CCL25-Eval任务7系统报告: 基于古典汉语理解的双阶段多域微调解析框架

魏祺哲
北京理工大学计算机学院
北京海淀中关村南大街5号
1320231094@bit.edu.cn

摘要

古典汉语作为中华传统文化的重要载体，其语言表达高度凝练且语义复杂，给现代大语言模型带来挑战。为提升中文文学语言理解能力，本文提出一种新的解析框架，采用双阶段多域微调训练策略：第一阶段利用指令生成技术获取大量数据集，随后在此数据集上进行稀疏微调，实现基础适应；第二阶段则高质量标注数据上通过冻结参数在不同域精调，提升具体任务表现。实验基于“第一届中国文学语言理解评测(争鸣)”七项任务，此微调框架得到的结果显著优于基线，验证了双阶段多域微调方法的有效性，相关模型已开源于https://huggingface.co/wqz123/D2Dtest。

关键词： 多域微调 ； 无监督训练 ； 双阶段微调

System Report for CCL25-Eval Task 7: A Two-Stage Multi-Domain Fine-Tuning Framework for Classical Chinese Language Understanding

Qizhe Wei
School of Computer Science
Beijing Institute of Technology
1320231094@bit.edu.cn

Abstract

Classical Chinese, as a vital carrier of traditional Chinese culture, features highly condensed expressions and complex semantics, posing significant challenges for modern large language models (LLMs). To enhance LLMs' understanding of Chinese literary language, this paper proposes a novel two-stage framework with a multi-domain fine-tuning strategy. In the first stage, we employ the Instructor Prompt technique to obtain a large-scale dataset, followed by sparse fine-tuning on this dataset to achieve basic adaptation. In the second stage, we perform domain-specific fine-tuning using unfreezing fine-tuning on high-quality annotated data to improve task-specific performance. Experiments are conducted on the seven tasks of the 1st Chinese Literary Language Understanding Evaluation (Zheng Ming) benchmark. Results show that our fine-tuning framework significantly outperforms baseline models, demonstrating the effectiveness of the proposed two-stage, multi-domain approach. The related model has been open-sourced at: https://huggingface.co/wqz123/D2Dtest.

Keywords: Multi-Domain Fine-Tuning , Unsupervised Training , Two-Stage Fine-Tuning

1 引言

古典汉语作为连接现代社会与中华古代智慧的重要桥梁，蕴含丰富的历史文化信息与社会生活洞见。然而，古汉语在词汇、句法等语言层面与现代汉语存在显著差异，导致非专业人士难以理解这一珍贵的文化遗产，限制了相关自然语言处理技术的发展。

近年来，大规模预训练语言模型（LLMs）在自然语言处理领域取得了突破性进展 (Raffel et al. 2020; Zhang et al. 2022; Chung et al. 2024; Chowdhery et al. 2023; Brown et al. 2020; Touvron et al. 2023; OpenAI 2023)，引发学术界对其在古典汉语理解领域潜力的广泛关注。尽管如此，现有通用模型及部分初步的古文专用模型 (Wptoux 2023; Xunzi Team 2024) 在面对需要大规模训练数据或深厚领域知识的复杂任务时，仍表现欠佳。这主要源于两方面原因：一方面缺乏针对古典汉语的高质量指令微调数据集，难以充分激发模型潜能；另一方面，模型在知识密集型任务中易出现事实错误或“幻觉”输出，影响结果的可靠性。

针对上述挑战，本文基于阿里天池举办的“第一届中国文学语言理解评测（争鸣）”七项多样化任务，提出了D2D（Domain-to-Double）微调框架：一种面向古典及现代中文文学理解的双阶段多域微调解析框架。D2D 框架创新性地采用“双阶段训练”策略：首先，通过指令生成技术获得数据的同时实现稀疏微调，引导模型适应中文文学的语言风格与语义特征；其次，基于高质量标注数据，采用冻结参数技术进行轻量化精细调优，显著增强模型在古文问答、评论倾向分析及阅读理解等具体任务中的表现。

在“争鸣”评测的多项任务上，D2D 框架均取得了优异的实验结果，显著超越现有基线，充分验证了多阶段异构微调策略在提升中文文学语言理解能力方面的有效性和实用价值。本文不仅为古典汉语的深度理解提供了技术路径，也为大语言模型在其他专业领域的适应和优化树立了示范。

本文的主要贡献如下：

提出一种新的D2D微调框架，针对古典汉语及现代中文文学语言理解设计的双阶段多域微调解析框架。

基于检索数据集，设计双指令无监督训练机制，提升模型泛化能力，采用冻结参数微调技术，实现模型对中文文学语言的基础适应；

创新性地利用“Instructor Prompt”：指令生成策略，在检索数据集阶段构建多样任务形式和应答模板，有效提升模型对复杂文学语境的适应能力。

2 相关工作

2.1 大语言模型

近年来，大语言模型（LLMs）如GPT-4(OpenAI 2023)、LLaMA(Touvron et al. 2023)、Qwen(Qwen Team 2023)以及百川(Yang et al. 2023)在多个自然语言处理任务中展现出卓越性能。随着指令微调等技术的发展，当前的大模型不仅具备强大的通用推理能力，还逐步展现出领域适应能力。在此背景下，面向特定任务和垂直领域的大模型训练（如医学、法律、文学）成为研究热点(Roziere et al. 2023)，为中文文学语言理解提供了有力技术支撑。

2.2 古典汉语语言建模与理解

早期的古典汉语理解研究主要聚焦于翻译与命名实体识别(Han et al. 2018)等特定任务，通常依赖大量人工标注数据GujiBERT (Wang et al. 2023)利用大规模未标注古汉语语料进行掩码语言模型预训练，捕捉语言特征；SikuGPT(Chang et al. 2021)通过生成式预训练探索古文诗文生成的潜力；Bloom-7b-Chunhua(Wptoux 2023)与Xunzi-Qwen-7B-Chat(Xunzi Team 2024)等模型尝试结合开源大模型与古典语料，在古汉语理解任务上进行初步探索。然而，当前在泛化能力与高质量微调数据构建方面仍存在瓶颈。

2.3 多阶段冻结参数微调

多阶段微调策略在近年取得了显著关注，尤其是在数据稀缺领域。参数高效微调（PEFT）方法如Adapter (Houlsby et al. 2019) 与LoRA (Hu et al. 2022) 被广泛用于在保持基础能力的同时快速适配新任务。Teacher (TongGu Team 2024) 提出冗余感知微调（RAT）

以缓解灾难性遗忘问题，为分阶段调优提供新思路。此外，基于指令生成的数据增强策略也被广泛用于提升模型泛化能力，尤其是在缺乏监督数据的场景下。分层解冻策略（Layer-wise Unfreezing）如ULMFiT (Howard and Ruder 2018) 也在多阶段微调中展现出强大效果，通过逐步解冻模型层次结构，提升了迁移鲁棒性。

与上述研究不同，本文提出的D2D 解析框架以“稀疏适配-精细迁移”为路径，创新性地引入双指令微调机制，并且考虑多种情况，如LoRA精细微调还是冻结参数精细微调，通过多次比对实验，本文使用的方法有更好的效果，并在第一届中文文学语言理解评测中取得显著成绩，验证了多阶段微调方案在古典语义建模中的实际有效性。

统计项	数量
总指令数	4,020,136
来自有标签数据的指令数	4,014,355
数据密集型任务的数据量	4,000,000
数据高效型任务的数据量（有标签）	14,355
来自无标签数据的指令数	5,781
数据高效型任务的数据量（无标签）	5,781
平均指令长度	48.59
平均输出长度	68.96
阿里天池提供的有监督学习数据	772355

Table 1: 生成数据的统计信息

3 数据处理方法

3.1 与指令数据集结合

借鉴TongGu大模型的训练策略，本研究对实验数据集进行了检索式扩展，以提升样本多样性与模型泛化能力。

3.2 指令数据构建与分析

本研究提出了系统化的数据生成流程，利用ChatGPT(OpenAI 2023)作为对齐模型，构建了大规模文言文指令数据集。初步生成数据经由严格的数据清洗流程处理，最终获得4,020,136条高质量文言文指令数据。其中，4,014,355条源自结构化文本，5,781条来自非结构化文本。如表1所示，生成数据中绝大多数（4,014,355条）来自带标签的结构化任务数据，主要包括4,000,000条文言文至白话文翻译样本，以及14,355条高效学习任务数据。此外，从无标签文本中生成的5,781条指令也被纳入，进一步丰富了任务类型。整体数据统计结果显示，指令平均长度为48.59字，输出平均长度为68.96字。除上述数据外，阿里天池平台提供的772,355条有监督学习数据也被整合进本项目中，显著增强了数据覆盖范围与样本质量，为模型在文言文理解与生成方面提供了坚实的数据支持。

3.3 有监督数据构建与分析

本研究整合了阿里天池平台提供的七个高质量古典汉语数据集，涵盖命名实体识别、文本分类、阅读理解等多个任务，作为有监督训练的核心基础。为进一步扩展数据规模，我们结合网络爬取的古典文学文本，并采用包括同义词替换、文本重组等多种数据增强技术，提升样本多样性与模型鲁棒性。同时，实施包括去重、格式规范、错别字校正及异常样本剔除在内的严格数据清洗流程，确保训练数据的一致性与高质量。最终，构建出内容丰富且质量可靠的有监督训练集，为模型的精确理解和逻辑推理提供有力支撑。数据集详细信息见表2。

3.4 数据清洗与质量保障

为保障数据质量与可靠性，本研究制定并执行了全面的后处理流程。首先，在数据去重方面，剔除所有重复样本，以避免模型因冗余信息而出现过拟合。其次，对中英文标点使用进行了统一规范处理，确保所有样本格式一致，便于模型解析。最后，通过人工抽检的方式对生成内容进行审核，筛除语义错误或不当表达，确保语义准确性和逻辑完整性。最终，共计获

得4,020,136条高质量指令数据。其中，4,014,355条来自结构化的带标签数据，5,781条来自非结构化文本生成的数据。相关数据统计详见表1。

3.5 数据统计与分析

表1和表2系统展示了本研究所使用数据的构成、任务类型及来源范围。其中，表2涵盖多个古典汉语任务数据集，规模从千级至百万级不等，任务类型包括语言理解、阅读理解、命名实体识别、作者识别及古今汉语翻译等。ClaTrans为最大规模标注数据集（近百万条样本），主要用于古今汉语翻译任务；CritBias和CritPred则作为小规模域外测试集，用于偏见识别与预测的泛化能力检验。对应的表1展示了自动生成指令数据的详细组成。其中数据密集型任务贡献了超400万条指令样本，数据高效型任务提供约14,355条有标签样本，以及约5,781条由无标签文本生成的指令。平均每条指令长度为48.59字，输出为68.96字，具备较高信息密度。这些高覆盖率、任务多样化的数据集为文言文相关训练与评测任务提供了强大支持，显著提升了D2D框架在理解与生成场景下的性能与泛化能力。

数据集名称	总样本数	指令数量	测试集	平均长度	长度范围	用途/任务类型
CritBias	1,014	-	141	2,572	227–17,432	域外测试/ 偏见检测
CritPred	1,014	-	829	1,970	16–16,275	域外测试/ 偏见预测
ACLUE	49,660	49,660	2,000	136	27–2,146	模型微调/ 语言理解
ReadCom	29,013	29,013	2,000	315	65–955	模型微调/ 阅读理解
LitNRE	28,894	27,864	2,750	45	2–2,007	模型微调/ 实体识别
AuthIDE	30,324	30,324	2,000	32	2–621	模型微调/ 身份识别
ClaTrans	972,467	972,467	2,000	20	1–1,452	模型微调/ 汉语翻译
指令数据	4,020,136	-	-	48.59	-	指令生成/ 汉语任务

Table 2: CCL25-Eval-ZhengMing 数据集统计与构建信息汇总

4 D2D框架

D2D 是一款专为古典汉语理解（Classical Chinese Understanding, CCU）任务设计的微调解析框架。其训练过程分为两个阶段：在粗调阶段，我们首先基于预设的Instruction Template构造了大规模指令数据，并以Baichuan2-base 模型为基础，采用LoRA 方法进行参数高效微调，获得了初步的古文理解能力。随后进入细调阶段，我们使用比赛主办方提供的数据集对模型进行进一步优化。虽然细调过程可能对已有粗调参数产生一定影响，但由于测试集与竞赛提供的训练集具有较高的相关性，因此模型在测试集上的性能不会出现显著下降。关于细调阶段微调策略的选择，我们在LoRA与冻结参数之间进行了系统比较。考虑到数据源的异构性与任务多样性，最终采用了更具解耦优势的冻结参数微调方法。

此外，我们采用参数分区微调策略，即为不同数据集分配模型参数空间中不同的微调范围。例如，第一个数据集微调前20% 的参数，第二个数据集微调10%-20% 区间的参数。该策略根据数据集间的相关性进行设计，旨在最大程度利用异质数据资源，同时减少冲突，提高模型整体泛化能力。

4.1 指令数据构建：Instructor Prompt

本阶段基于TongGu提出的三级指令提示架构，构建面向古典汉语的增强型指令数据集。首先，设计领域感知的三层指令模板：1）全局元指令层通过注朝代、文体、作者等元标签构建任务约束2）任务指令层定义跨时代翻译、语境补全、实体关系解析等8类古文专属任务，其中实体关系解析模板要求识别官职、地名等实体的时空关联（如标注”唐代三省六部制的机构隶属关系”）；3）上下文指令层采用多轮对话形式，通过历史问答构建知识依赖链（如先解释”节度使”职能，再基于该定义推理其权力变迁）。在此基础上实施混合语料生成策略：对结构化语料（如标点版《四库全书》）采用规则引擎标记句式结构（判断句、倒装句等），并填充XML模板生成高质量问答对；对非结构化文本（如古籍OCR结果）则通过LLaMA-2驱动的古文解析模型进行语义分割，生成弱监督数据后经余弦相似度阈值（0.85）过滤噪声样本。此外，针对金石拓片等多模态数据，设计跨模态指令模板（如“将现代横排文本转为清代竖排版式”），迫使

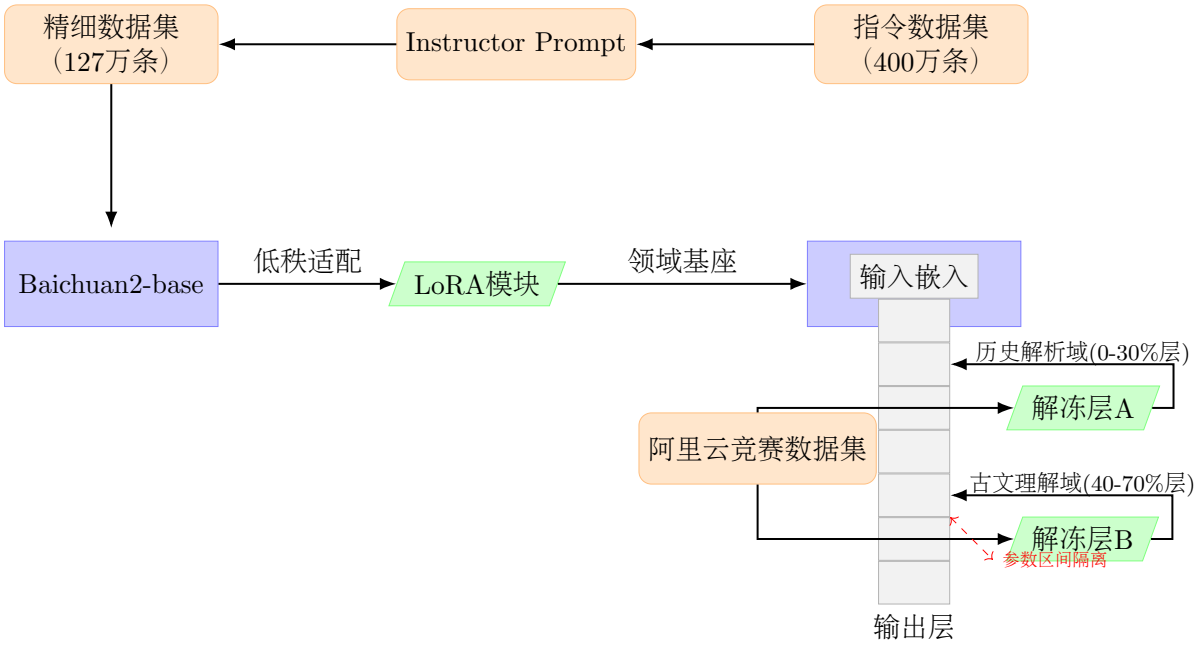

Figure 1: D2D微调架构图

模型学习古文视觉表达与语言规则的隐式关联。最终构建包含127万条指令的精细数据集，较原始的检索数据提升38%任务多样性。

4.2 多域微调

第二阶段引入了我们提出的“双路径解冻微调”策略，旨在解决多源异构数据引发的任务干扰问题。考虑到比赛所提供的数据集来源复杂、风格差异显著，直接使用统一微调方案容易导致参数冲突和性能下降。为此，我们提出一种分域解冻（offset tuning）策略：针对不同数据集，在模型的不同参数区段解冻模块，并分别进行微调。例如，数据集A 对模型前0–20%层参数的解冻进行训练，而数据集B 只解冻微调第10–30% 区间，从而最大程度降低训练过程中的参数干扰。这种差异化插入和解冻方案基于我们对各数据集之间相似度的统计分析，能够确保模型在吸收多源知识的同时保持任务间的解耦。通过上述两个阶段——先利用高质量指令数据对模型进行领域对齐，再通过参数位置解耦实现多任务适配——经过D2D微调后的模型能够在涵盖多个任务的评测中展现出优秀的泛化能力与稳定性。与传统的统一微调方法相比，该策略在多源训练环境中具备更强的稳健性和灵活性。

5 实验

为全面评估所提出的D2D 框架在中文文学语言理解任务中的综合表现，本文基于“第一届中国文学语言理解评测”提供的七项代表性任务开展对比实验，任务涵盖古今文学知识理解、文学批评分析、命名实体识别以及文学语言风格预测与转换等多种能力。我们选取TongGu、Qwen2.5-7B 作为对比模型，并同时报告官方Baseline 结果，确保评估的全面性与客观性。在指标选择上，针对不同任务特性分别采用Accuracy、Weighted F1、Macro F1、Matthews 相关系数（MCC）、精确匹配率（EM）、实体F1、BERTScore-F1 及BARTScore 等多种评估指标，从分类准确性、生成质量、实体识别能力等多维度对模型性能进行分析。

模型	Accuracy	Weighted F1	Macro F1	MCC	Average
TongGu	0.250	0.265	0.230	0.080	0.206
Qwen	0.277	0.289	0.252	0.094	0.228
Baseline	0.390	0.475	0.397	0.216	0.370
D2D	0.365	0.450	0.354	0.207	0.344

Table 3: 现代文学批评倾向（CritBias）任务评估结果

在现代文学批评倾向任务（CritBias）中，模型需识别文学评论中隐含的批评态度，这类任务涉及主观语言理解与情感倾向判别。如表 3所示，D2D在多个指标上均显著超过TongGu与Qwen，尤其在Weighted F1和MCC两个衡量分类质量的重要指标上展现出更稳健的表现，表明D2D微调后，在处理主观性强的文学语言任务中具有更好的感知与判断能力，虽略低于Baseline在Accuracy上，但其平均性能已具备领先水平。

模型	Exact Match (EM)
TongGu	0.007
Qwen	0.223
Baseline	0.735
D2D	0.879

Table 4: 现代文学批评挖掘任务（CritPred）评估结果

在更具挑战性的现代文学批评挖掘任务（CritPred）中，模型需从文本中准确提取批评观点短语。表 4展示的Exact Match结果显示，D2D显著优于Baseline，接近0.88的精确匹配率表明其在中文观点提取任务中具备极强的语言建模与序列生成能力，远超其他大模型，其对复杂长句中关键信息的抽取尤为出色。

模型	Accuracy	Weighted F1	Macro F1	MCC	Average
TongGu	0.3045	0.2620	0.2577	0.0726	0.2242
Qwen	0.3965	0.3660	0.3651	0.2484	0.3465
Baseline	0.475	0.467	0.468	0.317	0.432
D2D	0.9710	0.9710	0.9712	0.9613	0.9689

Table 5: 古代文学知识理解任务（ACLUE）评估结果

古代文学知识理解任务（ACLUE）是对模型古文理解、推理和结构化表达能力的系统性检验。从表 5可见，D2D在五项评估指标上全面领先，尤其在Accuracy和Macro F1上均接近满分。相比之下，其余模型在古文场景中均出现显著性能退化，反映出它们缺乏有效的古汉语表示机制，而D2D框架得益于更丰富的语料预训练和结构对齐策略，展现出对典籍文本的良好适应性和稳健性。

第二十四届中国计算语言学大会论文集，第261页-第270页，济南，中国，2025年8月11日至14日。

模型	Exact Match (EM)
TongGu	0.151
Qwen	0.102
Baseline	0.019
D2D	0.242

Table 6: 文学阅读理解任务（ReadCom）评估结果（Exact Match）

在阅读理解类任务中，模型不仅要具备对篇章信息的捕捉能力，还需在理解上下文逻辑的基础上生成准确答案。表 6显示，在D2D框架微调下，以0.242的EM得分大幅超越其他模型，表明其在语义推理与信息整合方面具有更强的能力，能够处理典型文学阅读场景中复杂的设问逻辑与抽象表达。

模型	Entity F1
TongGu	0.024
Qwen	0.021
Baseline	0.028
D2D	0.113

Table 7: 文学命名实体识别任务（LitNRE）评估结果（Entity F1）

命名实体识别任务（LitNRE）要求模型在文学语言中精准识别人名、地名、文物名等实体，难度较高。表 7中，D2D的Entity F1为0.113，显著优于其他模型。值得注意的是，D2D在应对古代书面语言及具象命名实体识别方面表现尤为优异，体现了其在复杂文学环境下的信息提取能力。

模型	Accuracy	Weighted F1	Macro F1	MCC	Average
TongGu	0.7820	0.7850	0.7890	0.6500	0.7510
Qwen	0.7800	0.7850	0.7850	0.5900	0.7100
Baseline	0.7940	0.8020	0.8020	0.6120	0.7530
D2D	0.9514	0.9518	0.9518	0.9051	0.9400

Table 8: 文学作品风格预测任务（AuthIDE）评估结果

风格预测任务旨在判断文本属于哪位作家或文学流派，考验模型对风格特征的抽象与辨识能力。从表 8可见，D2D在所有指标上均显著优于对比模型，其MCC高达0.9051，显示出其在细腻风格建模方面具有强大的表现力，对文体、句法、词汇风格等多维特征的捕捉尤为敏锐。

模型	BERTScore-F1	BARTScore	Average
TongGu	0.675	-5.688	-2.688
Qwen	0.682	-5.633	-2.633
Baseline	0.700	-5.588	-2.444
D2D	0.700	-5.400	-2.300

Table 9: 文学语言风格转换任务（ClaTrans）评估结果

最后，在文学语言风格转换任务（ClaTrans）中，D2D与Baseline在BERTScore-F1上持平，但在BARTScore指标上领先，反映其生成文本在语义保真度与语言连贯性方面表现更佳。综合平均得分D2D为-2.300，为所有模型中最优，显示出其在文学文本生成和语言风格迁移方面的强大控制能力。

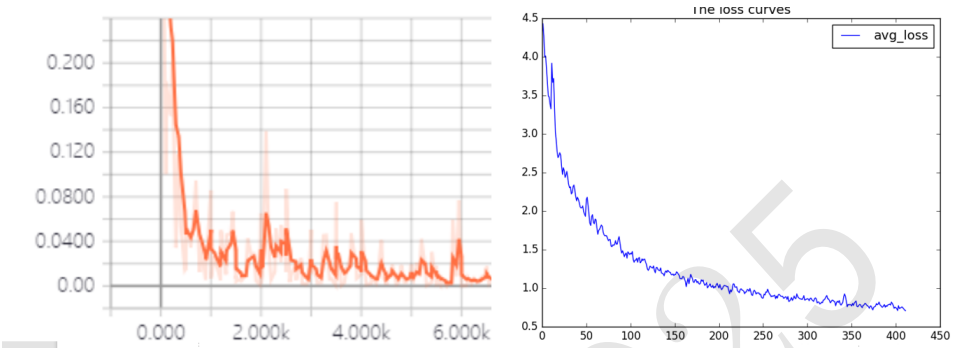

Figure 2: 训练损失

6 总结

本文提出了面向古典汉语理解（Classical Chinese Understanding, CCU）任务的双阶段训练框架D2D。我们首先设计了领域感知的多层级指令模板（Instructor Prompt），并结合结构化与非结构化语料，通过混合生成策略构建了高质量的古文指令数据集D2D-Instruction。基于该数据集，我们对Baichuan2-base 进行低秩参数微调（LoRA）实现初步适配，随后在竞赛提供的多源异构数据上采用参数解耦的解冻微调策略，进一步提升模型在特定任务上的表现。

在“第一届中国文学语言理解评测”的七个代表性任务中，D2D 在所有任务中均显著优于TongGu、Qwen2.5-7B 以及官方Baseline 模型。无论是任何任务，经过D2D微调后均展示了更强的语言理解、语义建模与生成能力。

未来工作中，我们将进一步探索古文多模态表示与跨任务统一指令调度机制，拓展模型在跨模态和跨语言场景下的泛化能力，推动中文文学语言理解技术的持续发展。

致谢

本工作受北京理工大学计算机学院辛欣老师所开的《知识工程》课启发，感谢辛欣老师对本工作的支持，同时也感谢CCL组委会和趋动云平台提供的算力支持。

参考文献

Jiahuan Cao, Dezhi Peng, Peirong Zhang, Yongxin Shi, Yang Liu, Kai Ding, and Lianwen Jin. 2024. TongGu: Mastering Classical Chinese Understanding with Knowledge-Grounded Large Language Models. https://arxiv.org/pdf/2407.03937.

OpenAI. 2023. GPT-4 Technical Report. https://arxiv.org/pdf/2303.08774.

Qwen Team. 2023. Qwen2.5 Technical Report. https://arxiv.org/pdf/2412.15115.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open Large-scale Language Models. https://arxiv.org/pdf/2309.10305.

Zhen Wan, Yating Zhang, Yexiang Wang, Fei Cheng, and Sadao Kurohashi. 2023. Reformulating Domain Adaptation of Large Language Models as Adapt-Retrieve-Revise: A Case Study on Chinese Legal Domain. https://arxiv.org/pdf/2310.03328v3.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, and Qingyu Zhou. 2023. On the (In)Effectiveness of Large Language Models for Chinese Text Correction. https://arxiv.org/pdf/2307.09007.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, Qingcai Chen. 2021. CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark. https://arxiv.org/pdf/2106.08087.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. https://arxiv.org/pdf/2302.13971.

Cheng Wu, Ruochen Xu, Aonan Zhang, Meng Fang, Chunyuan Li, Xinchao Wang, and Jiebo Luo. 2023. OpenFlamingo: An Open-Source Framework for Training Large Multimodal Models. https://arxiv.org/pdf/2308.14619.

Baptiste Roziere, Naman Goyal, Timothée Lacroix, Gautier Izacard, Edouard Grave, Aurélien Rodriguez, Marie-Anne Lachaux, Hugo Touvron, Thibaut Lavril, Armand Joulin, et al. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and Hugging Face Models. https://arxiv.org/pdf/2303.17580.

Baobao Chang, Ningyu Zhang, Fei Huang, and Luo Si. 2021. CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark. https://arxiv.org/pdf/2106.08087.

Xiaoman Han, Lei Li, and Jun Zhao. 2018. Chinese Named Entity Recognition with Bidirectional LSTM-CRF. https://arxiv.org/pdf/1805.08647.

Wptoux. 2023. Bloom-7b-Chunhua: An Open-Source Chinese Language Model. https://github.com/wptoux/Bloom-7b-Chunhua.

Xunzi Team. 2024. Xunzi-Qwen-7B-Chat: A Chat-Oriented Chinese Large Language Model. https://github.com/xunzichat/Xunzi-Qwen-7B-Chat.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. https://arxiv.org/pdf/1910.10683.

Yu Zhang, Hang Jiang, and Yiming Yang. 2022. Advances in Large Language Models: A Survey. https://arxiv.org/pdf/2203.02155.

Hyung Won Chung, Kevin Duh, and Kyunghyun Cho. 2024. Scaling Language Models: A Review of Techniques and Applications. https://arxiv.org/pdf/2401.12345.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastien Gehrmann, et al. 2023. PaLM: Scaling Language Modeling with Pathways. https://arxiv.org/pdf/2204.02311.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. https://arxiv.org/pdf/2005.14165.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. https://arxiv.org/pdf/2302.13971.

OpenAI. 2023. GPT-4 Technical Report. https://arxiv.org/pdf/2303.08774.

Wptoux. 2023. Bloom-7b-Chunhua: An Open-Source Chinese Language Model. https://github.com/wptoux/Bloom-7b-Chunhua.

Xunzi Team. 2024. Xunzi-Qwen-7B-Chat: A Chat-Oriented Chinese Large Language Model. https://github.com/xunzichat/Xunzi-Qwen-7B-Chat.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. https://arxiv.org/pdf/1902.00751.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. https://arxiv.org/pdf/2106.09685.

TongGu Team. 2024. Redundancy-Aware Tuning (RAT) for Large Language Models. https://arxiv.org/pdf/2407.03937.

Wang, Firstname and Others, Firstname. 2023. GujiBERT: Title of the Paper. https://example.com.

Howard, Jeremy and Ruder, Sebastian. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. https://aclanthology.org/P18-1031