763-1768.

[作者简介]



曾祥（1997– ），男，国防科技大学博士生，主要研究方向为网络内容安全、开源情报分析、自然语言处理。



王晔（1989– ），女，哈尔滨工业大学（深圳）教授，主要研究方向为网络内容安全、开源情报分析、知识图谱推理。



杨博捷（2003– ），男，国防科技大学硕士生，主要研究方向为自然语言处理、网络内容安全。



周斌（1971– ），男，国防科技大学研究员、博士生导师，主要研究方向为网络内容安全、开源情报分析、自然语言处理。



王志超（1992– ），男，湖南四方天箭信息科技有限公司高级工程师，主要研究方向为多模态情报数据融合、情报数据推理与决策生成。



贾焰（1960– ），女，国防科技大学教授、博士生导师，主要研究方向为网络空间安全、大数据分析和人工智能。