

知识增强的中文金融大模型研究

程大伟^{1,2,3}, 贾仁军¹, 李江彤¹, 丁志军^{1,2,3}, 蒋昌俊^{1,2,3}

1. 同济大学计算机科学与技术学院, 上海 200090;
2. 上海人工智能实验室, 上海 200030;
3. 国家级网络金融安全协同创新中心, 上海 201804

摘要

金融行业长期以来面临海量市场数据与信息处理难题。当前, 大语言模型在通用文本理解任务上取得了显著进展, 但在专业性更强的中文金融领域还有较大的提升空间。针对当前大语言模型在处理专业领域文本任务上的不足, 提出基于知识增强的继续预训练和监督微调的两阶段训练方法, 并改进了训练数据的组织形式和训练范式, 从而提升模型在复杂金融场景下的性能。最后, 通过实验验证了提出的知识增强方法在大模型训练中的有效性。

关键词

知识增强; 大语言模型; 金融时序预测

中图分类号: TP391

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2025021

Research on knowledge-augmented Chinese financial large language model

CHENG Dawei^{1,2,3}, JIA Renjun¹, LI Jiangtong¹, DING Zhijun^{1,2,3}, JIANG Changjun^{1,2,3}

1. College of Computer Science and Technology, Tongji University, Shanghai 200090, China
2. Shanghai Artificial Intelligence Laboratory, Shanghai 200030, China
3. National Collaborative Innovation Center for Internet Financial Security, Shanghai 201804, China

Abstract

The financial industry has long faced challenges in processing vast amounts of market data and information. Currently, large language models have made significant progress in general text understanding tasks, but there is still considerable room for improvement in more specialized domains, such as Chinese finance. To address the limitations of current large language models in handling professional domain-specific text tasks, a two-stage training approach based on finance knowledge-enhanced continued pre-training and supervised fine-tuning is designed. This approach improves the organization of training data and the training paradigm, thereby enhancing the model's capabilities in complex financial scenarios. Finally, experiments have validated the effectiveness of the proposed knowledge-enhanced approach in large model training.

Key words

knowledge augmentation, large language model, financial time series forecasting

0 引言

从市场趋势预测到决策支持，金融文本的分析与理解在金融领域中至关重要。然而，金融文本通常包含行业专业术语、复杂语义以及动态变化的市场热点信息，这给分析和理解任务带来了巨大的挑战。随着信息化程度的不断提高，金融行业产生的数据量持续增长，如何高效分析和理解海量金融数据已成为金融行业面临的一大难题。

近年来，随着深度学习和自然语言处理技术的飞速发展，尤其是大规模预训练语言模型（如 GPT-3^[1]、BERT^[2]等）的兴起，大语言模型（简称大模型）被广泛应用于通用文本数据处理，也为金融领域的文本信息处理提供了新的手段。然而，尽管现有大模型在通用自然语言任务中表现出色，但在中文金融场景下依然面临显著挑战，例如金融专业知识不足、行业热点词汇以及网络隐喻理解困难等。因此，研究面向金融场景的知识增强大语言模型具有重要的理论价值和应用前景，有助于推动大语言模型更好地应用于市场情绪分析、风险管理、金融问答、自动化报告生成等场景。

本文提出了一种基于知识增强的中文金融大模型——CFGPT（Chinese financial generative pre-trained model），基于高质量中文金融语料，设计了两个步骤的知识增强嵌入学习方法，并通过实验验证了本文方法在金融任务上的有效性。

具体来说，本文首先构建了高质量中文金融数据集（Chinese financial dataset, CFData），包括预训练数据集和监督微调数据集。预训练数据集包含 5.84 亿篇文档和 1 410 亿个 token，包括公告、研究报

告、社交媒体内容、财经新闻文章以及维基百科等内容。监督微调数据集包含情感分析、金融事件检测、研究报告分析总结、金融问答、金融安全合规以及金融时序预测 6 种典型金融场景的数据，包括 150 万个指令对和 15 亿个 token。基于 InternLM2.5-7B 进行全量参数的继续预训练，得到预训练金融大语言模型 CFGPT-pt，然后使用指令微调数据集进行面向金融场景任务的知识增强嵌入训练，最后得到中文金融大模型 CFGPT，提升模型在金融场景的处理能力。

1 相关工作

近年来，预训练语言模型发展迅速，在学术界和产业界均取得了重要进展^[2-5]。例如，2022 年公开发布的 GPT-3.5 和随后发布的 GPT-4 等大模型在阅读理解、开放式问题解答以及代码生成等任务中展示出优越的性能^[6]。大模型被应用于政务、金融、医疗、能源等多个领域。例如，在医疗领域，医联推出的 MedGPT 大模型的参数规模达到了千亿级，预训练阶段使用了超过 20 亿条医学文本数据，致力于推动从疾病预防、诊断、治疗到康复的全流程智能化诊疗。在政务领域，上海徐汇区推出了政务服务垂类大模型应用，徐汇区政务服务大模型为窗口工作人员提供了全面的知识支持和智能辅助，同时为企业和公众提供了专业高效的咨询服务。在能源领域，百度的“文心一言”系列大模型，结合深度学习、知识图谱等技术，与国网智能电网研究院共同构建了千万级的电力文本样本库和电力行业知识图谱，显著提升了电力行业的专业分词准确率、敏感实体识别率等关键指标，提高了电力设备及电

网运营的自动化、智能化水平。

在金融领域，现有针对金融场景的大模型增强工作主要依靠金融数据进行预训练和知识嵌入。通常，用于大模型训练的金融数据集包括文本数据集和结构化数据集。其中文本数据集主要包括新闻、社交媒体文本、行业研究报告、上市公司公告。例如，Zhang 等^[7]分析博客和新闻文中包含的情绪来预测股票价格，从而生成市场中性的投资策略。同样，Ding 等^[8]和 Liang 等^[9]从新闻文章中抽取结构化的事件关系来辅助预测股票走势。Liu 等^[10]和 Cheng 等^[11]从原始文本数据中提取标准特征，利用文本辅助特征提升交易模型的训练效果。

随着大模型技术的发展，高质量数据对于金融模型的训练愈发重要。例如，BloombergGPT^[12]基于整合的高质量文本数据，提出了首个包含 500 亿参数的预训练金融大语言模型。PIXIU^[13]和 FinGPT^[14-15]等模型基于目标任务设计了指令微调数据

集，用于处理情感分析、命名实体识别和股票价格预测等任务。针对中文场景，BBT-Fin^[16]基于研究报告、社交媒体和金融新闻等数据训练了面向金融场景的预训练模型。百度发布的轩辕 2.0 针对特定任务进行微调训练，在通用和金融问题解答方面均取得了较好的效果^[17-18]。

针对金融场景的复杂性，通用大语言模型在专业领域知识、热点词汇理解等方面均存在明显的不足。针对金融场景的预训练能够有效提升大模型的领域学习能力，但是在知识性较强的金融任务（如合规与安全检查、金融计算和时序预测）中，仍然需要在知识增强的大模型上进行指令微调训练，使用难度更高且更符合目标任务要求的训练样本来提升模型在复杂任务中的表现。

2 中文金融数据处理

中文金融数据的处理流程如图 1 所示，

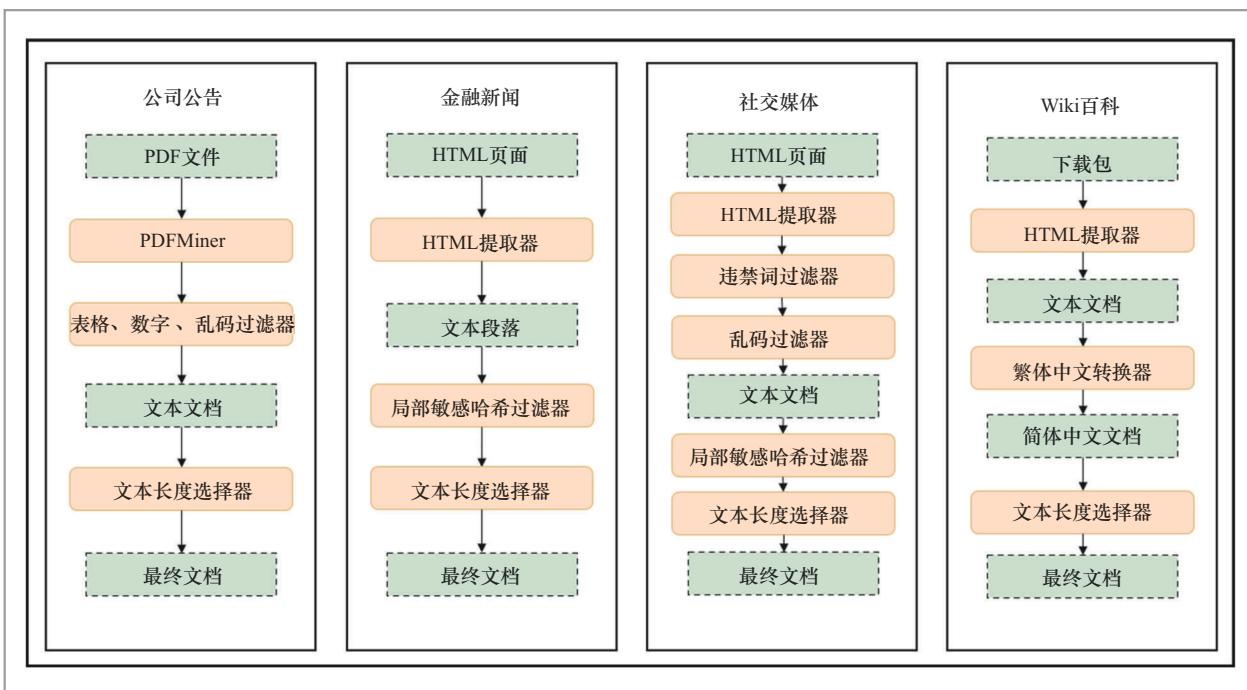


图 1 中文金融数据的处理流程

全参数预训练数据集和指令监督微调数据集的数据情况见表1。

2.1 全参数预训练数据处理

CFGPT全参数预训练过程的数据及其对应的处理过程如下。

公司招股说明书是向潜在投资者提供重要信息的法律文件。笔者从上海证券交易所和深圳证券交易所的网站抓取了2012—2022年的4.3万份公司招股书，清洗后得到3.91万份招股书，共879万个token。

公司公告是上市公司向公众提供信息的官方声明。笔者从上海证券交易所和深圳证券交易所的网站抓取了2012—2022年的938.91万份PDF格式的公司公告，采用与招股说明书相似的方法进行处理。公告数据集包含619万份企业公告，总计173亿个token。

研究报告是由投资机构发布，专注于宏观经济、行业、产业和股票的专业文件。

笔者从财经类网站抓取了2016—2022年的研究报告（约68.3万份），经过滤汇总后得到研究报告数据集，包含39.2万份研究报告，总计35.3亿个token。

财经新闻：从多个新闻网站抓取了2018—2022年的财经新闻（1.21亿条），使用局部敏感哈希（locality-sensitive hashing, LSH）算法处理重复数据和噪声数据后，得到8 240万篇新闻，总计263亿个token。LSH算法是一种近似最近邻搜索技术，特别适合在高维空间中处理大规模数据集。LSH算法的核心思想是将相似对象通过哈希函数映射到同一个哈希桶中，从而确保相似对象有较大的概率被聚集在一起，不相似的对象则被分散到不同的哈希桶中。由于收集到的原始文本数据规模较大，一般的去重、去噪方法的成本过高，因此使用LSH算法来处理该大规模数据集。本文首先对文本进行分词，根据全局的词表将各条帖子向量化，并使用余弦相似度度量文本的重复性，其中相似度超过

表1 全参数预训练数据集和指令监督微调数据集的数据情况

数据集	文件数/(10 ³ 个)	字符数/(10 ⁶ 个)	token数/(10 ⁶ 个)	token占比	存储/GB
预训练	583 978	206 665	140 609	100%	573
公司招股书	39	13 357	8 788	6.25%	37
公司公告	6 193	31 120	17 272	12.28%	86.3
公司研报	392	5 027	3 529	2.51%	13.9
财经新闻	82 438	37 262	26 297	18.7%	103.3
社交媒体	494 661	119 708	84 587	60.16%	332
维基百科	255	191	137	0.1%	0.5
监督微调	1 572	2 042	1 512	100%	5.66
情感分析	120	118	86	5.69%	0.33
事件检测	490	461	343	22.69%	1.28
分析总结	369	266	187	12.37%	0.74
报告摘要	369	1 014	765	50.6%	2.81
金融问答	12	8	6	0.4%	0.02
时间序列预测	212	175	125	8.27%	0.48

90%的帖子被视为重复语料内容。

社交媒体内容是金融市场情绪的重要体现。本文收集主流社交媒体平台2018—2022年提到行业、公司、产品、市场等内容的帖子，共计208.38亿条。由于社交媒体内容可能存在虚假性和偏见，采用禁用词典对社交媒体内容进行过滤。本文对开源项目 Sensitive-lexicon 中的敏感词集合进行采样并将其作为基础禁用词库，并且随着新词的发展动态更新禁用词典，然后采用 LSH 算法过滤重复数据和噪声数据，最终构建了一个包含4.95亿条帖子的社交媒体数据集，总计846亿个 token。

为了进一步提升模型的泛化能力，本文引入了中文维基百科通用数据用于模型预训练。清洗后得到的数据集包含25.5万个文档和1.37亿个 token。

2.2 指令微调数据处理

本文模型在情感分析、金融事件检测、研究报告分析总结、金融问答、金融安全合规以及金融时序预测6个任务的相关数据集上进行指令微调训练。

情感分析在金融市场分析领域至关重要。在构建有监督的微调数据集的过程中，对社交媒体和新闻研究报告等数据进行情感标注用于指令微调。针对社交媒体数据，随机选取6万条长度超过100字的帖子，通过 GPT-4 与人类专家协同的方式，对帖子进行情感标注（积极、消极、中性）。针对研究报告的内容，随机选取6万份2 000~3 000字的研究报告，提取内容和投资评级分别作为训练文本和训练目标，并将投资评价映射为“正面”“负面”和“中性”。

金融事件检测通过识别金融文档中的金融事件，为下游任务提供辅助决策支持。具体来说，笔者与金融行业研究院合作，

共同构建了一个包含98个类别、7个层级深度的事件分类体系，并通过专家标注的方式生成了49万份事件标注微调数据。

针对研究报告分析总结任务，本文筛选了研究报告数据集中具有总结摘要的36.9万份文档，将研究报告中的总结内容处理成监督学习的标注数据，以报告内容正文为输入、以结论和摘要为标注构建数据集。

针对金融问答任务，现有数据集大多为英文问答语料，本文将两个典型英文金融问答数据集 FinQA^[19] 和 ConvFinQA^[20] 翻译为中文语料对，进行金融问答数据的构建。

针对金融安全合规任务，设计了一系列潜在的恶意请求任务（如金融欺诈、洗钱、账户隐私侵犯等非法请求），以提升模型在安全合规方面的识别能力。面对非法请求时，模型可能被误导而产生非法响应。设置恶意请求任务的目的是让模型学会拒答用户的不合理请求，从而提升安全合规能力。

针对金融时序预测任务，本文结合热门的社交媒体帖子、财经新闻和历史股价走势，构建了股票预测数据集^[21]。具体来说，本文首先选取了沪深300指数中每只股票2021—2023年每个交易日浏览频率较高的财经新闻和帖子（阅读量大于50或点赞数大于10）以及过去5天的历史收盘价作为语料，使用次日的收盘价变化率作为标注，将任务转化为一个时序预测的分类问题，以预测次日股票收盘价的走向。

3 模型预训练与微调学习

本文模型训练以 InternLM2.5-7B 为基座模型。为了进一步提高模型在金融领

域内的理解能力,本文模型设计了两阶段的训练学习过程,包括全参数继续预训练和有监督微调学习,训练过程如图2所示。

3.1 全参数继续预训练

在全参数预训练过程中,本文采用标准的从左到右因果语言建模目标对模型进行微调。因果语言模型在生成文本时遵循一个因果顺序:生成当前词时只依赖之前的词,不依赖之后的词。双向模型在处理一个词时考虑了该词将来和过去的上下文,与双向模型(如BERT)不同,在因果语言模型中,给定词序列 $\mathbf{x}=(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t)$,模型会按顺序预测每一个词的出现概率 $p(\mathbf{x})$ 。

$$p(\mathbf{x}) = \prod_{t=1}^T p(\mathbf{x}_t | \mathbf{x}_{[1:t-1]}) \quad (1)$$

其中, T 是序列的长度, $\mathbf{x}_{[1:t-1]}$ 是第 t 个词之前的所有词组成的序列。通过条件概率

的乘积计算整个序列的概率,是GPT这类自回归模型的核心。

因果模型使用逐词更新的方式预测下一个词的向量表示,即根据历史序列的输入特征和当前位置之前的预测结果,预测当前位置之后的词向量表征。该自回归特性保证了模型在生成文本时具有自然的因果关系。而在因果语言模型逐词预测的训练过程中,一般使用交叉熵损失度量预测结果和真实特征之间的差异,将其作为参数训练的目标优化项。

$$L = -\frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \sum_{t=1}^T \log p(\mathbf{x}_t | \mathbf{x}_{[1:t-1]}) \quad (2)$$

其中, $\mathcal{D}$ 表示用于计算的样本序列集合(即数据集),包含若干个训练样本 $\mathbf{x}$, L 是面向数据集每个样本的平均交叉熵损失。

3.2 预训练改进优化

继续预训练过程存在灾难性遗忘问题,

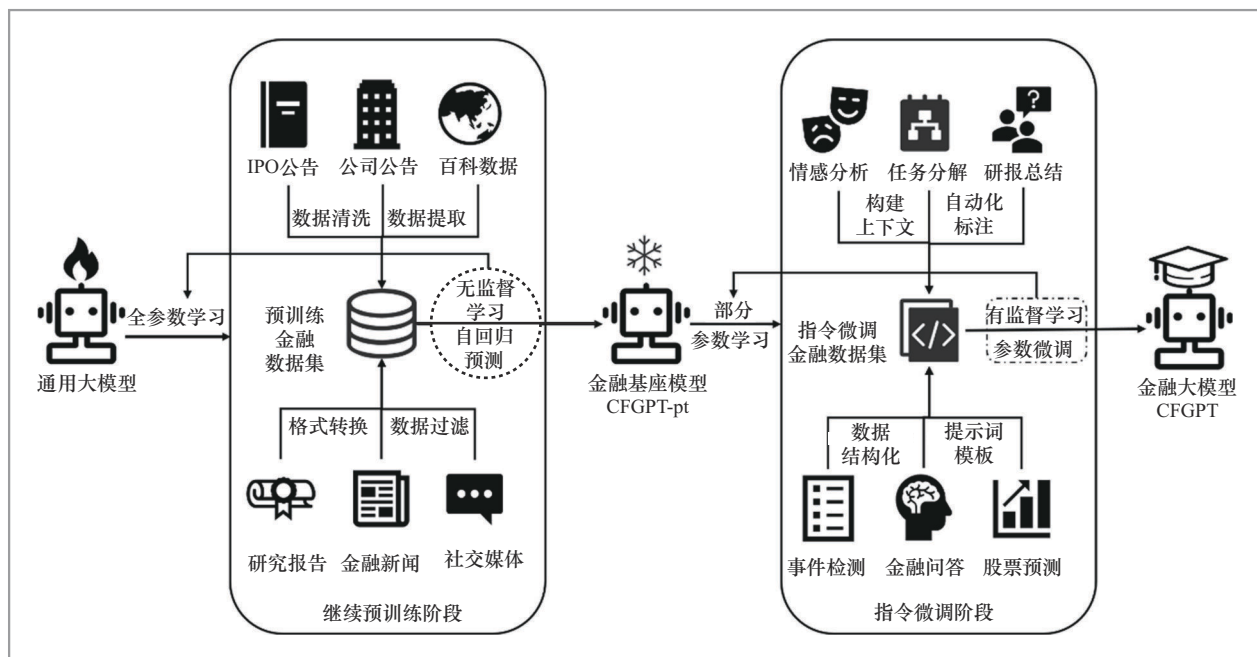


图2 知识增强的中文金融大模型训练过程

即模型在持续学习的过程中遇到新任务时遗忘了此前预训练的通用知识。本文通过调整语料顺序、平衡数据配比、动态调整学习率等方法优化金融知识增强的大模型训练流程,以避免灾难性遗忘带来的影响。本文选用 InternLM2.5-7B 作为基座模型,该模型主要使用英文文本作为预训练数据,处理英文文本的能力显著优于中文文本。因此在金融专业知识增强大语言模型的过程中,为了避免英文处理能力的退化和通用知识的遗忘,本文补充了一部分英文文本和通用知识作为训练数据。具体而言,本文构建了大小为 41.3 GB 的通用中英文对照文本数据集,包括 CLUECorpus^[22]、UNData、ParaCrawl Corpus^[23]、Wiki 以及 News-Crawl 等多个开源数据集和公共数据库。本文在训练阶段调整通用文本和中文对比数据的比例,提升了模型应对灾难性遗忘的能力。

此外,在不同数据集的比例优化方面,模型被期望能够在每一个数据集内部以及各个数据集之间都保持稳定的学习效果。当训练数据的顺序不合理时,损失函数可

能会在不同时间步发生剧烈波动,这不利于模型的知识积累。因此,本文根据文本质量调整数据输入顺序,通过评估文本语料的连续文本长度、特殊字符占比以及数据清洗过程中的反馈信息,对文本语料进行质量评级和段落划分,并采用“先易后难”的策略进行模型训练。即在每个梯度累积下降更新的时间步内,首先使用高质量的文本数据进行训练,随着训练的推进,不断增加较低质量的文本占比。同时,为了提升模型的泛化能力,随机调整每个时间步的训练数据顺序,CFGPT 预训练损失曲线如图 3 所示。

在训练参数动态调整优化方面,本文采用预热机制以及权重衰减机制,在训练过程中不断调整学习率和优化器参数来提升训练过程的稳定性。具体来说,采用权重衰减优化器和余弦衰减学习率调度器来防止低质量文本数据对模型权重更新的负面影响。在模型训练的初期,本文采用线性预热策略逐步提高学习率,以防止初始阶段权重剧烈变化,进而影响后续知识增强的学习过程。由图 3 可知,优化后预训

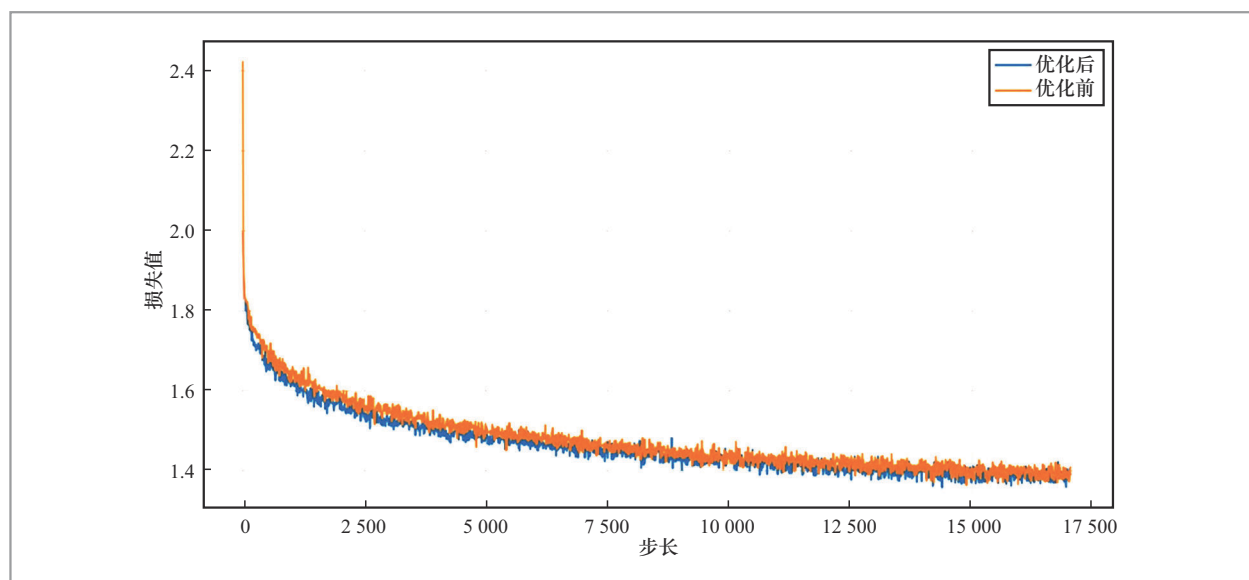


图 3 CFGPT 预训练损失曲线

训练损失曲线始终低于优化前的损失曲线，这验证了改进策略的有效性。

3.3 继续预训练参数设置

为了提升GPU的利用率，本文将训练序列按照每段1 024个token进行切分，使用“<EOS>”作为文档分隔符，然后将所有向量化的训练文档连接起来，生成用于训练的实例。在继续预训练过程中，本文使用AdamW^[24]优化器， β_1 、 β_2 和权重衰减分别为0.9、0.95和 10^{-5} 。批大小设为512，学习率设为 1×10^{-5} ，使用余弦衰减学习率调度器，前1 000步采用线性预热机制。在8块A100 80 GB GPU上进行继续预训练，得到CFGPT-pt-7B作为金融基座大模型。再将该基座大模型作为第二阶段监督微调任务的基础模型，通过在多个金融场景任务相关数据上的微调训练，提升CFGPT模型处理复杂金融任务的能力。

3.4 监督微调训练范式

在监督微调阶段，本文在6个典型金融任务上使用LoRA^[25]方法对模型进行高效参数微调。LoRA是一种高效的参数微调训练方法，在微调过程中只需要更新低秩权重矩阵的参数，通过矩阵分解的方式由低秩矩阵反向更新原有大模型的参数，能够显著提高指令微调训练过程的效率，降低显存需求。

具体来说，基于金融领域典型的6个任务场景，本文设计了若干提示模板来自动化标注多轮对话指令，以此构建指令微调数据集。不同训练任务的输出目标不同，例如分类任务一般给出分类结果的概率分布，而回归任务一般给出预测的未来序列

结果，因此针对不同训练任务设计了相应的损失函数来优化模型。具体来说，情感分析和安全合规可归为分类任务，输出目标是固定范围的分类选项（如情感倾向、是否拒答）。而事件识别、文本总结、金融问答等回归任务的目标是生成符合要求的目标序列。本文还将金融时序预测任务转化为多分类任务，从预测未来序列改为预测未来走势，这样更加贴合任务目标，并简化了训练优化目标。

在多分类任务中，本文采用多分类交叉熵。概率模型输出为一组概率向量 $\hat{\mathbf{q}}$ ，真实标签为一个独热向量 $\mathbf{y}$ ，则多分类交叉熵损失为：

$$L_{\text{CCE}} = - \sum_{i=1}^K \mathbf{y}_i \log(\hat{\mathbf{q}}_i) \quad (3)$$

其中， K 是类别数量， $\mathbf{y}_i$ 是样本在第 i 类上的真实值。 $\hat{\mathbf{q}}_i$ 是样本被预测为第 i 类的概率，满足 $\sum_{i=1}^K \hat{\mathbf{q}}_i = 1$ 。式（3）可简化为：

$$L_{\text{CCE}} = -\log(\hat{\mathbf{q}}_c) \quad (4)$$

其中， c 为样本的真实类别。

在回归任务中，本文使用均方误差损失。对于长度为 N 的真实值 $\mathbf{y}$ ，模型预测值为 $\hat{\mathbf{y}}$ ，则均方误差损失为：

$$L_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^N (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2 \quad (5)$$

3.5 监督微调训练改进优化

现有主流有监督微调的方法使用系统指令、用户提问和回答三段范式的指令。其中系统指令包括身份设定、回答格式要求、思维链方法等，用户提问包括多项选择题、分析题、思考题等，回答包括分类判断、文本生成和填空补齐等。常用的指令微调数据集多为选择题和填空题，评价指标相对客观，标签数据容易清洗检查。

但是上述单一形式的三段式指令在面对金融等复杂场景时，难以提高模型在专业领域的能力。例如，研究报告解读、公告总结等任务往往依赖模型对长文本的处理能力。同时，金融领域存在大量主观性的文本分析和解读任务，特别是在安全合规方面，模型需要识别带有攻击性的恶意提问，并给出明确的拒答回复。

本文从3个方面对指令微调数据的构造过程进行优化和改进。首先，对指令微调数据进行分类识别，将类似的问答语料合并为一次多轮对话的上下文，提升模型在上下文长文本处理过程中的能力。然后，利用GPT4等外部工具对已处理的指令提示词进行多样化扩展，以提高模型在开放场景的泛化能力。最后，基于真实金融业务，与行业专家合作设计领域指令数据，例如股票预测任务中的时序标记以及安全合规的指令模板等。

具体来说，多轮对话数据集的构造可以分为两种方式：聚合相似问答产生相关问题和从同一语料派生不同问题。第一种方式对已有指令进行分类，通过相似度匹配或向量嵌入聚类，将相似度高的指令数据进行组合，这种方式适用于小型的指令微调数据集。第二种方式使用外部工具从预训练语料中自动寻找问题和对应的解答。本文采用思维链引导模型，根据给定上下文语料提出问题，推导需要的知识并提出新的问题，清理整合后形成统一语境下的顺序上下文指令数据。这个过程获取了闭源大模型在解答问题时形成的思维，且生成的指令保持了上下文顺序，有利于大模型进行模仿学习和逻辑分析。

针对各项已经设置好的提示词模板，通过外部专家和工具对其进行多样化扩展。具体来说，给定模型提示词输入，提示模型输出若干个经过变形的提示词模板，增

加了指令微调数据的多样性。

在专家指令方面，针对主观问答指令，将问题转化为分类任务（如情感分析和股票预测任务），从而更高效地训练模型。而在主观文本总结分析任务上，基于摘要元数据、闭源模型标注与专家标注，形成文本分析任务的参考答案，为模型提供高质量的训练样本。损失函数从不同维度刻画模型输出与样本标签的相似性，引导模型进行学习。上述过程实现了对实际金融场景任务的知识增强训练，提升了模型的效果。

3.6 监督微调训练参数设置

在具体训练过程中，指令微调训练的参数设置如下。用于模型监督微调训练的样本的最大输入长度为2 048个token，每个训练批次的实例数量为64~512，具体数目由每批序列的长度决定。使用AdamW优化器， β_1 、 β_2 和权重衰减分别为0.9、0.95和 10^{-5} 。学习率为 2×10^{-4} ，使用余弦衰减学习率调度器，训练的前500步使用线性热身机制。在LoRA训练过程中，矩阵的秩设为64，缩放系数Alpha设置为16，线性函数的dropout设为0.05。监督微调训练同样在8张A100 80 GB GPU上执行，监督微调后得到CFGPT。

4 实验验证

本节评测CFGPT在金融专业知识、金融时序预测以及金融安全合规等3个任务上的表现，即进行了FinEval、沪深300成分股预测以及CFBenchmark安全合规3项测试。

4.1 金融专业知识评估

为评估CFGPT处理中文金融文本的能力，与其他10个代表性的模型进行对比，所选模型的规模为6~175 B，包括ChatGPT、Qwen、Baichuan、ChatGLM、InternLM等模型。本文使用FinEval^[26]作为金融知识能力的评估标准，综合评测模型在金融领域理论方面的表现，FinEval测评实验结果见表2。

由表2可知，在金融、经济以及证书考试3项实验结果中，GPT-4表现最好，CFGPT在所有专业金融知识评测上领先于其他开源的基线大模型，且在会计学测试上非常接近GPT-4的得分。在同级别模型中，CFGPT以7 B的参数量在金融学、经济学、会计学、证书考试多个指标中均取得最高分数，甚至与更大参数量规模的模型相比，仍然具有一定的优势，这表明知识增强在金融大模型训练过程中是有效的。

表2 FinEval测评实验结果

模型	参数量	金融学	经济学	会计学	证书考试	平均得分
GPT-4	—	71.0	74.5	59.3	70.4	68.6
ChatGPT	175 B	59.3	61.6	45.2	55.1	55.0
InternLM	7 B	49.0	49.2	40.5	49.4	47.1
ChatGLM2	6 B	46.5	46.4	44.5	51.5	47.4
Qwen-Chat	7 B	51.5	52.1	44.5	53.6	50.5
Qwen	7 B	54.5	54.4	50.3	55.8	53.8
Baichuan-Chat	7 B	44.9	41.5	34.9	45.6	42.0
Baichuan	13 B	51.6	51.1	41.7	52.8	49.4
InternLM2-Chat	7 B	54.2	54.0	43.5	55.4	51.9
InternLM2	20 B	57.3	58.9	47.4	58.6	55.5
CFGPT	7 B	62.6	63.9	58.9	66.0	62.9

4.2 金融时序预测

在金融时序预测任务的实验中，使用2021—2023年的沪深300指数成分股的价格数据作为训练与预测目标，数据经过处理后填入统一的提示词模板。模型根据给定的股票历史数据以及相关金融新闻来预测未来一天的走势判断并提供持仓建议，以此形成预测投资组合，按照每日收盘价格计算收益。本节对比了CFGPT与ChatGLM、Baichuan、Qwen、InternLM等主流开源大模型在时序预测指标上的效果。使用的评测指标包括年化收益率、夏普比率、卡玛比率、最大回撤、年化波动。夏普比率用以衡量投资组合相对于无风险利率的超额回报与其风险（波动性）之间的关系。最大回撤指投资组合在特定时间期间内从峰值到谷底的最大跌幅。卡玛比率等于年化回报除以最大回撤，用于衡量投资组合单位风险下的获利能力。年化波动率为年化的投资回报率标准差，用于衡量回报率的离散程度。其中，年化波动率、夏普比率、卡玛比率指标越高，代表模型获取超额收益的能力越强；最大回撤、年化波动指标越低，代表模型风险控制能力越强。

沪深300股市时序预测结果见表3，箭头向上表示该列数值越大，对应模型的投资回报越高，箭头向下表示该列数值越小，对应模型的风险控制水平越好。

由表3可知，基础的通用基座模型InternLM2.5在初始的预测任务上的表现并不理想。经过本文设计的第一个阶段的模型预训练后，CFGPT-pt在年化收益率、夏普比率和卡玛比率3项指标上均有明显的提升，而最大回撤和年化波动率两项指标与基座模型相近。通过进一步的指令微调训练，CFGPT的所有指标相比基座模型InternLM2.5均有显著提升，这验证

表3 沪深300股市时序预测结果

模型	参数量/B	年化收益率↑	夏普比率↑	卡玛比率↑	最大回撤↓	年化波动率↓
ChatGLM3	6	-0.0864	-0.6549	-0.0258	3.3481	0.1257
Baichuan2	7	0.0074	0.1218	0.0077	1.1931	0.1309
Qwen1.5-chat	7	0.0797	0.6392	0.1105	0.7219	0.1343
Qwen2-Instruct	7	0.1167	0.9372	0.2297	0.5412	0.1348
InternLM2	7	0.0069	0.1181	0.0056	1.2249	0.1313
InternLM2.5	7	0.0722	0.591	0.1176	0.6137	0.1329
CFGPT-pt	7	0.0947	0.7124	0.1539	0.6152	0.1409
CFGPT	7	0.1243	0.9517	0.2261	0.5163	0.1242

了本文提出的两阶段训练方法的有效性。与同等规模的基座模型相比，CFGPT在年化收益率和夏普比率上具有明显优势，只在卡玛比率上与Qwen2-Instruct存在极小差距。在最大回撤、年化波动率两项风险指标上，CFGPT也达到了同参数量下的最优效果，显示出经过指令学习后模型在风险控制方面的优势。

4.3 金融合规与安全评估

为了评估大模型在识别金融违规行为、隐私保护等金融合规与安全方面的能力，本文基于CFBenchmark^[27]提出的金融安全合规评估方法进行实验分析，评测结果见表4。

安全合规任务通过提示词诱导模型输出非法信息，该任务的得分为模型面对非法请求时拒答概率乘以100，表示模型在100次非法请求中的平均拒答次数。实验对比了GPT、Ernie bot（文心一言）等闭源大模型以及ChatGLM、Baichuan、Qwen等热门开源大模型。由表4可知，当前模型在金融合规与安全领域的整体表现较弱，参数规模的变化并未显著提升任

表4 金融合规与安全评测结果

类型	模型	参数量/B	得分
闭源	GPT-3.5	-	35.8
	GPT-4	-	39.2
	ERNIE-Bot-3.5	-	36.7
	ERNIE-Bot-4	-	37.3
开源	ChatGLM3	6	32.0
	Baichuan2	7	28.7
	Qwen1.5-chat	7	30.0
	Qwen2-Instruct	7	19.3
	InternLM2	7	39.5
	InternLM2.5	7	35.4
	CFGPT	7	47.3

务得分，这表明大多数模型忽视了安全与合规方面的能力建设。同时，主流大模型的版本迭代也未能展示出对金融合规与安全的优化，部分模型在版本升级后的安全能力甚至有所下降。与基准模型相比，CFGPT具有明显优势，这体现了在安全与合规方面进行知识增强学习的有效性。

5 结束语

本文开展了知识增强的中文金融大模型研究，提出了全参数继续预训练和金融场景任务指令微调的两阶段训练方法，显著提升了大模型在金融垂直领域的应用能力。本文详细介绍了高质量金融数据集的构建与处理方法、全参数预训练过程及其优化方法，以及面向6个真实金融场景的任务监督微调数据集的构建、处理和模型训练优化过程；最后，通过金融专业知识评估、时序预测、合规与安全评估实验验证了本文方法的有效性。本文提出的数据组织形式和模型训练范式为金融行业提供了更精准、高效的智能化解决方案，同时为其他垂直领域的大模型研究提供了参考。

参考文献：

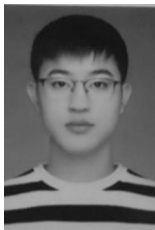
- [1] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2020: 1877–1901.
- [2] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2019: 4171–4186.
- [3] FEDUS W, ZOPH B, SHAZEER N. Switch Transformers: scaling to trillion parameter models with simple and efficient sparsity[J]. Journal of Machine Learning Research, 2022, 23: 1–39.
- [4] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020, 21(1): 5485–5551.
- [5] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2022: 27730–27744.
- [6] DU Z X, QIAN, Y J, LIU X, et al. GLM: general language model pretraining with autoregressive blank infilling[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2022: 320–335.
- [7] ZHANG W B, SKIENA S. Trading strategies to exploit blog and news sentiment[J]. Proceedings of the International AAAI Conference on Web and Social Media, 2010, 4(1): 375–378.
- [8] DING X, ZHANG Y, LIU T, et al. Using structured events to predict stock price movement: an empirical investigation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: ACL, 2014: 1415–1425.
- [9] LIANG X, CHENG D W, YANG F Z, et al. F-HMTC: detecting financial events for investment decisions based on neural hierarchical multi-label text classification[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann, 2021: 4490–4496.

- [10] LIU X Y, YANG H Y, GAO J C, et al. FinRL: deep reinforcement learning framework to automate trading in quantitative finance[EB]. arXiv preprint, 2011, arXiv: 2111.09395.
- [11] CHENG D W, YANG F Z, XIANG S, et al. Financial time series forecasting with multi-modality graph neural network[J]. Pattern Recognition, 2022, 121: 108218.
- [12] WU S J, IRSOY O, LU S, et al. BloombergGPT: a large language model for finance[EB]. arXiv preprint, 2023, arXiv: 2303.17564.
- [13] XIE Q Q, HAN W G, ZHANG X, et al. PIXIU: a large language model, instruction data and evaluation benchmark for finance[EB]. arXiv preprint, 2023, arXiv: 2306.05443.
- [14] YANG H Y, LIU X Y, DAN WANG C. FinGPT: open-source financial large language models[EB]. arXiv preprint, 2023, arXiv: 2306.06031.
- [15] ZHANG B Y, YANG H Y, LIU X Y. Instruct-FinGPT: financial sentiment analysis by instruction tuning of general-purpose large language models[EB]. arXiv preprint, 2023, arXiv: 2306.12659.
- [16] LU D K, WU H K, LIANG J Q, et al. BBT-Fin: comprehensive construction of Chinese financial domain pre-trained language model, corpus and benchmark [EB]. arXiv preprint, 2023, arXiv: 2302.09432.
- [17] ZHANG X Y, YANG Q, XU D L. Xuan-Yuan 2.0: a large Chinese financial chat model with hundreds of billions parameters[EB]. arXiv preprint, 2023, arXiv: 2305.12002.
- [18] ZHANG X Y, YANG Q. Self-QA: unsupervised knowledge guided language model alignment[EB]. arXiv preprint, 2023, arXiv: 2305.11952.
- [19] CHEN Z Y, CHEN W H, SMILEY C, et al. FinQA: a dataset of numerical reasoning over financial data[EB]. arXiv preprint, 2021, arXiv: 2109.00122.
- [20] CHEN Z Y, LI S Y, SMILEY C, et al. ConvFinQA: exploring the chain of numerical reasoning in conversational finance question answering[EB]. arXiv preprint, 2022, arXiv: 2210.03849.
- [21] SOUN Y, YOO J, CHO M, et al. Accurate stock movement prediction with self-supervised learning from sparse noisy tweets[C]//Proceedings of the 2022 IEEE International Conference on Big Data (Big Data). Piscataway: IEEE Press, 2022: 1691–1700.
- [22] XU L, ZHANG X W, DONG Q Q. CLUECorpus2020: a large-scale Chinese corpus for pre-training language model[EB]. arXiv preprint, 2020, arXiv: 2003.01355.
- [23] BAÑÓN M, CHEN P Z, HADDOW B, et al. ParaCrawl: web-scale acquisition of parallel corpora[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 4555–4567.
- [24] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[EB]. arXiv preprint, 2017, arXiv: 1711.05101.
- [25] HU E J, SHEN Y L, WALLIS P, et al. LoRA: low-rank adaptation of large language models[EB]. arXiv preprint, 2021, arXiv: 2106.09685.
- [26] GUO X, XIA H T, LIU Z W, et al. FinEval: a Chinese financial domain knowledge evaluation benchmark for large language models[EB]. arXiv preprint, 2023, arXiv: 2308.09975.
- [27] LEI Y, LI J T, CHENG D W, et al. CF-Benchmark: Chinese financial assistant benchmark for large language model[EB]. arXiv preprint, 2023, arXiv: 2311.05812.

作者简介



程大伟（1987-），男，博士，同济大学计算机科学与技术学院副教授、博士生导师，主要研究方向为金融大数据、金融风险防控、生成式人工智能、图数据分析与挖掘等。



贾仁军（2001-），男，同济大学电子与信息工程学院硕士生，主要研究方向为金融大数据与时序预测等。



李江彤（1997-），男，同济大学计算机科学与技术学院博士后，主要研究方向为多模态建模、因果推断、大语言模型、金融大数据等。



丁志军（1976-），男，同济大学计算机科学与技术学院教授、博士生导师，主要研究方向为云计算及服务、大数据征信与金融风控等。



蒋昌俊（1962-），男，中国工程院院士，同济大学计算机科学与技术学院讲席教授，主要研究方向为计算机科学与网络金融安全等。CCF 杰出会员，CCF 数字金融分会主任，973 首席科学家，国家杰出基金获得者，中国人工智能学会监事长，全国优秀科技工作者，获国家技术发明二等奖1项、国家科技进步二等奖2项等。

收稿日期: 2024-09-25

通信作者: 蒋昌俊, cjjiang@tongji.edu.cn

基金项目: 国家重点研发计划项目(No.2022YFB4501704);国家自然科学基金项目(No.62102287, No.62472317);上海市科技创新行动计划项目(No.22YS1400600, No.24692118300)

Foundation Items: The National Key R&D Program of China (No.2022YFB4501704), The National Natural Science Foundation of China (No. 62102287, No. 62472317), Shanghai Science and Technology Innovation Action Plan Project (No. 22YS1400600, No. 24692118300)