计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(12): 2951−2961, 2025

基于大语言模型的知识库查询风格自适应转换

付钧渤　赵国帅　钱学明
（西安交通大学电子与信息学部　西安　712046）
（junbofu@stu.xjtu.edu.cn）

Adapative Query Style Transfer for Large Language Models as Knowledge Bases

Fu Junbo, Zhao Guoshuai, and Qian Xueming
（*Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an* 712046）

Abstract　The increasing capabilities of large language models (LLMs) in knowledge storage have underscored their potential utility as knowledge bases. However, it's important to note that any given prompt can merely offer a lower-bound estimate of the knowledge encompassed by the language model. Prior prompt learning methods in the context of language models as knowledge bases (LMs-as-KBs) have overlooked the influence of query style. We have unveiled a significant revelation—there are indeed learnable preference within LLMs pertaining to query style. Leveraging this distinctive model characteristic, we introduce the adaptive query style transfer (ARES) method to improve the performance of LMs-as-KBs by adapting LLM's preference. ARES initiates by presenting a candidate set of queries, achieved through paraphrasing to incorporate various expression styles. Subsequently, an evaluator is trained to learn and discern LLM's preferences for query styles, ultimately evaluating the candidate set and selecting the potentially optimal query. Experiments conducted across multiple datasets have convincingly demonstrated the efficacy of our approach in enhancing question answering accuracy on LMs-as-KBs scenarios. Furthermore, incremental comparison of the original model and three baseline methods shows an average improvement of 2.26%, 1.68%, 1.19%, and 1.17%, respectively, indicating ARES can be effectively utilized in conjunction with other approaches, leading to enhanced performance and optimization across different dimensions.

Key words　large language models (LLMs); language models as knowledge bases; promote learning; text style transfer; in-context learning

摘　要　大语言模型（large language models, LLMs）在知识存储方面不断增强的能力展示了其作为知识库的潜在效用. 然而, 任何给定的提示只能提供大语言模型所涵盖知识的下限估计. 在语言模型即知识库（language models as knowledge bases, LMs-as-KBs）的场景中, 先前的提示学习方法忽略了查询风格对模型表现的影响. 揭示了大语言模型确实具有与查询风格相关的可学习偏好, 并且利用大语言模型的这种特性引入了查询风格自适应转换（adaptive query style transfer, ARES）方法, 通过适应大语言模型的偏好来增强其知识查询的表现. ARES 方法从构造查询候选集开始, 通过改写实现多种表达风格的纳入. 随后, ARES 训练一个评估器来学习并识别大语言模型对查询风格的偏好, 评估查询候选集并选择潜在最优查询. 在多个数据集上进行的实验表明了该方法在提高大语言模型即知识库服务上查询准确率的有效性,

收稿日期: 2024-06-20; **修回日期**: 2025-04-09
基金项目: 国家自然科学基金项目(62372364); 陕西省技术创新引导计划项目(2024QCY-KXJ-199); 咸阳市重点研究计划项目(L2023-ZDYF-QYCX-002)
This work was supported by the National Natural Science Foundation of China (62372364), the Project of Shaanxi Provincial Technology Innovation Guidance (2024QCY-KXJ-199), and the Key Research and Development Plan of Xianyang City (L2023-ZDYF-QYCX-002).
通信作者: 赵国帅 (guoshuai.zhao@xjtu.edu.cn)

增量对比原始模型与3个基线方法分别实现了平均2.26个百分点,1.68个百分点,1.19个百分点,1.17个百分点的提升,这表明ARES可以与其他方法有效地结合使用,从而实现多角度的查询表现增强.

关键词 大语言模型;语言模型即知识库;提示学习;文本风格转换;上下文学习

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202440549 **CSTR:** 32373.14.issn1000-1239.202440549

随着预训练方法的不断进步及训练语料规模与模型参数量的不断增加,大语言模型(以下简称大模型)展示出了其在知识存储方面越来越强大的能力.在预训练过程中,大模型学习预训练语料中的事实性知识,并以大量的参数隐性地存储它们[1].基于此,一些研究提出将大模型作为知识库使用.相较于传统的知识库,大模型知识库具备一些独特的优势.在构建时,它们通过无监督训练节省了人工知识提取的成本,并实现了广泛的知识覆盖.在使用时,它们拥有灵活的查询方式、上下文理解能力和推理能力,从而能够更准确、自然地回答用户问题.因此,大模型可以在一些场景下代替结构化数据库,在知识密集型任务中发挥作用[2].

与传统结构化知识库的固定查询方式不同,大模型知识库的查询方式是较为灵活的自然语言查询[3].然而,如果没有给出恰当的输入,大模型所知道的知识事实可能无法被检索到.换句话说,任何给定的提示都只能获取大模型中所含知识的下界估计[4].因此,如何调整大模型作为知识库输入的提示是一个关键的问题.目前,可用于语言模型即知识库场景下的提示学习研究可分为2类:白盒优化和黑盒优化.白盒优化考虑在模型参数为可访问的前提下优化提示.例如在输入序列中添加并训练提示向量[5-7].然而,出于商业因素、安全问题以及隐私保护的考虑,大模型有时也会作为接口呈现给用户[8],在这种情况下则只能采用黑盒优化方法.例如,在输入中添加针对特定任务语义的固定提示模板[4],或从候选集中检索相关示例添加至提示中[9-10].然而,这些方法均忽视了查询本身表达方式的重要性.事实上,查询的表达方式也是一个可优化的因素.

当使用基于大模型的知识库时,在保持查询语义一致的前提下,可以有多种方式来表达一个查询.有趣的是,作为输入中一个可变因素,查询的风格可以影响大模型根据其固有的知识储备提供准确答案的能力.图1展示了大模型面对风格不同但语义一致的查询时的表现,该示例中使用的模型为MPT-7B,查询的正确答案应为“Democratic Party”.当大模型接收到查询风格1时,它生成的回复是包含幻觉的;当模型接收到查询风格2时,其掩盖了自己知道正确答案的事实;只有当模型接收到查询风格3时,其才根据自己固有的知识储备提供了正确答案.更重要的一点是,本文的研究揭示了在大模型被作为知识库使用时,其对于查询本身的表达方式存在着可学习的偏好.并且,本文的实验同时表明,这些偏好与人类对自然语言的常见偏好,例如信息性、连贯性和流畅性等不完全一致.

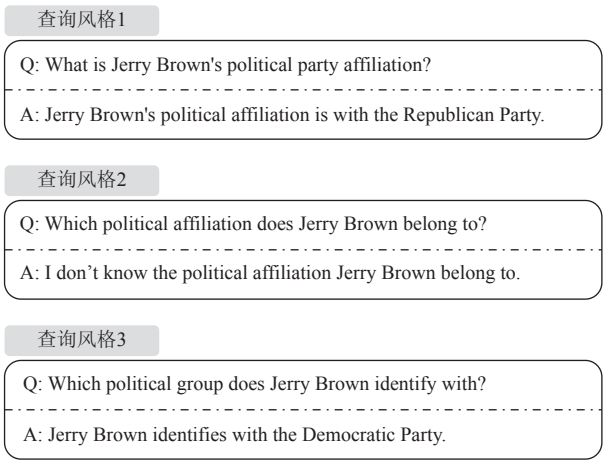


Fig. 1 LLMs' responses when faced with different styles of queries

图1 大语言模型面对不同风格查询时的回复

在本文中,我们提出了一种称为查询风格自适应转换(adaptive query style transfer, ARES)的方法,通过让查询的风格对齐大模型对表达方式的偏好来改善其作为知识库时的表现.值得注意的是,该方法也属于一种黑盒方法,即能够适用于模型参数不可访的场景.具体来说,ARES包括查询释义和查询选择2个主要步骤.查询释义通过修改原始查询的表达方式来生成一组候选查询.本文中,我们通过上下文学习增强了基于生成的查询释义方法.随后,查询选择依据大模型对候选集中每个查询的相对偏好选择一个适当的查询.在训练阶段,ARES首先收集大模型对不同风格查询的回答,并为每个查询依据其回答的正确与否进行标注,以建立评估器的训练集.然后训练评估器来学习大模型对不同查询样式的相对偏

好. 在推理阶段, ARES 利用训练后的评估器来评估大模型对每个查询的相对偏好, 并选择最有可能使大模型提供准确回答的潜在最优查询. 在多个数据集上进行的实验表明, ARES 能够显著提高大模型在作为知识库时的问答准确性. 此外, 本文还进行了增量式的对比试验, 证明 ARES 也可以与其他的提示学习方法有效地协同使用. 本文工作的主要贡献可以总结为 3 点:

1)探索并揭示了不仅大模型准确回答查询的能力会受到查询风格的影响, 而且对查询风格表现出可学习的偏好. 本文分析并证明了这种偏好与人类对自然语言的常见偏好具有一定差异.

2)介绍了 ARES, 一个旨在增强大模型知识库查询表现的优化方法. 我们利用大模型的反馈对释义生成方法进行了改进, 并设计了一种能够学习大模型对查询风格偏好的评估器. 据我们所知, 这是在大模型即知识库服务场景下考虑查询表达方式优化的工作.

3)在多个数据集上进行了广泛的实验, 以验证 ARES 的有效性. 本文的实验同时表明, ARES 还可以与其他类型的提示学习方法有效地结合使用, 从而实现不同角度上查询表现的优化.

1 相关工作

1.1 语言模型即知识库服务

虽然大模型近年来发展迅速, 但由于算力需求与商业因素等, 让其适用于所有的组织和个人仍是一个挑战[11]. 因此, Sun 等人[12] 提出了语言模型即服务, 即语言模型以接口的形式提供给用户来完成各种任务, 并进一步考虑了模型被视作黑盒时如何提高其表现. 目前, 将大模型作为服务发布已经成为一种常见的做法, 例如 GPT-4[13], Claude 3.0[14], Palm 2.0[15] 等.

同样地, 将大模型用作知识库时, 我们也更推荐将其作为一种服务呈现给使用者. 首先, 对于一些组织和个人来说, 部署与微调大型模型的成本是昂贵的, 而通过接口使用大模型的知识库服务则更为可行. 更重要的是: 一方面, 一些特定领域的微调模型包含的知识可能涉及隐私; 另一方面, 由于无法确定模型直接给出的回答是否是符合法律规范和道德伦理的, 直接将模型提供给用户使用具有一定的安全隐患[16-17], 而将大模型知识库作为接口开放则可以使知识库的外部输出管理更为可控. 本文提出的方法便考虑在这种场景下优化模型表现, 即模型被作为一个知识库服务提供给用户, 其输入和输出均为自然语言, 输入的优化只能在模型参数为不可访问的前提下完成.

1.2 提示学习相关方法

一般来说, 问答可以作为一个通用的方式来实现开放域上对语言模型的知识提取. 因此, 针对问答任务的提示优化通常可以用于语言模型即知识库服务的场景下. 例如 Jiang 等人[4] 在 2020 年提出 LAPQA, 该方法通过基于回译与挖掘的方式将任务相关的模板加入到问答的提示中. 此后, 又有一些自动构建连续型提示的方法被提出, 这些方法可以直接在模型的词嵌入空间中构建并优化提示, 但大多需要获取模型参数来完成反向梯度传播. 如 Li 等人[5] 在 2021 年提出的 Prefix Tuning, 该方法在输入前附加一系列连续的任务特定向量, 并在其训练时保持语言模型的参数冻结. 类似地, Lester 等人[6] 提出了 Prompt Tuning, 与 Prefix Tuning 相比, 它仅在嵌入层中插入任务特定的向量. 与此同时, Liu 等人[7] 引入了 P-Tuning, 该方法利用少量连续的自由参数作为提示注入到语言模型中, 并使用梯度下降进行优化, 作为对离散提示搜索的替代方法. 随着大模型参数量的增加和理解能力的增强, 其逐渐具备了上下文学习(in-context learning, ICL)的能力, 因此, 上下文学习也被当作一种提示学习方法应用于问答提示的优化中. 近年来, 一些工作聚焦于优化上下文学习中示例的检索. 如 Rubin 等人[9] 在 2022 年提出的 EPR, 该方法训练了一个基于学习的密集检索器, 以检索更好的单个上下文示例. 而 Ye 等人[10] 在 2023 年提出的 CEIL 则通过确定性点过程来建模给定输入和多个上下文示例之间的交互.

1.3 释义生成

释义生成是指构造与原句表达方式不同但句子含义相同的新句子[18], 被广泛地应用于自然语言处理的多个任务领域当中, 如机器翻译、信息检索及对话系统等. 现阶段, 句子改写的策略主要分为 2 种: 基于规则的方法和基于生成的方法. 早期基于规则的方法聚焦于基于特定规则对原始句进行有针对性的更改, 例如同义词替换[19]、句法转换[20]、模板嵌入[21] 以及表达式替换[22] 等. 尽管这些方法能对句子表达方式的改变提供精确的控制, 但会在保持句子流畅性上面临一定的挑战. 随着深度学习在过去 10 年的发展, 一些研究提出了基于神经网络的生成式句子改写方法. 生成式句子改写通过文本生成模型直接生成新句子, 依据生成过程可以分为端到端的改写[23] 与基于回译的改写[24], 依据改写角度则可以

分为泛改写与面向特定角度[25]的改写，如正式性[26]、复杂性[27]、礼貌性[28]等. 在大模型时代，语言模型的生成能力进一步增强，通过向模型提供一些简单的提示便可以让模型完成改写任务，且其效果能够超过大多数传统方法.

1.4 文本生成的机器评估指标

近年来，有一种趋势是使用神经网络模型来代替人工评估从多方面评价自动生成的文本[29]. 这个过程可以概括描述为：首先，标注人员依据人类自身对文本各方面特性的偏好，对生成的文本相较于人工构建的参照文本进行打分，以构建训练数据. 然后，训练神经网络模型通过标注数据来学习人类对于文本各方面特性的偏好，以用于自动评估生成的文本.

Sellam 等人[30]在 2020 年提出了 BLEURT，这是一种基于 BERT 的文本评估指标，可以依据数千个具有潜在有偏见的训练示例模拟人类的判断. 随后，Yuan 等人[29]在 2021 年提出了 BARTScore，该方法以无监督方式训练基于 BART 的评估模型，可以灵活应用于不同角度的文本评估. Wan 等人[31]在 2022 年进一步提出了 UniTE 方法，一个能够处理多形式评估任务的统一框架，该方法考虑了多种评估任务之间的共性，并提出了统一的预训练方法来更好地适应多任务学习. 以上这些机器评估指标均已经验证了它们在各种人类评估标准上的有效性.

2 ARES 方法

在基于大模型的知识库服务中，对输入的优化方法决定了能在多大程度上激发模型挖掘其内在知识的潜力. 先前的研究工作已经表明，由于语言模型在预训练过程中所学习的知识文本具有特定的表达方式，其对于语义一致但表达方式不同的输入会有不同的反馈[2]. 我们认为，对于语言模型在某一语义子空间上的预训练文本，其在表达方式上的分布应该存在着特定方向的“偏差”. 因此，ARES 的目标就是学习和适应这些“偏差”，也可称为语言模型在特定语境中对查询风格的“偏好”.

图 2 展示了 ARES 进行训练和推理的流程. 在训练阶段，ARES 通过大模型的回复获取查询候选集的对应标签，以构建评估器的训练集，然后以语言建模的方式训练评估器. 在推理阶段，ARES 通过查询释义-查询选择的过程来实现查询风格的自适应转换，并将潜在的最优查询提供给大模型以获取查询结果. 在这个过程中，查询释义决定了查询优化的上限，而查询选择则决定了其下限. 给定原始查询 q_0，ARES 首先构造一组查询候选集 $Q=\{q_0,q_1,\cdots,q_n\}$，然后使用经过训练的评估器对 Q 中的查询进行评估并根据评估结果挑选出潜在最优查询 q_{best}.

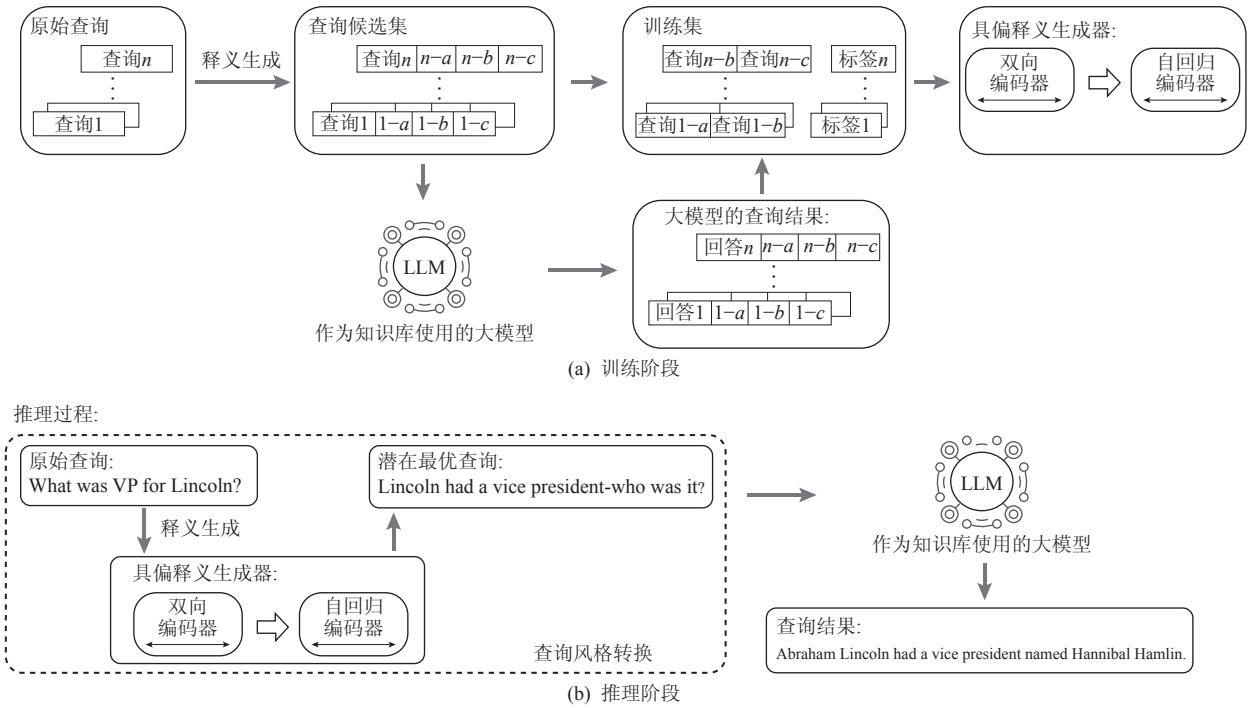


Fig. 2 The process of training and inference for ARES

图 2 ARES 进行训练和推理的流程

2.1 模型偏好学习

2.1.1 查询释义

ARES 采用了基于生成的释义方法，通过向语言模型中输入特定提示来构造查询的候选集，每个候选集中包含了原始查询以及释义后的查询. 我们希望释义生成器所生成的释义中能包含足够多模型所偏好的查询风格，以此来帮助评估器学习，因此采取了基于模型反馈的上下文学习来增强提示.

在对数据集正式释义前，首先要构建一个示例池，图 3 中展示了示例池的构建流程. 其中，E 代表示例池中的示例数量，T 代表预先设置的阈值，当 $E\geqslant T$ 时，示例池构建阶段结束，正式的释义过程开始. 具体而言，该流程会遍历数据集来为每个查询构建伪候选集，并获取大模型对伪候选集中每个查询的反馈. 对于一个伪候选集，如果其中的原查询无法使大模型回答出正确的答案而释义后的查询可以，则可以认为该候选集的反馈满足筛选条件并将其加入示例池中. 为了保证示例池中示例的数量能够满足模型对查询风格多样性的需求，这个过程可以迭代执行. 在构建示例池及正式释义时，我们均从示例池中随机挑选 k 个示例加入提示中，其中 k 取决于释义生成器可以接受的合适输入长度.

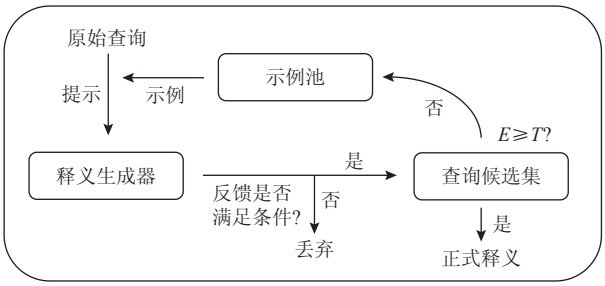


Fig. 3 Continuously performing the generation of paraphrases and the selection of examples

图 3 循环执行释义生成和示例选择

2.1.2 查询评估

为了挑选出潜在最优查询，首先要训练一个用于评估查询的评估器来学习大模型在各种语境下对表达方式的偏好. 由于每个查询最优表达方式的基本真值子集是无法获取的，因此不能应用传统的最大似然估计法来学习评估器的参数. 受以往的文本生成机器评估指标的启发，我们选择训练评估器学习大模型对表达方式的相对偏好.

第 1 步. 评估器训练集的构建. 首先，获取大模型对每一个候选集 Q^j 中的各个查询 q_i^j 的回复，其中 $j\in|M|,|M|$ 表示候选集的个数，$i\in|N|,|N|$ 表示每个候选集中包含的释义查询个数. 然后根据查询的答案判断大模型的回答是否正确来为查询赋予标签 $\hat{l}_i^j$，可以为“正”或“负”. 接着，使用标有标签的释义候选集来构建评估器的训练集. 该步骤的目标是构建一个训练集，其输入数据为一个有序查询对 $\{q_a^j,q_b^j\}$，且句对中的 2 个查询必须位于同一候选集中且标签不同. 即查询对中 2 个查询的标签仅可能为(负，正)或(正，负)，分别对应着不同的标签词 $\hat{w}_+$ 和 $\hat{w}_-$. 具体而言，我们会将每个候选集 Q^j 中所有标签为正的查询构建为正候选集 Q_{P}^j，所有标签为负的查询构建为负候选集 Q_{N}^j，然后将两者做正反笛卡儿积后取并集来构成最终的训练集：

$$D_{\mathrm{train}}=(Q_{\mathrm{P}}^j\times Q_{\mathrm{N}}^j)\cup(Q_{\mathrm{N}}^j\times Q_{\mathrm{P}}^j). \tag{1}$$

第 2 步. 参考一些先前工作的设定[32-33]，我们选择将我们的训练任务形式化为语言建模，这种方式能够有效缓解语言模型在预训练任务和下游任务上优化目标的不一致带来的性能下降. 我们采用 BART[34] 模型作为评估器的骨干，该模型保留了 Transformer 的编码器-解码器结构，我们希望通过其中的交叉注意力机制来捕捉不同风格的查询间的差异，以帮助评估器更好地学习语言模型的偏好. 对于一条训练数据中的查询对 $\{q_a^j,q_b^j\}$，ARES 将句对中的前者与提示 p_{former} 进行拼接并输入到 BART 的编码器中：

$$\boldsymbol{h}_{\mathrm{enc}}^j=ENCODER\left(p_{\mathrm{former}}+q_a^j\right). \tag{2}$$

根据 Transformer 并行训练中的 Teacher Force 机制，在解码器生成第 k 个 token 时，$\boldsymbol{h}_{\mathrm{enc}}^j$ 和目标输出序列的前 $k-1$ 个 tokens $t_{1:k-1}^{\mathrm{obj}}$ 作为输入，产生基于注意力的表示[35]：

$$\boldsymbol{h}_{\mathrm{dec}}^{j,k}=DECODER\left(\boldsymbol{h}_{\mathrm{enc}}^j,t_{1:k-1}^{\mathrm{obj}}\right). \tag{3}$$

token t_k 的条件概率定义为

$$P\left(t_k|t_{1:k-1},p_{\mathrm{former}}+q_a^j\right)=softmax\left(\boldsymbol{h}_{\mathrm{dec}}^{j,k}\boldsymbol{W}+\boldsymbol{b}\right), \tag{4}$$

其中 $\boldsymbol{W}\in\mathbb{R}^{d_h\times|\mathcal{V}|}$，$\boldsymbol{b}\in\mathbb{R}^{|\mathcal{V}|}$，$d_h$ 代表模型的隐藏维度，$|\mathcal{V}|$ 代表词典大小，两者均为可学习的参数，对于损失函数，我们采用解码器的输出和通过提示 p_{later}、句子 q_b^j 以及标签词 $\hat{w}^j$ 拼接所构成的 Ground Truth 的交叉熵. 本文中的 p_{former} 与 p_{later} 均为人工构造的用于帮助模型学习的提示，具体设置详见 3.1.2 节.

2.2 推 理

在测试时，对于一个查询 q_{test}^0，首先对其进行释义来生成一个查询候选集 Q_{test}. 然后获取 Q_{test} 中每个释义后查询 q_{test}^i 与原查询 q_{test}^0 构成的句对 $\left\{q_{\mathrm{test}}^0,q_{\mathrm{test}}^i\right\}$ 作为输入时评估器输出的标签词 $\hat{w}$ 对应的条件概率：

$$f(\hat{w},i)=p_{\hat{w}}\left(t_c|p_{\text{later}}+q^i_{\text{test}},q^0_{\text{test}}\right), \quad (5)$$

其中$\hat{w}\in\{\hat{w}_+,\hat{w}_-\}$, 具体设置详见 3.1.2 节.

接着, 通过训练后的评估器计算q^i_{test}相对于q^0_{test}受语言模型的偏好程度:

$$R_score(i)=\underset{\hat{w}_+}{softmax}\left(\left[f(\hat{w}_+,i),f(\hat{w}_-,i)\right]\right)-1/2. \quad (6)$$

然后, 选取释义后的查询中相对于原查询的R_score得分最高且为正的查询作为潜在最优查询, 如果所有释义后的查询相对于原查询的R_score得分均为负, 则选择保留原查询作为潜在最优查询. 至此, 从原查询到潜在最优查询的风格自适应转换已完成.

最终, 将风格转换后的查询添加到大模型的提示模板中, 获取对知识库查询的输出结果.

3 实 验

3.1 实验设置

3.1.1 数据集

我们将向大模型知识库查询知识的过程视为一个问答过程. 因此, 在本文中, 我们选取了 3 个常见的开放域问答数据集进行实验.

1)WebQuestionsSP(WebQSP)数据集[36]是由 Yih 等人[36]创建的用于语义解析的问答数据集. 该数据集通过专家使用可行的SPARQL 查询对 WebQuestions[37]中的实例进行注释构建而成, 包括 4 737 个涉及 Freebase 中实体自然语言问题, 分为 3 098 个训练问题和 1 639 个测试问题. 我们从训练集中选择了 2 000 条数据作为实验的训练集, 并以完整的验证集作为实验的测试集.

2)ComplexQuestions(CompQ)数据集[38]是一个由 Bao 等人[38]构建的多约束问答数据集, 包含了通过挖掘搜索查询日志中具有重叠术语的查询而获得的问题-答案对, 类似于 WebQuestions[37]和 SimpleQuestions[39]数据集. 该数据集中的问题均经过手动的答案标注, 包含 1 300 条训练数据和 800 条测试数据. 我们直接遵循数据集的原始划分作为实验所用的训练集和测试集.

3)ComplexWebQuestions(CWQ)数据集[40]是一个由 Talmor 等人[40]构建的问答数据集. 他们利用 WebQSP[36]中的部分数据通过程序生成更复杂的伪自然语言形式查询, 再由众包工作者改进为自然语言问题. CWQ 共包含 35 000 个复杂问题, 包括 27 639 条训练数据、3 519 条验证数据及 3 531 条测试数据. 我们从训练数据中选择了 2 000 条数据作为实验的训练集. 而由于测试集的答案是不可见的, 我们选择使用完整的验证集进行测试.

3.1.2 实现细节与评价指标

我们采用 MPT-7B[41]作为实验中提供知识库服务的主要大模型, 该模型是一个参数为 7×10^9 的解码器风格 Transformer, 在 1TB 规模的英文文本和代码语料上进行了预训练. 此外, 我们还考虑了使用 GPT-neo-1.3B[42], 一个在 The pile[43]预训练语料上训练的类 GPT-3 架构的大模型, 以及 BLOOMZ-7b1[44], 一个在跨语言任务数据集上微调了 BLOOM 的多语言大模型. 我们选取了 CoEdIT-XL[45], 一个具有 3×10^9 参数的在文本编辑任务上微调了 Flan-T5[46]的语言模型作为我们的释义生成器. 对于所有语言模型的生成解码, 我们均使用波束搜索作为解码方式, 波束数量设置为 4 并使用随机采样.

我们选择精确匹配(exact match, EM)作为问答任务的评价指标. 对于评估器的训练, 我们选用 BART-Large 作为骨干架构, 训练时的学习率均设置为 E−5, 批大小为 32, 训练轮次数为 10. 此外, 我们将提示模板p_{former}与p_{later}分别设置为“Compared to the expression of _”和“The expression of _ is”, 并将标签词$\hat{w}_+$和$\hat{w}_-$分别设置为“better”和“worse”. 本文的所有实验均在 2 张 NVIDIA-A100 GPU 上完成.

对于采用了 ARES 方法的实验结果, 我们报告了 5 次重复实验的平均值和标准差.

3.2 增量式对比实验

我们使用 MPT-7B 作为提供知识库服务的大模型, 在 3 个开放域问答数据集上进行了增量式对比实验, 实验结果如表 1 所示. 我们从 2 类提示学习方法中选择了 3 种可用于语言模型即知识库服务场景的方法.

1)LPAQA[4]. 该方法向输入中加入任务相关的固定提示. 我们遵循原论文的设置, 使用基于挖掘的方法生成多样化的提示模板, 然后对训练集使用原文中的 Top-1 Prompt Selection 选择最佳提示添加到输入中.

2)TopK-BM25[47]. 该方法通过向输入添加与查询内容相关的示例来优化提示, 属于一种上下文学习方法. TopK-BM25 使用经典的稀疏检索方法 BM25, 在训练集中选择得分最高的前 k 个示例作为上下文.

3)CEIL[10]. CEIL 是由 Ye 等人[10]在 2023 年提出的一种方法, 也属于上下文学习方法. 该方法通过确定性点过程建模给定输入与上下文示例之间的交互, 并通过精心设计的对比学习目标优化检索模型.

Table 1 Experimental Results of QA Accuracy of Different Methods on Three Open Domain QA Datasets

表 1 在 3 个开放域问答数据集上各方法的问答准确率实验结果 %

方法	WebQSP	CompQ	CWQ	均值
LLM	42.71	26.63	29.16	32.83
LLM + ARES（本文）	44.80±0.25	29.95±0.12	30.52±0.77	35.09
提升值	+2.09↑	+3.32↑	+1.36↑	+2.26↑
LLM + LPAQA	43.56	29.75	30.32	34.54
LLM + LPAQA + ARES（本文）	46.07±0.38	31.67±0.16	30.92±0.30	36.22
提升值	+2.51↑	+1.92↑	+0.60↑	+1.68↑
LLM + LPAQA + BM25	47.65	31.25	34.90	37.93
LLM + LPAQA + BM25 + ARES（本文）	48.24±1.09	33.07±0.26	36.05±0.21	39.12
提升值	+0.59↑	+1.82↑	+1.15↑	+1.19↑
LLM + LPAQA + CEIL	47.90	31.63	37.99	39.17
LLM + LPAQA + CEIL + ARES（本文）	**49.13±0.54**	**33.22±0.13**	**38.69±0.11**	**40.35**
提升值	+1.23↑	+1.59↑	+0.70↑	+1.17↑

注：黑体和下划线的数据分别为表现最优和次优的结果，↑代表结合了 ARES 的方法相较于原方法的提升. “±”前后数据分别为 5 次重复实验的平均值和标准差.

我们将这 3 种方法及其组合作为对比基线，并将 ARES 与这些对比基线结合使用以进行增量式对比.

实验结果显示，ARES 在 3 个数据集上对比结合了手工构造提示的输入平均提高了 2.26 个百分点. 这表明 ARES 确实能够有效地提高大模型的知识查询准确率. 此外，需要清晰认识到的一点是，当将 2 种不同的提示学习方法结合使用时，它们的优化空间可能会重叠. 但当我们的方法与 LPAQA, LPAQA+BM25, LPAQA+CEIL.

结合使用时，实验结果仍显示出了相对基线方法 1.68 个百分点，1.19 个百分点，1.17 个百分点的准确率提升. 这表明 ARES 不仅在单独使用时有效，在与其他提示学习方法结合时仍能实现可观的准确率提升. 这也从侧面说明了 ARES 确实考虑并优化了以往研究所忽视的一个方面，即查询本身的风格，从而在新的维度上实现了模型查询表现的提升.

3.3 实验分析

3.3.1 ARES 的潜力

为了探究表达方式的优化在大模型即知识库服务场景下对知识查询效果提升的潜力，我们在 WebQSP 数据集上进行了 2 个方面的实验. 一方面，我们探究随着释义数量的增加，ARES 查询准确率的变化. 另一方面，我们探究在假设每次都能命中最优查询的前提下，查询准确率随释义数量的变化，我们将其称之为查询的潜在准确率. 实验结果如图 4 所示. 可以看到，随着释义数量的增加，潜在准确率的增长趋势虽然逐渐放缓，但一直在提升. 而查询准确率在释义数量大于等于 5 后虽有波动但总体趋于不变. 我们将此归因于随着释义数量的增多，虽然候选集中出现能使模型正确回答问题的查询的概率越来越大，但将其在众多释义中挑选出来的难度也越来越大. 由此可见，在大模型作为知识库场景下，表达方式的优化尚有很大待发掘的潜力，如何将其充分挖掘则有待后续工作的探索.

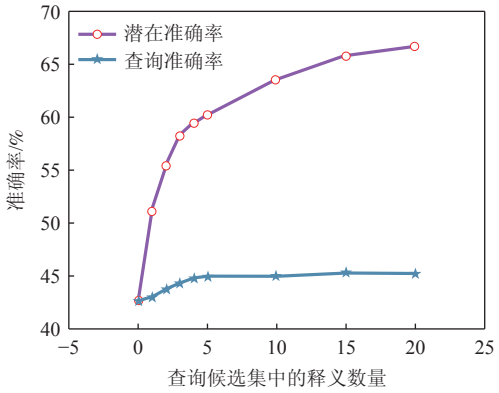


Fig. 4 Query accuracy and potential accuracy change with the number of interpretations in the candidate set

图 4 查询准确率和潜在准确率随候选集中释义数量的变化

3.3.2 数据集迁移实验

本文所使用的数据集虽然均为开放域数据集，但其相互之间也存在着一些方面上的差异. 其中，WebQSP 和 CWQ 两者的知识域相同，而与 CompQ 不完全重合；WebQSP 和 CompQ 两者的问题句式相对简单，而 CWQ 的问题句式则较为复杂. 我们将每个在特定数据集上训练的评估器同时应用于 3 个数据集的测试集进行测试，实验结果如图 5 所示. 实验表明，在不同数据集上训练的评估器在跨数据集使用时均能实现查询准确率的提升，这表明 ARES 所训练的评估器具有一定跨数据集的能力. 此外，可以观察到，在 CWQ 数据集上训练的评估器在另外 2 个测试集上的提升效果较为可观，而在 CompQ 和 WebQSP 数据集上训练的评估器在 CWQ 测试集上的提升作用则不明显. 我们认为，这是由于模型在复杂句式数据集上所学习到的偏好可以有效地迁移到简单句式的查询中，而在简单句式数据集上学习到的偏好则难以迁移到复杂句式的查询上.

3.3.3 模型迁移实验

我们在 WebQSP 数据集上使用不同参数规模和

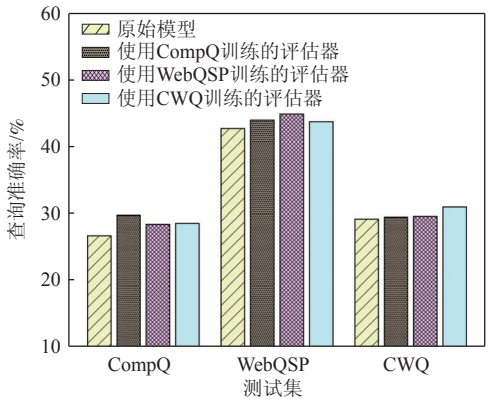


Fig. 5 Query accuracy of evaluators trained by different training sets in cross-test set testing

图 5 不同训练集所训练的评估器在跨测试集测试时的查询准确率

架构的大模型验证了 ARES 在模型迁移方面的有效性, 包括 GPT-neo-1.3B, BLOOM-7B, MPT-7B 以及 MPT-30B, 实验结果如图 6 所示. 可以观察到, 无论是应用在早期、参数量较小的模型还是近期、中等及较大参数量的模型上, ARES 均实现了查询准确率的有效提升.

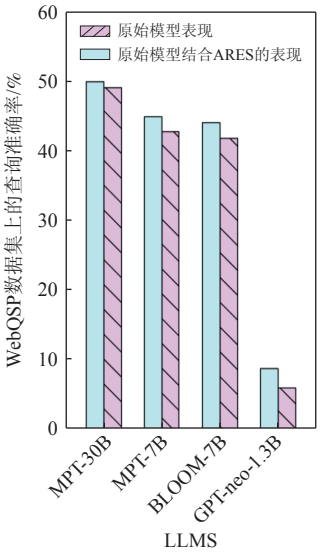


Fig. 6 Performance of ARES on LLMs with different structures and parameter scales

图 6 ARES 在不同结构和参数规模的大语言模型上的表现

3.3.4 大模型对查询风格的偏好性表现

描述作为知识库服务的大模型对查询风格的偏好是一个复杂且难以明确回答的问题. 这是由于模型会对不同的语义表现出多样化的偏好, 而这些偏好难以用精确的术语直接概括. 因此, 我们选择基于潜在最优查询中使用的释义类型来具体化模型的偏好. 具体而言, 我们将模型的偏好表示为释义后能够使模型正确回答而释义前不能正确回答问题的释义类型分布. 对于释义类型的分类, 我们遵循 Wahle 等人[48] 提出的标准, 该标准将释义类型划分为 7 个一级类别和 26 个二级类别.

我们利用 WebQSP 数据集的测试集来分析模型的偏好, 表 2 展示了占比最高的前 5 种释义类型. 从一级分类的角度来看, 模型对表达方式的偏好主要体现在其对基于词法变换的释义、基于句法变换的释义和其他类别 3 种释义类型的敏感性上. 从二级分类的角度来看, 模型对表达方式的偏好主要体现在其对格式转换、基于语义的变化和顺序变换 3 种释义类型的敏感性上. 其中, 模型对格式转换表现出显著的敏感性. 该实验也同时体现了 ARES 的自适应性, 即评估器能够根据不同的原始查询语义自适应地选择合适的释义方式, 而不是根据固定的规则进行释义.

Table 2 Five Interpretation Types with the Highest Proportion on WebQSP Dataset

表 2 在 WebQSP 数据集上占比最高的 5 种释义类型

一级释义类型	二级释义类型	占比/%
词法变换	格式转换	44.71
其他类型	语义释义	14.82
其他类型	顺序变换	12.96
词法变换	拼写更改	6.62
形态变换	情态动词变换	2.77

3.3.5 ARES 在推理上的额外时间成本

ARES 消耗的额外推理时间由查询释义时间和查询选择时间 2 部分构成. 由于评估器的评估执行仅需一次解码, 且其参数规模远小于提供知识库服务的大模型, 因此在考虑额外推理时间时可以忽略查询选择时间, 主要关注释义生成所需的时间. 设原问题的序列长度为 l_q, 释义生成数量为 n, 假设生成释义的平均序列长度与原问题的序列长度相同, 当释义生成模型的推理速率为 k(单位为 token/s)时, ARES 消耗的额外推理时间大致可表示为 $l_q \times n/k$.

ARES 的主要对比方法名为上下文学习, 所消耗的额外推理时间由查询编码时间和向量搜索时间 2 部分构成. 由于编码执行仅需 1 次解码, 且其参数规模远小于提供知识库服务的大模型, 因此查询编码时间可以忽略. 向量搜索所需时间主要取决于所采用的近似算法的时间复杂度及示例池中示例的数量. 鉴于本文实验所使用的示例池规模仅在 10^4 数量级, 因此向量搜索的时间也可以忽略.

由于模型在实际中的推理速度会受到计算量、

访存量、硬件特性、软件实现和系统环境等诸多因素的影响，我们在此仅讨论在本文实验环境下 ARES 的额外推理时间消耗表现. 对于提供知识库服务的大模型 MPT-7B，以及不同规模的释义生成器 CoEdIT-L 和 CoEdIT-XL，我们在 WebQSP 数据集上测试得到的推理速度如表 3 所示. 根据该实验结果，可以认为当提供知识库服务的大模型所消耗的推理时间为 T 且遵循主实验中所设置的释义数量 $n=5$ 时，ARES 在 CoEdIT-XL 和 CoEdIT-L 上的额外推理时间分别为 0.94T 和 0.30T. 尽管与上下文学习方法相比，ARES 的额外推理时间成本较高，但 ARES 与上下文学习方法的优化空间并不冲突并且可以结合使用，因此在对推理速度需求较低但对查询质量要求较高的场景下，ARES 是一种可行的方法.

Table 3 Average Inference Speed of Different Models on WebQSP Dataset

表 3 WebQSP 数据集上各模型的平均推理速度 token/s

模型	MPT-7B	CoEdIT-XL	CoEdIT-L
推理速度	0.46	2.44	7.72

3.4 消融实验

3.4.1 模型反馈对上下文学习的影响

本节分析了大模型的反馈对上下文学习中的示例筛选的作用. 具体而言，我们将释义生成中的示例添加方法替换为在所有示例中随机挑选示例添加至输入中来进行模型反馈的消融实验. 表 4 给出了在 WebQSP 数据集上释义数量 1~5 时的实验结果. 可以发现，基于模型反馈的上下文学习实现的效果在释义数量等于 3 时便超越了使用随机上下文学习在释义数量等于 5 时的效果，这反映了模型的反馈对示例能起到有效的筛选作用，从而提升评估器训练集的质量. 此外，使用了随机上下文学习的查询准确率随着释义数量的增加而上升，我们认为这种现象可以归因为一部分低质量的释义影响了训练过程. 而使用了基于模型反馈的上下文学习的查询准确率随着释义数量的增加平稳上升，因此可以认为 ARES 有助于平衡一个候选集内释义的质量.

Table 4 Comparison of Query Accuracy Between ARES' s Estimator and Other Evaluation Methods on Different Datasets

表 4 ARES 的评估器与其他评估方法在不同数据集上的查询准确率比较 %

评估方法	WebQSP	CompQ	CWQ
No Transfer	42.71	26.63	29.16
BLUERT	42.40	23.38	28.37
UniTE	43.38	27.13	28.34
ARES（本文）	44.80	29.95	30.52

3.4.2 评估器对释义挑选的影响

为了验证 ARES 的评估器相较于传统基于人类偏好的文本生成机器评估指标的有效性，我们在 3 个数据集上将 ARES 的评估器替换为传统的文本生成机器评估指标 BLUERT 以及 UniTE 来进行消融实验，实验结果如图 7 所示. 可以看到，对于较早的文本生成机器评估指标 BLUERT 而言，其对应的实验结果在 3 个数据集上均没有实现查询准确率的提升，这表明 BLUERT 作为评估指标无法有效挑选出候选集中模型所偏好的查询. 而对于改进的文本生成机器评估指标 UniTE 而言，其对应的实验结果在查询句式较为简单的数据集 WebQSP 和 CompQ 上实现了较小的准确率提升，但在查询句式较为复杂的数据集 CWQ 上未实现准确率的提升，这表明 UniTE 作为评估指标在查询句式较为简单的语境下，具有一定的挑选潜在最优查询的能力，但在句式复杂的查询语境下则无法完成挑选. 相比之下，我们提出的评估方法在 3 个数据集下均对查询准确率实现了可观的提升，这证实了 ARES 的有效性，同时也侧面说明了模型对查询问题的表达方式存在着特定方向上的偏好，且这种偏好方向与人类对自然语言的常见偏好具有差异.

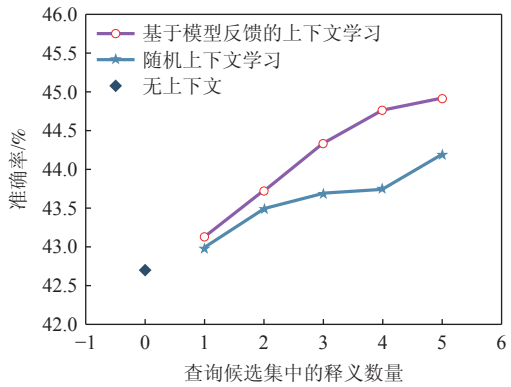


Fig. 7 Impact of adding model feedback on accuracy of query in WebQSP dataset

图 7 在 WebQSP 数据集上添加模型反馈对查询准确率的影响

4 总 结

我们提出了 ARES，一个能够自适应地转换查询风格，以增强大模型作为知识库时查询表现的方法. ARES 可以有效地与其他调优方法结合使用，在查询中实现效果的进一步提升. 此外，我们还证明了

ARES在不同问题复杂程度的问答数据集和不同规模及架构的语言模型上的有效性. 总的来说, ARES是应用在语言模型即知识库场景下的提示学习方法的一个重要补充, 提供了一种新颖的黑盒调优方法来实现大模型查询表现的优化.

作者贡献声明: 付钧渤负责论文的总体设计、实验的设计与分析、论文撰写及修改; 赵国帅对研究的整体思路进行指导, 参与研究设计, 对论文进行修改与润色; 钱学明提供了重要的理论支持和指导, 参与研究讨论并提供关键性的反馈意见.

参 考 文 献

[1] Cao Boxi, Lin Hongyu, Han Xianpei, et al. Knowledgeable or educated guess? Revisiting language models as knowledge bases[C]// Proc of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing. Stroudsburg, PA: ACL, 2021: 1860–1874

[2] Heinzerling B, Inui K. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries[C]// Proc of the 16th Conf of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2021: 1772–1791

[3] Petroni F, Rocktaschel T, Riedel S, et al. Language models as knowledge bases?[C]//Proc of the 24th Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing (EMNLPIJCNLP). Stroudsburg, PA: ACL, 2021: 1860–1874

[4] Jiang Zhengbao, Xu F, Araki J, et al. How can we know what language models know?[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 423−438

[5] Li X, Liang P. Prefix-tuning: Optimizing continuous prompts for generation[C]//Proc of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing. Stroudsburg, PA: ACL, 2021: 4582–4597

[6] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning[C]//Proc of the 26th Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2021: 3045–3059

[7] Liu Xiao, Ji Kaixuan, Fu Yicheng, et al. P-Tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2022: 61–68

[8] Wang Xiaochen, Zhang Kun, Zhang Peng. Large model safety and practice from multiple perspectives[J]. Journal of Computer Research and Development, 2024, 61(5): 1104−1112 (in Chinese)
(王笑尘, 张坤, 张鹏. 多视角看大模型安全及实践[J]. 计算机研究与发展, 2024, 61(5): 1104−1112)

[9] Rubin O, Herzig J, Berant J. Learning to retrieve prompts for in-context learning[C]//Proc of the 16th Conf of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2022: 2655–2671

[10] Ye Jiacheng, Wu Zhiyong, Feng Jiangtao, et al. Compositional exemplars for in-context learning[C]//Proc of the 40th Int Conf on Machine Learning. New York: JMLR. org, 2023: 39818–39833

[11] Chen Huimin, Liu Zhiyuan, Sun Maosong. The social opportunities and challenges in the era of large language models[J]. Journal of Computer Research and Development, 2024, 61(5): 1094−1103 (in Chinese)
(陈慧敏, 刘知远, 孙茂松. 大模型时代的社会机遇与挑战[J]. 计算机研究与发展, 2024, 61(5): 1094−1103)

[12] Sun Tianxiang, Shao Yunfan, Qian Hong, et al. Black-box tuning for language-model-as-a-service[C]//Proc of the 39th Int Conf on Machine Learning. New York: JMLR. org, 2022: 20841–20855

[13] Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report[R/OL]. [2023-03-27]. https://cdn.openai.com/papers/gpt-4.pdf

[14] Anthropic. The Claude 3 model family: Opus, sonnet, haiku[R/OL]. [2024-03-04]. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

[15] Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models[C]//Proc of the 36th Int Conf on Neural Information Processing Systems. New York: Curran Associates Inc, 2022: 2176

[16] Yi Xiaoyuan, Xie Xing. Unpacking the ethical value alignment in big models[J]. Journal of Computer Research and Development, 2023, 60(9): 1926−1945 (in Chinese)
(矣晓沅, 谢幸. 大模型道德价值观对齐问题剖析[J]. 计算机研究与发展, 2023, 60(9): 1926−1945)

[17] Li Nan, Ding Yidong, Jiang Haoyu, et al. Jailbreak attack for large language models: A survey[J]. Journal of Computer Research and Development, 2024, 61(5): 1156−1181 (in Chinese)
(李南, 丁益东, 江浩宇, 等. 面向大模型的越狱攻击综述[J]. 计算机研究与发展, 2024, 61(5): 1156−1181)

[18] Madnani N, Dorr B. Generating phrasal and sentential paraphrases: A survey of data-driven methods[J]. Computational Linguistics, 2010, 36(3): 341−387

[19] Wei Jason, Zou Kai. EDA: Easy data augmentation techniques for boosting performance on text classification tasks[C]//Proc of the 24th Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: ACL, 2019: 6382–6388

[20] Androutsopoulos I, Malakasiotis P. A survey of paraphrasing and textual entailment methods[J]. Journal of Artificial Intelligence Research, 2010, 38: 135−187

[21] Su Yixuan, Vandyke D, Wang Sihui, et al. Planthen-generate: Controlled data-to-text generation via planning[C]//Proc of the 26th Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2021: 895–909

[22] Mairesse F, Walker M. Controlling user perceptions of linguistic style: Trainable generation of personality traits[J]. Computational Linguistics, 2011, 37(3): 455−488

[23] Jegadeesan M, Kumar S, Wieting J, et al. Improving the diversity of unsupervised paraphrasing with embedding outputs[C]//Proc of the 1st Workshop on Multilingual Representation Learning. Stroudsburg, PA: ACL, 2021: 166–175

[24] Prabhumoye S, Tsvetkov Y, Salakhutdinov R, et al. Style transfer

through back-translation[C]//Proc of the the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 866–876

[25] Jin Di, Jin Zhijing, Hu Zhiting, et al. Deep learning for text style transfer: A survey[J]. Computational Linguistics, 2022, 48(1): 155−205

[26] Hovy E. Generating natural language under pragmatic constraints[J]. Journal of Pragmatics, 1987, 11(6): 689−719

[27] Tan S S L, Goonawardene N. Internet health information seeking and the patient-physician relationship: A systematic review[J]. Journal of medical Internet research, 2017, 19(1): e9

[28] Madaan A, Setlur A, Parekh T, et al. Politeness transfer: A tag and generate approach[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 1869–1881

[29] Yuan Weizhe, Graham Neubig, Liu Pengfei. BARTScore: Evaluating generated text as text generation[C]//Proc of the 35th Int Conf on Neural Information Processing Systems. New York: Curran Associates Inc, 2021: 2088

[30] Sellam T, Das D, Parikh A. BLEURT: Learning robust metrics for text generation[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 7881–7892

[31] Wan Yu, Liu Dayiheng, Yang Baosong, et al. UniTE: Unified translation evaluation[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2022: 8117–8127

[32] Han Xu, Zhao Weilin, Ding Ning, et al. PTR: Prompt tuning with rules for text classification[J]. AI Open, 2022, 3: 182−192

[33] Schick T, Schmid H, Schutze H. Automatically identifying words that can serve as labels for few-shot text classification[C]//Proc of the 28th Int Conf on Computational Linguistics. Stroudsburg, PA: ICCL, 2020: 5569–5578

[34] Lewis M, Liu Yinhan, Goyal N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 7871–7880

[35] Ashish V, Noam S, Niki P, et al. Attention is all you need[C]//Proc of the 31st Int Conf on Neural Information Processing Systems. New York: Curran Associates Inc, 2017: 6000–6010

[36] Yih W, Richardson M, Meek C, et al. The value of semantic parse labeling for knowledge base question answering[C]//Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2016: 201–206

[37] Berant J, Chou A, Frostig R, et al. Semantic parsing on freebase from question-answer pairs[C]//Proc of the 18th Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2013: 1533–1544

[38] Bao Junwei, Duan Nan, Yan Zhao, et al. Constraint-based question answering with knowledge graph[C]//Proc of the 26th Int Conf on Computational Linguistics. Stroudsburg, PA: ICCL, 2016: 2503–2514

[39] Bordes A, Usunier N, Chopra S, et al. Large-scale simple question answering with memory networks[J]. arXiv preprint, arXiv: 1506.02075, 2015

[40] Talmor A, Berant J. The Web as a knowledge base for answering complex questions[C]//Proc of the 16th Conf of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 641–651

[41] MosaicML NLP Team. Introducing MPT-7B: A new standard for open-source, commercially usable LLMs[R/OL].[2024-02-05]. https://www.databricks.com/blog/mpt-7b

[42] Black S, Gao L, Biderman S, et al. GPT-neo: Large scale autoregressive language modeling with mesh-tensorflow[CP/OL]. [2024-03-17]. https://github.com/EleutherAI/gpt-neo

[43] Gao L, Biderman S, Black S, et al. The pile: An 800 GB dataset of diverse text for language modeling.[J]. arXiv preprint, arXiv: 2101.00027, 2015

[44] Muennighoff N, Wang T, Sutawika L, et al. Crosslingual generalization through multitask finetuning[C]//Proc of the 54th Annual Meeting of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2023: 15991–16111

[45] Raheja V, Kumar D, Koo R, et al. CoEdIT: Text editing by task-specific instruction tuning[C]//Proc of the 28th Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 5274–5291

[46] Chung W, Hou L, Longpre S, et al. Scaling instruction-finetuned language models[J]. Journal of Machine Learning Research, 2024, 25(70): 153

[47] Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond[J]. Foundations and Trends® in Information Retrieval, 2009, 3(4): 333−389

[48] Wahle P, Gipp B, Ruas T. Paraphrase types for generation and detection[C]//Proc of the 28th Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 12148−12164



Fu Junbo, born in 2001. Master. His main research interests include natural language processing and information retrieval.

付钧渤, 2001 年生. 硕士, 主要研究方向为自然语言处理、信息检索.



Zhao Guoshuai, born in 1991. PhD, associate professor, PhD supervisor. His main research interests include social multimedia data mining, information retrieval, and personalized recommendation.

赵国帅, 1991 年生. 博士, 副教授, 博士生导师. 主要研究方向为社会多媒体数据挖掘、信息检索、个性化推荐.



Qian Xueming, born in 1974. PhD, professor, PhD supervisor. His main research interests include computer vision, and data mining and retrieval.

钱学明, 1974 年生. 博士, 教授, 博士生导师. 主要研究方向为计算机视觉、数据挖掘与检索.