文章编号:1005-0523(2026)02-0115-12

基于增强全局特征提取的分类大模型框架

陈可纬[1,2],刘建华[1,2],陈治铭[1],柯添赐[1],徐 戈[2]

(1. 福建理工大学计算机科学与数学学院,福建 福州 350118;
2. 闽江学院福建省信息处理与智能控制重点实验室,福建 福州 350108)

摘要:大语言模型(LLMs)通常采用指令微调方法适应下游任务,以增强其泛化能力,然而该方法针对LLMs的分类任务存在一定的性能局限性,有时无法满足任务需要。针对上述问题,提出一种全局特征提取分类大模型框架。该框架使用本文提出的全局特征提取增强方法,在注意力层释放全局特征,再对特征进行增强,并在微调的过程中运用低秩微调优化损失,最后构建一个全局特征提取的分类大模型。与基线模型RoBERTa相比,在通用情感分析数据集SST-2和AGNews上,准确率分别提升1.44个百分点和0.95个百分点。与基线模型PIQN模型相比,在通用命名实体识别(NER)数据集OntoNotes和CoNLL2003中,F_1分数分别提升0.79%和1.99%。实验结果表明,在不需要复杂的提示工程或外部知识的条件下,使用该框架的大模型性能显著优于其数倍规模的LLMs。

关键词:大语言模型;分类任务;命名实体识别;情感分析

中图分类号:TP399 **文献标志码:**A

本文引用格式:陈可纬,刘建华,陈治铭,等. 基于增强全局特征提取的分类大模型框架[J]. 华东交通大学学报,2026,43(2):115-126.

A Classification Framework Based on Enhanced Global Feature Extraction for Large Models

Chen Kewei[1,2], Liu Jianhua[1,2], Chen Zhiming[1], Ke Tianci[1], Xu Ge[2]

(1. College of Computer Science and Mathematics, Fujian University of Technology, Fuzhou 350118, China;
2. Fujian Provincial Key Laboratory of Big Data Mining and Applications, Fuzhou 350108, China)

Abstract: Large language models (LLMs) are often trained with instruction fine-tuning to adapt to downstream tasks to enhance their generalization ability, but this method has certain performance limitations for LLMs' classification tasks, and sometimes cannot meet the task requirements. To address this issue, a global feature extraction classification large model framework is proposed. This framework uses the global feature extraction enhancement method proposed in this paper to release global features in the attention layer, then enhance the features, and apply the depth low-rank fine-tuning optimization loss proposed in this paper during fine-tuning. Finally, a global feature extraction classification large model is constructed. Compared with the baseline model RoBERTa, the accuracy on the general sentiment analysis dataset SST-2 and AGNews was improved by 1.44 and 0.95 percentage points, respectively. Compared with the baseline model PIQN, the F_1 score on the general named

收稿日期:2024-12-24
基金项目:国家自然科学基金项目(62172095);福建省自然科学基金项目(2023J01349);闽江学院福建省信息处理与智能控制重点实验室开放课题(MJUKF-IPIC2024402)

entity recognition (NER) dataset OntoNotes and CoNLL2003 was improved by 0.79% and 1.99%, respectively. The experimental results show that, under the condition of not requiring complex prompt engineering or external knowledge, the performance of the large model using this framework is significantly better than that of its several times larger LLMs.

Key words: LLMs; classification task; named entity recognition (NER); sentiment analysis

Citation format: CHEN K W, LIU J H, CHEN Z M, et al. A classification framework based on enhanced global feature extraction for large models[J]. Journal of East China Jiaotong University, 2026, 43(2): 115−126.

近年来，尽管大语言模型（large language models，LLMs）在NLP领域表现显著[1]，但在结构化分类任务中仍面临特征判别弱、领域适应低等瓶颈。例如，GPT-3在SST-2和AGNews数据集的准确率仅为76.0%和43.9%[2]，远不及小模型gMLP的94.8%[3]；ChatGPT在多个NER数据集的F_1得分低于90.00%[4]，而小规模的PIQN则达90.96%[5]。这表明LLMs的通用优势未能有效转化为特定任务的性能。

当前提升LLMs下游泛化能力的核心路径为指令微调与提示工程[6]。在指令调整方面，如Hyper-Tuning框架[7]、构建领域抽取指令集等方法，虽增强了特定场景执行力，但高度依赖专家级、高逻辑性的指令数据。在提示工程方面，研究者提出了提示编程、引入Python解释器增强推理[8]、Reflexion框架[9]及Self-refine技术[10]等，此类方法虽具备零参数调整优势，但响应质量极易受提示词影响，导致性能波动巨大，在特定领域难以适配[11]。

针对上述大模型特征提取转化能力不足的短板，本文结合RoBERTa[12]强大的特征提取能力与Li等[13]的大模型微调思想，提出基于增强全局特征提取的分类大模型框架。该框架引入旋转位置嵌入以提升特征关注力，在注意力层调整因果掩码释放全局特征，并使用SwiGLU进行增强。受BERT微调启发[14]，本文使用近百万样本结合低秩自适应（LoRA）技术优化损失。结果表明，该框架在情感分析和NER基准测试中均表现出显著的性能提升。

1 大语言模型基本结构

LLMs通过在大量文本数据上进行预训练，能够深入学习语言的结构和语义信息，在各种NLP任务中表现出卓越的性能，包括但不限于文本生成、文本分类、问答系统和机器翻译。

LLMs模型的基本结构如图1所示，其核心基于Transformer架构，它包括编码器（encoder）和解码器（decoder）两部分，但在LLMs中主要关注的是解码器部分，用于处理输入序列。LLMs模型通过堆叠多个Transformer解码器层来建立模型，每一层都进一步提升了模型对语言结构的理解能力。自注意力机制使得模型能够基于输入序列的不同部分计算出不同的注意力权重，动态地聚焦于信息量最丰富的部分。

由于Transformer架构中没有显式地处理序列顺序信息，所以LLMs在输入层中引入了位置编码（positional encoding），用于赋予每个词汇在输入序列中的位置信息，使得模型能够区分不同位置的词汇。而在输出层中引入softmax分类器，用于根据模型的输入生成下一个词的概率分布，使得模型能够被用于生成文本、语言理解和各种NLP任务。

LLMs模型的预训练采用无监督方式，在大规模文本数据上进行。预训练任务通常包括掩码语言模型（masked language model，MLM）和下一句预测（next sentence prediction，NSP）两种任务。在MLM任务中，模型学习根据上下文预测被随机掩盖的单词；而在NSP任务中，模型则需要判断两个句子是不是连续的文本；两种任务的结合有助于模型对语言结构和语义信息的理解。

预训练完成后，LLMs模型可以通过微调（Fine-tuning）的方式应用于特定的下游NLP任务。微调过程涉及在预训练模型基础上，使用较小的、特定任务的数据集进行再训练，以使模型参数更好地适应该任务。本文使用的LLaMA、ChatGLM和Qwen模型作为开源LLMs的典型代表，其基本结构也包

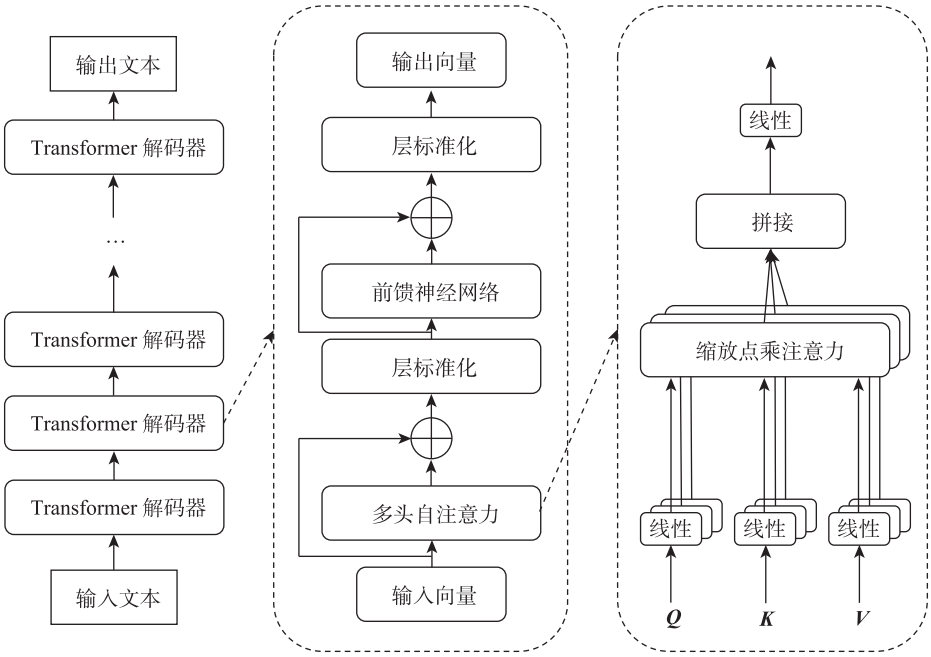

图1　LLMs模型结构

Fig. 1　Structure of LLMs model

括了多层堆叠的Transformer编码器、自注意力机制、前馈神经网络、位置编码等组件。这些组件相互作用，使得模型能够有效地处理序列数据，其捕捉复杂语言特征和细微语义信息的能力，使其成为当今最先进的自然语言处理模型之一。

2　增强全局特征提取的分类大模型框架

首先提出增强全局特征提取的分类大模型框架，其总体框架如图2所示，将框架应用于大模型中构建增强全局特征提取的分类大模型。然后对方法的原理进行详尽分析，接着介绍所提出模型各模块的方法函数，最后对本文使用的参数高效微调技术进行说明。

2.1　增强全局特征提取的分类大模型

提出了增强全局特征提取的分类大模型（FS-LLMs），其总体结构如图3所示。首先通过分词器将输入文本解析为离散的token序列$\boldsymbol{S}$，随后通过词嵌入将每个token映射为连续向量，最终形成向量序列$\boldsymbol{T}$。如式（1）、式（2）

$$\boldsymbol{S}=\mathrm{Tokenizer}(text) \tag{1}$$

$$\boldsymbol{T}=\mathrm{Embedding}(\boldsymbol{S}) \tag{2}$$

将式（2）得到的向量序列$\boldsymbol{T}$使用前置层归一化方法RMSNorm后传递进入模型的注意力模块，在自注意力层中使用RoPE编码将相对位置信息嵌入，并在注意力计算的过程中对这些向量进行旋转变换，以便模型更好地处理序列数据中的位置特征。模型通过对因果掩码矩阵的动态调整，实现了注意力模块的全局特征提取。再对模型提取出的序列分类潜在特征$\boldsymbol{H}$进行池化操作，获取用于序列分类的向量$\boldsymbol{h}$

$$\boldsymbol{h}=\mathrm{Pooling}(\boldsymbol{H}) \tag{3}$$

最后向量$\boldsymbol{h}$在进入全连接层前，再次使用了前置层归一化方法RMSNorm，这种方法通过减少计算量、降低梯度波动，使训练过程更加稳定。

此外，全连接层使用SwiGLU激活函数，增强了模型捕捉输入中有用信息的能力，提高了模型的非线性表达能力。最后通过全连接层后映射至标签空间并得到输出。在此基础上，基于输出和真实标签计算交叉熵损失，并通过LoRA[15]对FS-LLMs模型进行微调，仅对低秩矩阵进行训练，最大化正确标签的概率[16]。以上过程得到优化模型效果。

在微调过程中，本文通过使用旋转位置嵌入位置信息，提升模型对特征的关注能力。在注意力层调整因果掩码，释放全局特征，使模型得到基本的双向信息特征补充，让所有标记都可以相互依赖。再使用SwiGLU对特征进行增强，在微调的过程中运用低秩自适应（LoRA）优化损失，最后构建一个全局特征提取的分类大模型。随着全局注意力的恢复，

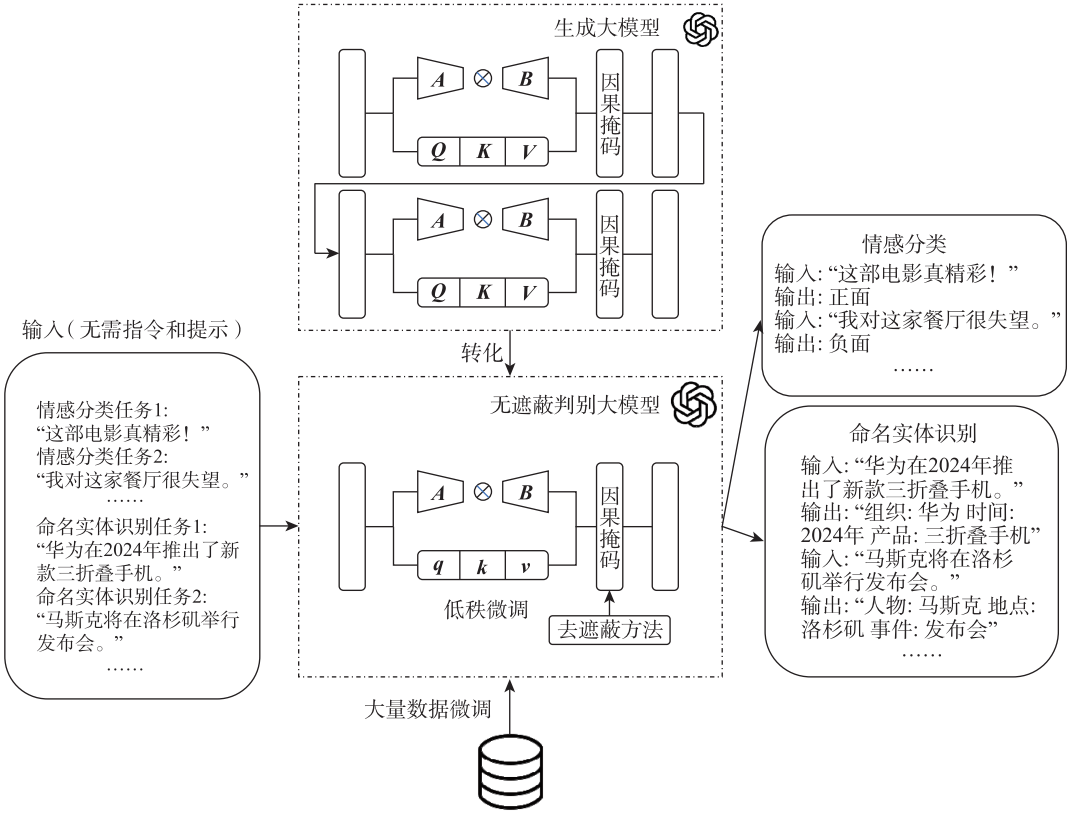

图2 全局特征提取大模型框架

Fig. 2 Large model framework for global feature extraction

对特征的提取能力增强，模型的准确率上升。

2.2 全局特征释放方法

大语言模型注意力层的因果掩码（causal masking）是为了在生成序列时确保当前位置的预测不依赖于该位置之后的任何信息。这在序列生成任务中是非常关键的，但在分类任务中因果掩码对模型起到了干扰作用，导致模型丢失了文本序列的关键特征。为增强模型对特征的提取能力，本文对LLMs中的模块进行调整。

因果掩码防止了信息的泄露，解码器只能依赖于生成文本中较早的位置。自注意力层的双向依赖提取被减少为单向，导致在标记级别存在关键信息的丢失。调整前的解码器模块中的因果掩码方程式$\boldsymbol{M}$如式（4）所示

$$\boldsymbol{M}=\begin{bmatrix}0 & -inf & -inf & \cdots & -inf & -inf & -inf\\ 0 & 0 & -inf & \cdots & -inf & -inf & -inf\\ 0 & 0 & 0 & \cdots & -inf & -inf & -inf\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & 0 & -inf & -inf\\ 0 & 0 & 0 & \cdots & 0 & 0 & -inf\\ 0 & 0 & 0 & \cdots & 0 & 0 & 0\end{bmatrix} \tag{4}$$

这里用$\boldsymbol{M}$表示因果掩码矩阵。其避免了任务中向前信息泄露，但也阻止了双向依赖关系的提取。x为模型的输入序列，y为模型的输出序列，则解码器的输出y_i可以通过以下方式计算

$$y_i=\mathrm{softmax}\left(\boldsymbol{W}\cdot x_i+\boldsymbol{U}\cdot y_{i-1}\right) \tag{5}$$

式中：$\boldsymbol{W}$为输入到输出的权重矩阵；$\boldsymbol{U}$为输出到输出的权重矩阵；x_i为输入序列的第i个元素；y_{i-1}为输出序列的前一个元素。由于因果掩码矩阵的存在，模型在计算y_i时只能考虑x_i以及y_{i-1}之前的信息，确保了生成的序列是符合预期的。

本文的实验研究表明，使用有因果掩码的标记表示进行预测，在标记分类任务中表现明显不佳。为了解决这个问题，本文对原始的因果掩码矩阵进行了调整，构建一个加权的掩码矩阵。其权值公式如下

$$w(d)=\mathrm{e}^{-\lambda d} \tag{6}$$

式中：d为不同token到当前token的距离；λ为常量。对于距离d=0（即当前token），$w(0)$=1，表示当前token的权重最大；对于距离d=1（即下一个token），$w(1)$会小于1，表示未来token的权重被衰减；

对于距离 d=2，w(2)更小，以此类推。在这个矩阵中，当前token可以访问自己和之前的token，而未来的token会根据它们与当前token的距离受到不同程度的遮掩，从而允许模型在自注意力机制中更灵活地捕捉序列中任意位置的信息。因果掩码调整后，自注意力层中的特征提取方式发生了重要的变化，专注于提取输入数据中的判别性特征，以便更好地进行分类任务。其核心操作公式为

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{dk}}+\boldsymbol{M}\right)\boldsymbol{V} \quad (7)$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分别为从输入表示通过线性变换得到的查询、键、值矩阵；$\boldsymbol{M}$为因果掩码矩阵。通常在自回归生成任务中，因果掩码矩阵$\boldsymbol{M}$确保模型在计算当前token的注意力时，只能看到前面的token，防止模型看到未来的信息。因果掩码矩阵调整后，模型的自注意力机制从单向特征提取转变为全局特征提取，这增强了模型对上下文信息的捕捉能力，在情感分析中模型可以充分利用整个句子的信息来预测情感类别；对于命名实体识别，模型能够同时关注某个词的前后文，进而更准确地识别出该词的实体类别。

在式(7)中可知，原始的输出 y_i 依赖于前一个输出 y_{i-1}，这个自回归结构保证了生成过程的因果性。当$\boldsymbol{M}$上三角矩阵调整后，模型不再限制 y_i 只能访问 y_{i-1}，整个输出序列都可以参与计算。于是，公式中的 y_i 不仅可以依赖前一个时间步 y_{i-1}，还可以依赖其他时间步的 $y_j(j\neq i)$，这相当于取消了序列生成中的顺序依赖。调整后解码器的输出 y_i 计算方式如式(8)

$$y_i=\mathrm{softmax}(\boldsymbol{W}\cdot x_I+\boldsymbol{U}\cdot Y) \quad (8)$$

式中：$Y=[y_1,y_2,\cdots,y_T]$是整个输出序列，而不再仅仅依赖 y_{i-1}。在这种情况下，模型可以基于整个序列中的信息进行计算，从而生成每一个 y_i。

而对于标签空间的映射，不再建模有利于生成任务的输入数据和标签的联合分布$p(x,y)$，用更适合分类任务的条件分布$p(x|y)$取代。潜在表示原来的映射不仅考虑标签，还考虑生成数据本身，通常包含隐变量或生成过程。现在通过从输入数据中提取特征来直接映射到标签空间，专注于区分不同的标签。

2.3　前置层归一化函数RMSNorm

如图3所示，FS-LLMs模型在多头自注意力层和全连接层两个模块前分别引入了前置层归一化方法RMSNorm，相较于通用归一化函数LN,不计算均值和标准差，只使用均方根，因此计算量更小，在处理大规模数据和长序列时更高效。并且由于RMSNorm的计算方式减少了梯度波动，能更好地应对梯度爆炸和梯度消失等问题，使模型的训练过程更加稳定,同时在多头自注意力层与全连接层之后进行残差连接。针对输入向量$\boldsymbol{a}$，RMSNorm 函数计算公式如式(9)、式(10)

$$\mathrm{RMS}(\boldsymbol{a})=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{a}_i^2 \quad (9)$$

$$\bar{\boldsymbol{a}}_i=\frac{\boldsymbol{a}_i}{\mathrm{RMS}(\boldsymbol{a})} \quad (10)$$

函数提供了输入向量的规范化，确保每一层中向量之间的交互是平稳且均衡的，减少数值不稳定带来的干扰。

2.4　全连接层函数SwiGLU

如图2所示，模型的全连接层采用SwiGLU的激活函数。相较于ReLU函数，SwiGLU的激活函数帮助网络更好地捕捉输入中的有用信息，提高网络的特征表达能力，使之能够灵活地针对输入数据建立非线性关系模型。其计算公式如式(11)~式(13)

$$\mathrm{FFN}(x,\boldsymbol{W},\boldsymbol{V},\boldsymbol{W}_2)=\mathrm{SwiGLU}(x,\boldsymbol{W},\boldsymbol{V})\boldsymbol{W}_2 \quad (11)$$

$$\mathrm{SwiGLU}(x,\boldsymbol{W},\boldsymbol{V})=\mathrm{Swish}_{\beta}(x\boldsymbol{W})\times x\boldsymbol{V} \quad (12)$$

$$\mathrm{Swish}_{\beta}(x)=x\sigma(\beta x) \quad (13)$$

式中：$\sigma(x)$为Sigmoid函数；当β趋近于0时，Swish函数趋近于线性函数$y=x$，当β趋近于无穷大时，Swish函数趋近于ReLU函数。

SwiGLU则通过门控机制和非线性激活，增强了特征向量的表达能力和交互的复杂度，使得模型能够处理更加复杂的模式。

2.5　位置编码嵌入方法RoPE

模型在使用分词器进行标记化处理时嵌入位置编码，采用旋转位置嵌入(rotary positional embeddings，RoPE)代替原有的绝对位置编码。RoPE借助了复数的思想，出发点是通过绝对位置编码的方式实现相对位置编码，其目标是通过式(14)的运算来给q，k添加绝对位置信息。经过上述操作后，q和k就带有位置m和n的绝对位置信息

$$\tilde{q}_m=f(q,m),\ \tilde{k}_n=f(k,n) \quad (14)$$

RoPE改进了不同位置向量之间的交互，使得模型不仅基于向量的内容进行交互，还能够捕捉它

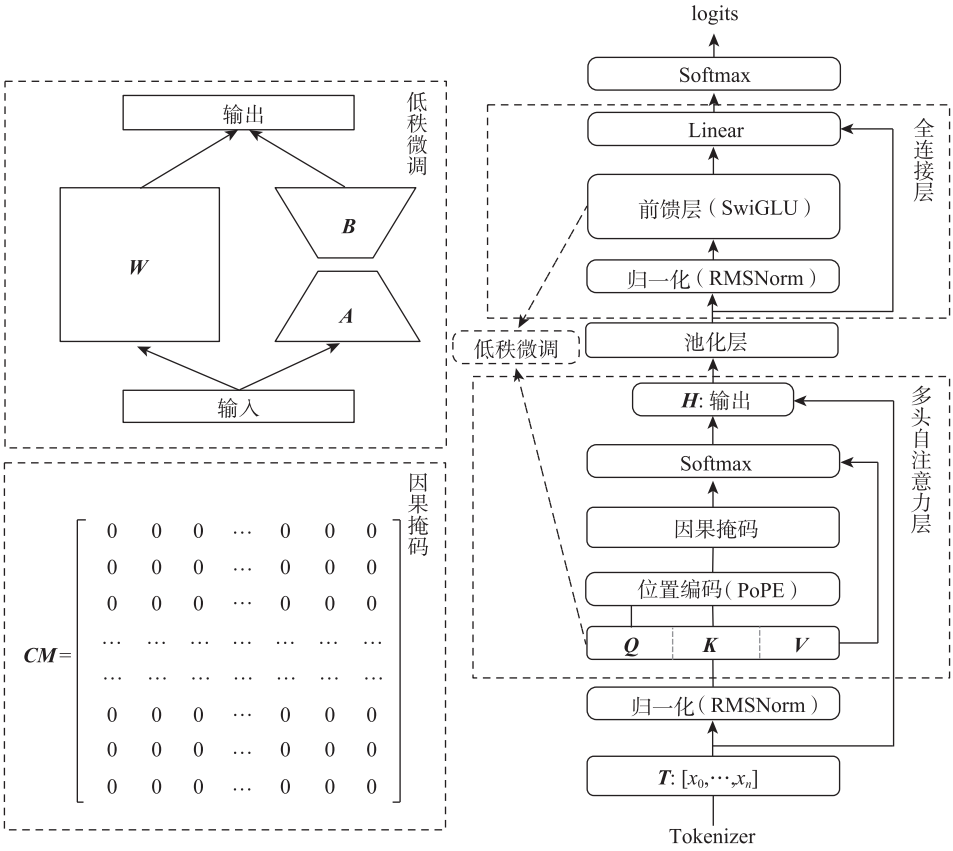

图3 全局特征提取大模型结构

Fig. 3 Global feature extraction large model structure

们的相对位置特征，特别是在长序列任务中表现优异。

3 实验与结果分析

3.1 实验任务数据和评价指标

3.1.1 情感分析数据集和评价指标

本研究在4个多领域英语数据集上实验，分别为一般情感分析的数据集SST-2、SST-5、新闻情感分析的AGNews和金融情感分析的Twitter-Fin。SST-2包含67 348个训练样本和8 729个测试样本，标签为积极和消极。SST-5包含8 500个训练样本和2 200个测试样本，标签为非常积极、积极、中性、消极和非常消极。AGNews包含120 000个训练样本和7 600个测试样本，标签为商业、科技、政治和体育。Twitter-Fin包含9 833个训练样本和2 369个测试样本，标签为看跌、中性和看涨。

评估情感分析模型的性能采用准确率评价指标，量化模型的表现；准确率表示模型正确分类的样本占总样本的比例，是最直观且易于理解的评价指标之一，其计算公式如式(15)

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

式中：TP为真正例(true positives)，表示模型正确地预测为正类的样本数；FP为假正例(false positives)，表示模型错误地将负类样本预测为正类的样本数；式(14)可以评估模型在正类预测中的准确性。

3.1.2 命名实体识别数据集和评价指标

本研究在OntoNotes和CoNLL2003[17]两个数据集上对命名实体识别任务进行实验，这两个数据集是常用的权威命名实体识别数据集。OntoNotes和CoNLL2003由于命名实体识别任务的复杂性，评估其性能的准确性和可靠性至关重要。评价指标F_1是比较常用的一种度量方法，综合考虑了精确率(Precision)和召回率(Recall)的信息，提供了一个平衡这两者的手段。

评价指标F_1是精确率和召回率的调和平均值，提供了一个单一的度量标准，评估模型的整体效

能，尤其是能权衡精确率和召回率，其计算方法为式（16）

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{16}$$

3.1.3　基线对比模型

为了评估FS-LLMs模型性能，本文选择了情感分析与命名实体识别任务中较经典的模型以及近期效果较好的模型进行对比。

GPT-3：基于Transformer架构，并通过无监督学习从大量互联网文本中进行训练，处理复杂的语境时表现出色，能够理解和分析带有情感的长文本

ChatGPT：能够处理各种形式的文本，包括非正式语言和口语，这使得它在非结构化数据中的命名实体识别表现良好。

BERT：使用双向Transformer进行编码，获得了更深层次的语言理解，能够提取文本中的关键情感特征。

RoBERTa：一种改进版的BERT模型，通过更大的训练数据、更长的训练时间和优化的训练策略，显著提升了自然语言处理任务的性能。

gMLP：gMLP模型是一种基于具有门控机制的多层感知机（MLP）的架构，旨在在多个语言和任务中达到甚至超过Transformers相同的表现。

AGN[15]：Adaptive Gate Network (AGN)方法能够建模文本中的复杂关系，例如单词之间的依赖关系，这有利于捕捉细腻的情感表达，对文本的特征提取更为准确。

GPT-3.5-Turbo[18]：能够提供快速的实体识别能力，并且保留了GPT-3的广泛知识和强大的语言理解能力，在NER任务中的表现依然强大。

PIQN[5]：Parallel Instance Query Network（PIQN）方法设置全局且可学习的实例查询，实例查询并非依赖外部知识构建，而是在训练过程中学习其不同的查询语义，以并行方式从句子中提取实体，在多个NER任务中表现优秀。

部分模型在Twitter-Fin、SST-5和CoNNL2003数据集上的分类结果缺失。这是由于相关文献中某些模型框架未在这些特定数据集上进行实验，导致缺乏相应的结果。本文引用了这些文献中的实验结果，为了保证实验结果与原始文献的一致性，本文未对缺失的部分进行填充。

3.2　实验环境和实验设置

本文沿用了文献[20]的部分实验设计，在此基础上，我们进一步扩展，提出了基于增强全局特征提取的分类大模型框架。将其应用于多种大模型，以探讨这些改动对性能的影响。通过实验结果对比，验证所提出方法的优越性和通用性。

本节实验针对文本分类和命名实体识别任务进行，验证本文提出的增强全局特征提取的分类大模型框架的有效性。实验中使用的大模型是最为先进的开源大模型LLaMA-2、Qwen-1.5和chatGLM-2版本，采用了标准的参数高效微调方法LoRA来微调模型。分别与零次和少次学习的大模型LLMs、指令微调的大模型LLMs和其他分类基线模型进行了比较。

实验使用PyTorch提供了灵活的模型开发和训练接口，并配备NVIDIA GTX 3090 GPU，能够支持较大模型和批量数据的训练。在大模型训练中批量大小设置为16，初始学习率为2e-4。并将LoRA主要参数lora_rank、lora_alpha和lora_dropout分别配置为8、32和0.1。批量大小设置为8，初始学习率为4e-5。在这种环境下利用PyTorch的自动微分和数据加载工具，结合GTX 3090的强大计算能力，进行高效的模型训练和调试。

3.3　实验结果

3.3.1　情感分析任务

在表1中，本文展示了不同模型在SST-2、AGNews、Twitter-Fin和SST-5数据集上的表现，其中包含了每个模型的精确率（Precision）和召回率（Recall）数据。

LLaMA-2-7B和LLaMA-2-13B在零样本设置下在SST-2和AGNews任务上，精确率仅有76.24%和37.06%。表明它们难以有效进行分类。GPT-3在零样本设置下的精度较低，在SST-2上仅为54.30%，但其在少样本设置下的表现大幅提升至93.40%。LLaMA-2-7B在指令微调后，性能大幅提高，尤其是在AGNews和SST-2任务上，精确率分别为52.40%和91.97%。这表明指令微调能够在任务特定的数据集上一定程度改善模型的分类能力。

本文提出的FSLLMs方法在多种任务和数据集上表现出色，尤其是在特定任务领域的应用中，精度提升显著。在SST-2、AGNews、Twitter-Fin和SST-5任务上，FSLLMs-LLaMA-2-7B在与RoBERTa-Large相比的表现中，分别提升了1.44个百分点、0.95个百分点0.47个百分点和1.18个百分点，这显

表1 多类别文本分类实验结果

Tab.1 Experimental results of multi-category text classification %

模型	SST-2		AGNews		Twitter-Fin		SST-5	
	精确率	召回率	精确率	召回率	精确率	召回率	精确率	召回率
LLaMA-2-7B（零样本）	76.24	73.36	37.06	36.67	23.29	21.65	38.82	36.91
LLaMA-2-13B（零样本）	69.14	67.23	59.82	58.25	27.96	27.93	37.14	36.13
LLaMA-2-7B（指令微调）	91.85	89.17	52.19	51.72	69.15	66.97	42.61	42.12
GPT-3-175B（零样本）	54.30	51.80	43.90	42.20	–	–	–	–
GPT-3-175B（少样本）	93.40	91.40	84.30	82.60	–	–	–	–
BERT-Base	92.78	90.78	94.51	91.51	88.19	86.44	55.07	54.09
BERT-Large	92.86	91.54	94.45	92.34	88.74	86.33	55.79	55.12
RoBERTa-Base	94.61	92.61	94.70	91.70	90.32	88.57	58.46	55.74
RoBERTa-Large	96.10	93.21	94.78	92.96	90.95	89.35	59.64	58.96
gMLP-Large	94.80	92.72	93.41	91.56	89.11	87.94	57.25	56.24
AGN	93.27	91.27	93.82	92.42	–	–	55.72	55.36
FSLLMs-ChatGLM-2-6B	96.10	93.47	94.95	93.97	90.23	88.64	59.87	59.01
FSLLMs-Qwen-1.5-7B	95.13	92.25	94.13	93.11	89.14	87.41	59.21	58.25
FSLLMs-LLaMA-2-7B	**97.54**	**94.36**	**95.73**	**94.44**	**91.42**	**89.79**	**60.82**	**59.79**
FSLLMs-LLaMA-2-13B	92.94	91.16	95.42	93.57	87.77	86.45	52.70	51.42

示出该方法在特定任务泛化上的优势，但是研究发现13B模型的性能显著下降，这是由于13B模型的参数更多，需要更多数据来有效调整权重。训练样本的不足，导致大模型过拟合。而7B模型因参数较少，可能更快收敛到局部最优解，13B模型因复杂度过高反而难以捕捉任务特定模式。本文模型FSLLMs-ChatGLM-2-6B和FSLLMs-Qwen-1.5-7B的性能低于微调后的RoBERTa-Large，同样是由于参数量远大于RoBERTa-Large，具有更高的容量，能够拟合更复杂的函数关系。但这也导致其方差更大，容易过拟合。从而需要充分且适配的训练样本来进行训练。

通过基于增强全局特征提取的分类大模型框架，本文提出的模型在多个任务上的性能明显优于现有的指令微调方法，尤其在需要特定领域知识的任务中显示出较为显著的提升。无论是在广泛的情感分析任务，还是在更加细化和领域特定的任务（如金融情感分析和新闻分类），FSLLMs都展现了其强大的适应能力，且改进无需复杂的提示工程或外部知识，进一步证明了其在实际应用中的潜力。这些结果为大规模语言模型在特定任务中的应用提供了有力的支持。

3.3.2 命名实体识别

表2展示了FSLLMs在CoNLL2003和OntoNotes V5数据集上的命名实体识别（NER）任务实验结果。与PIQN模型相比，FSLLMs-LLaMA-2-7B在F_1分数上分别提高了0.79%（CoNLL2003）和1.99%（OntoNotes V5）。这些结果表明，FSLLMs在NER任务中具有显著优势。

表2 命名实体识别任务F_1分数

Tab. 2 Named entity recognition task F_1 scores %

模型	CoNNL2003		OntoNotes V5	
	精确率	F_1	精确率	F_1
LLaMA-2-7B（零样本）	1.86	1.35	1.63	1.20
ChatGPT（零样本）	69.71	67.20	51.85	51.10
GPT-3.5-Turbo（零样本）	–	–	20.39	18.22
BERT-Base	92.91	92.40	88.90	88.88
RoBERTa-Base	92.64	92.13	91.13	91.55
PIQN	93.29	92.78	91.43	90.96
FSLLMs-ChatGLM-2B	93.19	92.27	93.57	91.56
FSLLMs-Qwen-1.5-7B	92.20	91.11	92.05	90.20
FSLLMs-LLaMA-2-7B	**94.08**	**93.51**	**94.66**	**92.77**
FSLLMs-LLaMA-2-13B	92.28	91.32	92.14	91.41

在零样本设置下，LLaMA-2-7B的表现较差，F_1分数分别为1.35%和1.20%，这表明LLaMA-2-7B在没有标注数据的情况下难以有效处理NER任务。相比之下，FSLLMs-LLaMA-2-7B模型在相同设置下的F_1分数分别为93.51%和92.77%，显著优于BERT-Base和RoBERTa-Base等PIQN传统模型。

但是在模型扩展到13B规模时，性能发生了下降，这是由于去除因果掩码后，模型需要重建双向注意力权重，这一过程对训练数据量敏感。13B模型因参数规模大，对数据需求更高。更大规模的模型在微调时可能面临梯度不稳定问题，尤其在训练后期出现过拟合反弹。

FSLLMs的优势可归因于其增强全局特征方法，该方法通过双向自注意机制有效恢复标记间的全局特征，弥补了因果掩码所带来的信息流限制。与传统模型相比，FSLLMs能够更好地捕捉句子中的全局语义信息，从而提高了在NER任务中的表现。FSLLMs-ChatGLM-2-6B和FSLLMs-Qwen-1.5-7B在性能上也显著优于BERT-Base和RoBERTa-Base等传统模型，这充分体现了本文方法的通用性与鲁棒性。

3.4　消融实验

3.4.1　情感分析消融实验

为了验证本文方法各个部分在分类任务中的有效性，本文设计在情感分析任务中移除每个组件并观察模型效果的变化。实验效果如表3和图4~图6所示。“w/o FS”表示不使用全局特征释放方法，在此限制下本文的FSLLMs模型的性能产生了一些变化。FSLLMs在SST-2和AGNews上的表现优于FSLLMs(w/o FS)，但在Twitter-Fin和SST-5上的性能反而不如移除组件的情况。本文认为从数据集大小中可以得出这些差异的原因。SST-2和AGNews明显比Twitter-Fin和SST-5具有更大的数据集规模，本文的模型在不同大小的数据集中可能表现不同。而在FSLLMs(w/o FS)中这一过程并不需要，因为它可以直接利用最后一个标记来进行分类任务。13B的模型在移除组件后性能不降反升，这是由于从FSLLMs中移除了全局特征释放方法，模型需要更多的训练样本来重建先前在预训练期间被隐藏的参数，所使用数据集的训练样本，不足以满足重建隐藏参数的需求。本文由此得出结论：FSLLMs(w/o FS)可以快速适应小规模数据集，而

表3　情感分析任务消融实验

Tab.3　Ablation study on sentiment analysis tasks　　%

模型	SST-2		AGNews		Twitter-Fin		SST-5	
	精确率	召回率	精确率	召回率	精确率	召回率	精确率	召回率
LLaMA-2-7B	76.26	73.36	37.39	36.67	23.40	21.65	39.05	36.91
LLaMA-2-13B	69.77	67.23	59.56	58.25	28.36	27.93	36.89	36.13
FSLLMs-LLaMA-2-7B（w/o LS）	77.84	74.40	38.50	39.95	25.67	22.67	42.26	38.84
FSLLMs-LLaMA-2-13B（w/o LS）	73.39	69.06	63.88	60.94	31.14	32.90	39.59	38.19
FSLLMs-LLaMA-2-7B（w/o FS）	96.71	93.67	95.51	94.17	**91.63**	**90.51**	**62.61**	**60.26**
FSLLMs-LLaMA-2-13B（w/o FS）	96.77	94.21	95.43	94.23	91.12	89.58	62.44	59.23
FSLLMs-LLaMA-2-7B	**97.54**	**94.36**	**95.73**	**94.44**	91.42	89.79	60.82	59.79
FSLLMs-LLaMA-2-13B	92.94	91.16	95.42	93.57	87.77	86.45	52.70	51.42

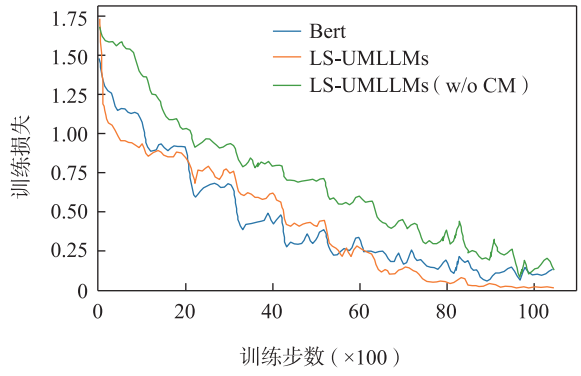

图4　SST-5训练损失比较图

Fig. 4　Training loss comparison chart for SST-5

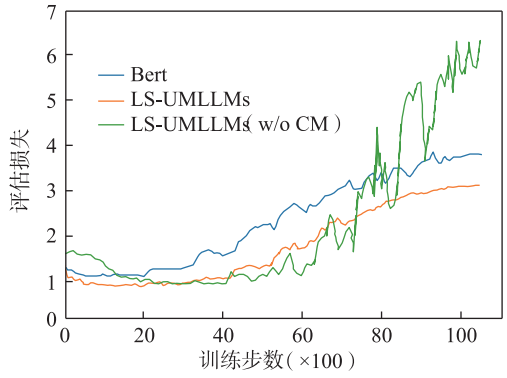

图5　SST-5评估损失比较图

Fig. 5　Assessment losses comparison chart for SST-5

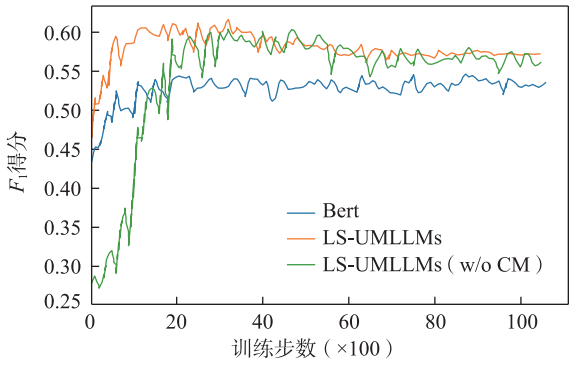

图6 SST-5 F_1得分比较图

Fig. 6 Scatter plot of SST-5 F_1 scores

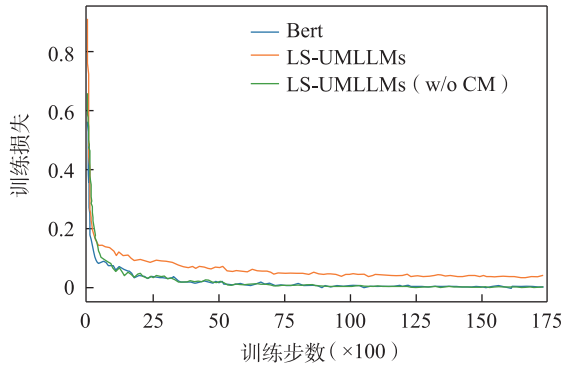

图7 CoNLL2003训练损失比较图

Fig. 7 Training loss comparison chart for CoNLL2003

FSLLMs在有充足的训练样本时可以取得更好的结果。“w/o LS”的模型表示不使用标签监督微调模型，由于缺乏标签信息的引导，模型无法有效地学习任务相关的特征，只能依赖自监督学习的信号，缺乏明确的分类或标注指引，容易导致错误的特征提取或注意力分配，从而极大地影响了模型的预测性能。

3.4.2 命名实体识别消融实验

为了验证本文模型各个部分在分类任务中的有效性，本文设计在命名实体识别任务中移除组件并观察模型效果的变化。实验效果如表4和图7~图9所示。“w/o FS”表示不使用全局特征释放方法，可以在实验结果中得出，FSLLMs的性能体现在NER任务中尤为突出。这种性能可以归因于全局特征在潜在表示中的关键作用。通过恢复双向

表4 命名实体识别实验消融实验

Tab.4 Ablation study on named entity recognition experiments %

模型	CoNNL2003		OntoNotes V5	
	精确率	F_1	精确率	F_1
LLaMA-2-7B	1.86	1.35	1.63	1.20
FSLLMs-LLaMA-2-7B（w/o LS）	16.15	16.36	15.50	20.68
FSLLMs-LLaMA-2-13B（w/o LS）	25.56	24.37	23.73	22.62
FSLLMs-LLaMA-2-7B（w/o FS）	77.07	74.42	78.67	77.77
FSLLMs-LLaMA-2-13B（w/o FS）	75.77	74.21	79.83	77.52
FSLLMs-LLaMA-2-7B	**94.08**	**93.51**	**94.66**	**92.77**
FSLLMs-LLaMA-2-13B	92.28	91.32	92.14	91.41

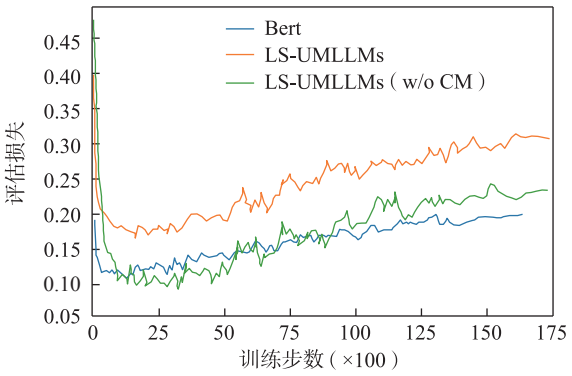

图8 CoNLL2003评估损失比较图

Fig. 8 Assessment losses comparison chart for CoNLL2003

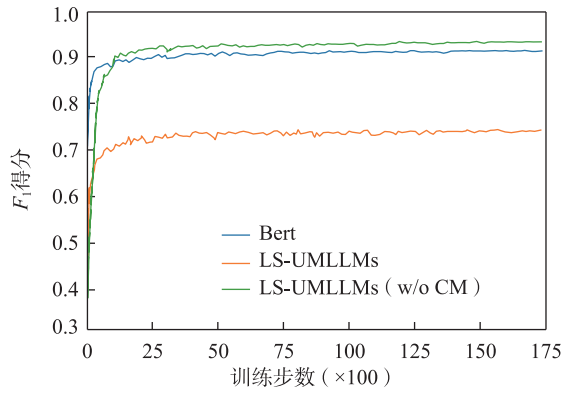

图9 CoNLL2003 F_1得分比较图

Fig. 9 Scatter plot of CoNLL2003 F_1 scores

自注意机制，FSLLMs能够有效捕捉和恢复这些全局特征，从而在处理NER任务时展现出更优越的性能。而这种特征在其他LLMs模型中则是缺失的。“w/o LS”的模型表示不使用标签监督微调模型，由于缺乏明确的实体类别标注，这使得模

型难以准确识别实体边界。此外,NER 任务依赖于上下文信息和序列依赖关系,而无监督模型无法充分利用这些特性。领域特定的知识也难以通过无监督学习自动获取,导致模型在特定领域表现不佳。

4 结论

本文标签监督方法,对大语言模型在分类任务上的性能进行研究,提出增强全局特征提取大模型框架,并进行实验,得到以下核心结论。

1）所提出框架在多个大模型的多种文本分类任务中展现了显著的性能提升,尤其是在情感分析和命名实体识别任务中。

2）通过对数据集和特征提取的深入分析,本文发现增强后的模型在处理需要全局上下文特征的任务时表现尤为优越。

3）目前,FSLLMs运用于多个开源大语言模型中,在通用的情感分析和NER任务数据集中验证了其效果,未来可以将其扩展到更多领域特定的任务（如医学文本分析、心理文本分析等）。跨领域的应用将为该方法的实用性提供更多的验证,并有助于其在工业界的落地。未来可以结合可解释人工智能（XAI）方法,深入分析模型决策过程,提升其透明度和可靠性。

参考文献：

[1] 朱君辉, 王梦焰, 杨尔弘, 等. 大模型生成回答与人类回答文本的语言特征比较研究[J]. 中文信息学报, 2024, 38(4): 17-27.
ZHU J H, WANG M Y, YANG E H, et al. A comparative study of language between artificial intelligence and human: a case study of ChatGPT[J]. Journal of Chinese Information Processing, 2024, 38(4): 17-27.

[2] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[3] LIU H, DAI Z, SO D, et al. Pay attention to mlps[J]. Advances in Neural Information Processing Systems, 2021, 34: 9204-9215.

[4] LI B, FANG G, YANG Y, et al. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness[J]. arXiv preprint arXiv, 2023: 2304.11633.

[5] SHEN Y L, WANG X B, TAN Z Q, et al. Parallel instance query network for named entity recognition[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland. Stroudsburg, PA, USA: ACL, 2022: 947-961.

[6] WANG Y Z, MISHRA S, ALIPOORMOLABASHI P, et al. Super-naturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates Kerrville: Association for Computational Linguistics 2022: 5085-5109.

[7] PHANG J, MAO Y, HE P, et al. Hypertuning: Toward adapting large language models without back-propagation[C]//International Conference on Machine Learning. PMLR, 2023: 27854-27875.

[8] CHEN W H, MA X G, WANG X Y, et al. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks[J]. arXiv preprint arXiv, 2022: 2211.12588.

[9] SHINN N, CASSANO F, BERMAN E, et al. Reflexion: Language agents with verbal reinforcement learning[J]. Advances in Neural Information Processing Systems, 2023, 36: 8634-8652.

[10] MADAAN A, TANDON N, GUPTA P, et al. Self-refine: Iterative refinement with self-feedback[J]. Advances in Neural Information Processing Systems, 2023, 36: 46534-46594.

[11] 曹义亲, 盛武平, 周会祥. 基于TF-IDF-MP算法的新闻关键词提取研究[J]. 华东交通大学学报, 2021, 38(1): 122-130.
CAO Y Q, SHENG W P, ZHOU H X. Research on news keyword extraction based on TF-IDF-MP algorithm[J]. Journal of East China Jiaotong University, 2021, 38(1): 122-130.

[12] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv, 2019: 1907.11692.

[13] LI Z, LI X, LIU Y, et al. Label supervised llama finetuning [J]. arXiv preprint arXiv, 2023: 2310.01208.

[14] DEVLIN J, CHANG M-W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv, 2018: 1810.04805 .

[15] HU E J, SHEN Y L, WALLIS P, et al. Lora: Low-rank adaptation of large language models[J]. Iclr, 2022, 1(2): 3.

[16] WEI J, WANG X Z, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.

[17] TJONG KIM SANG E F, DE MEULDER F. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition[C]//Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, Edmonton, Canada. Morristown, NJ, USA: ACL, 2003: 142-147.

[18] WEI X, CUI X Y, CHENG N, et al. Chatie: Zero-shot information extraction via chatting with chatgpt[J]. arXiv preprint arXiv, 2023: 2302.10205.

第一作者:陈可纬(2000—),男,硕士研究生,研究方向为自然语言处理、大语言模型。E-mail:1019375578@qq.com。

通信作者:刘建华(1967—),男,教授,博士,硕士生导师,研究方向为智能计算、机器学习。E-mail:jhliu@fjnu.edu.cn。

(责任编辑:李根)