基于古汉语大语言模型的多任务学习探究

姚欣宇[1,2], 王梦笛[1,2], 高原[1,2], 高歌[2,3,*], 陈波[1,2,*], 赵小兵[1,2]
[1]中央民族大学，信息工程学院，北京，100081
[2]国家语言资源监测与研究民族语言中心，北京，100081
[3]中央民族大学，中国少数民族语言与文学学院，北京，100081
{xinyu230101, wang_mengdi_wmd, gaoyuan_g_y}@163.com
ggao.bj@qq.com, chenbomuc@muc.edu.cn, nmzxb_cn@163.com

摘要

随着大语言模型在多任务学习领域展现强大泛化能力，其在低资源古汉语场景的应用价值亟待探索。本文基于LLaMA3-Chinese-8B利用21GB高质量古汉语语料进行增量预训练，接着进行十项任务微调（包括句读、词性标注、命名实体识别（NER）、事件识别、翻译、词语解释、反向词典、历史人物知识、诗歌赏析、诗歌生成），设计了单任务微调和双任务组合微调两种策略，通过55组实验量化了任务之间的正增益与负增益，首次系统揭示了古汉语多任务学习中的增益关系。实验结果表明，不同任务之间存在协同效应与任务干扰效应，并且具有不对称性。基础类古汉语任务之间表现出更强的协同效应，相比之下，翻译类和生成类任务之间协同效应表现较弱。同时，受双任务设定的影响，不同古汉语任务的稳定性存在明显差异。

关键词： 多任务学习；大语言模型；古汉语任务

Research on Multi-task Learning Based on Ancient Chinese Large Language Model

Xinyu Yao[1,2], Mengdi Wang[1,2], Yuan Gao[1,2], Ge Gao[2,3,*], Bo Chen[1,2,*], Xiaobing Zhao[1,2]
[1]School of Information Engineering, Minzu University of China, Beijing, 100081
[2]National Language Resource Monitoring & Research Center for Minority Languages, Beijing, 100081
[3]School of Chinese Ethnic Minority Languages and Literatures, Minzu University of China, Beijing, 100081

Abstract

As large language models (LLMs) demonstrate strong generalization abilities in the field of multi-task learning, their application potential in low-resource Classical Chinese scenarios requires further exploration. In this paper, we perform continuous pretraining on LLaMA3-Chinese-8B using 21GB of high-quality Classical Chinese corpora, followed by fine-tuning on ten tasks, including punctuation, part-of-speech tagging, named entity recognition (NER), event identification, translation, word explanation, reverse dictionary, historical figure knowledge, poetry appreciation, and poetry generation. We design two fine-tuning strategies: single-task fine-tuning and dual-task combined fine-tuning, and conduct 55 experimental configurations to quantify the positive and negative transfer effects between tasks. This paper systematically reveals the gain relationships in Classical Chinese multi-task learning for the first time. Experimental results show that both synergistic and interference effects exist between tasks, and these effects exhibit asymmetry. Basic Classical Chinese tasks exhibit stronger synergistic effects, whereas translation and generative tasks show relatively weaker interactions. Additionally, due to the dual-task setting, the stability of different Classical Chinese tasks varies significantly.

Keywords: Multi-Task Learning , LLMs , Classical Chinese Task

* 通讯作者

1 引言

随着大语言模型（Large Language Models, LLMs）在通用领域的突破性进展，其在古籍智能处理中的应用价值日益凸显。中华古籍作为承载传统文化的重要载体，具有词汇体系特殊、句法结构复杂、语义表达隐晦等特征，亟需适配性的语言模型支撑其数字化研究与传播。尽管主流大模型（如GPT-4 (Achiam et al., 2023)、LLaMA3 (Grattafiori et al., 2024)）在现代文本处理中表现卓越，但在古汉语领域仍面临显著局限：其一，模型预训练语料中古汉语占比不足，导致对文言文语法规则、专有词汇（如“鼎”“笏”等器物词）的泛化能力薄弱；其二，现有研究多聚焦于单任务优化（如句读或翻译），缺乏能够统一处理多任务的模型，难以应对古籍处理中的多任务实际需求。基于此，多任务协同训练作为提升古汉语处理模型能力的潜在方向，为破解古汉语任务数据稀疏性和模型任务单一性难题提供了新路径。它通过共享表征空间促进知识迁移，联合优化所有任务，并使单一训练模型适应所有任务 (Crawshaw, 2020)。现有研究表明多任务协同训练在机器翻译、文本分类等现代语言任务中的增益效应 (Wang et al., 2020; Lu et al., 2020; Escolano et al., 2021)，并发展出任务路由 (Ding et al., 2023; He et al., 2024)、梯度调制 (Peng et al., 2022; Zhang et al., 2024)等优化方法，但针对古汉语任务间知识迁移规律的系统性研究尚属空白，任务组合的增益方向与强度仍缺乏量化依据。

本文聚焦于古汉语任务间的增益关系这一核心问题。现有研究存在两大空白：第一，语义共享机制具有复杂性，古汉语任务间的知识交互受制于语言特性的多维差异。例如，词性标注需依赖句读结果以划分语法单位，而反向词典构建可能需要结合词义解释的上下文特征，此类任务间的依赖关系尚未被量化分析；第二，任务冲突具有不可预测性，部分任务组合可能因目标函数竞争导致负迁移效应。如诗歌生成注重创造性表达，而事件识别需严格遵循事实逻辑，二者在共享模型参数时可能存在语义冲突。因此，系统揭示古汉语任务间的增益机制，不仅关乎多任务模型性能优化，更对理解古汉语认知规律具有理论价值。

本研究的主要贡献体现在以下两个方面：

（1）古汉语多任务评测体系设计：构建包含句读、翻译、诗歌赏析、诗歌生成等十类古汉语任务的基准数据集，涉及基础类任务、翻译类任务和生成类任务，系统评估大语言模型在古汉语多任务场景下的表现，填补古汉语多任务评估资源的空白；

（2）古汉语任务增益矩阵与增益分析：通过对比单任务微调与双任务组合微调策略，首次构建古汉语任务增益矩阵，揭示任务间的协同与抑制效应，为多任务模型的结构优化与任务选择提供了数据驱动的理论依据。

2 相关工作

2.1 古汉语模型研究进展

随着大语言模型（LLMs）在通用领域的突破性进展，古汉语智能处理逐渐成为研究热点。然而，受限于古汉语的复杂性与数据稀缺性，相关模型的发展路径呈现单任务优化与多任务探索并行的特点。早期古汉语模型以任务独立优化为核心，研究聚焦于句读标点、分词、翻译等基础任务。隐马尔可夫模型（Hidden Markov Model, HMM）、条件随机场（Conditional Random Field, CRF）、支持向量机（Support Vector Machine, SVM）等方法被广泛应用在古汉语词性标注 (Huang et al., 2002)、古汉语断句 (Huang et al., 2010)、命名实体识别 (Yuan et al., 2019; Li, 2018; Meng et al., 2015)等单任务的处理及优化中，但受限于古汉语复杂的语法结构和语义歧义性，其泛化能力与准确率均存在瓶颈。随着深度学习技术的不断发展，循环神经网络（Recurrent Neural Network, RNN）及其变体为古籍单任务智能处理优化提供了新的解决方案，如古汉语翻译 (Lample et al., 2017)、诗词生成 (Yang et al., 2019)等。

针对古汉语标注数据稀缺但任务关联性强的特点，多任务学习逐渐成为研究焦点。大语言模型（LLMs）与多任务学习（MTL）的结合已成为推动古汉语自然语言处理的关键技术路径。专注于古籍文言文处理的TongGu (Cao et al., 2024)模型通过两阶段指令微调与检索增强生成（CCU-RAG）技术，实现有效处理24个不同文言文理解任务，成为理解文言文的有力工具。此外，“荀子”[1]、“AI太炎” (Li et al., 2024)、WenyanGPT (Yao et al., 2025)古籍大模型以及一系列基于BERT和GPT架构的古籍预训练模型，如“Guji” (Wang et al., 2023)系列模型、SikuBERT (Wang et al., 2022)、SikuGPT (Liu et al., 2024)等，均支持古汉语相关任务的

[1]https://github.com/Xunzi-LLM-of-Chineseclassics/XunziALLM

第二十四届中国计算语言学大会论文集，第259页–第272页，济南，中国，2025年8月11日至14日。

协同处理。但古汉语多任务协同机制的系统性研究仍是空白，缺乏对古汉语任务间知识共享与冲突的定量分析。

2.2 多任务学习

多任务学习（Multi-Task Learning, MTL）是一种归纳迁移机制，其核心思想是通过利用相关任务训练信号中包含的领域特定信息来提高泛化能力 (Caruana, 1997)。近年来，MTL在自然语言处理多个任务中展现出显著优势，尤其在任务数据有限或任务间具有关联性的场景下更为突出。

任务之间的相关性是多任务学习成功的关键因素之一 (Zhang et al., 2023)。高相关性的任务协同训练不仅有助于加快模型的收敛速度，还能显著提升整体性能 (Kandemir et al., 2014; Jaques et al., 2017; Guo et al., 2020)。Lu等人 (Lu et al., 2020)将MTL应用于情感分析任务，并引入了变分自编码器生成模型（MTVAE），在提升情感分类准确率的同时也促进了相关任务的性能提升。Lu等人 (Lu et al., 2019)在LSTM结构中引入了共享与任务特定的参数，成功实现了命名实体识别与词性标注等基础类任务的协同学习；Cao等人 (Cao and Yogatama, 2020)则在任务嵌入空间中建模潜在共享技能，提升了生成类任务如摘要生成和问答任务的协同表现。

在结构化预测任务中MTL的研究相对较少，但已有研究展示了其在资源受限场景下的潜力。例如，Luong等人 (Luong et al., 2015)指出，辅助任务的数据量若过大，可能导致负迁移效应；Benton等人 (Benton et al., 2017)则强调，在目标任务样本有限的情况下，MTL能显著提升性能。为进一步理解任务间的相互影响，Martínez Alonso与Plank (Alonso and Plank, 2016)系统性地分析了主要任务与辅助任务的组合条件，发现辅助任务标签的均匀分布是成功迁移的关键因素。此外，Mou等人 (Mou et al., 2016)也探讨了MTL与迁移学习的边界，分析了在哪些条件下这些技术在句子分类任务中是可行的。Joachim Bingel (Bingel and Søgaard, 2017)则从学习曲线角度出发，指出，如果主要任务在20-30%的百分位数内有平坦的学习曲线，但辅助任务的曲线仍然相对陡峭，那么MTL更有可能发挥作用，但尽管如此，在90项测试中，只有40项产生了增益。近期的大量研究强调，MTL模型的成功取决于这些辅助任务的正确选择 (Guo et al., 2019; Kung et al., 2021; Chen et al., 2022; Grégoire et al., 2024)。

为公平评估MTL模型架构本身的性能，现有研究通常在单任务和多任务设置下保持超参数不变，从而排除其他变量干扰 (Alonso and Plank, 2016; Bingel and Søgaard, 2017; Zhan and Zhang, 2024)，使实验结果更真实地反映MTL架构本身的优势或局限。

值得注意的是，近年来兴起的参数高效微调方法，如LoRA (Hu et al., 2022)，虽通过引入低秩约束，在某些场景中表现良好，但在任务差异较大的复杂多任务设置中，其建模能力和适应性可能受限，难以充分捕捉不同任务之间的细粒度差异 (Liu et al., 2025; Yang et al., 2025)。

3 基于微调古汉语大模型的多任务学习

3.1 古汉语预训练模型构建

预训练阶段使用的语料主要来自以下几个渠道：（1）古诗文网、文言古籍网等公开权威的古籍整理网站；（2）Github上开源的古汉语相关数据集与个人整理资源；（3）现有的开源古汉语任务基准和数据集，如ACCN-INS (Cao et al., 2024)数据集。所有语料在采集后均进行了统一的格式化处理和严格的清洗流程。该流程包括：去除HTML标签与脚注信息、统一段落结构、剔除特殊符号与无效字符，以及格式统一（如标点规范化）等。

通过上述处理后，我们获得了约21GB的纯净古汉语语料数据。语料内容广泛，涵盖中国传统“四部”分类体系中的经、史、子、集四大类，即包括《论语》《孟子》《史记》《汉书》《庄子》《韩非子》《昭明文选》等经典典籍，又收纳诸如笔记小说、戏曲变文、铭文碑刻等多种文学样式，还包括宗教经典、地方志书、谱牒、医书、兵书、农政典籍、科技文献等多元化内容，时间跨度自先秦延续至民国，涵盖繁体、简体、异体及古今字形，具备极高的语言多样性与历史深度。详细语料来源及数据规模如表1所示。

在模型结构方面，我们选用LLaMA3-Chinese-8B作为基础架构。该模型在性能与参数规模之间取得了良好的平衡，特别适合进行大规模中文文本的迁移预训练。我们在标准的因果语言模型（Causal Language Modeling, CLM）任务上对该模型进行了增量预训练，以便更好地适应古汉语语料的特点。

此外，为了进一步提升训练效率，我们引入了低秩自适应（LoRA） (Hu et al., 2022)微调技术，作为一种参数高效的训练方法。LoRA不直接修改预训练模型原有的稠密权重矩阵，而是引入额外的低秩矩阵来表示权重变化的部分，能够显著减少需要更新的参数数量，从而降低计算资源的消耗，同时保持模型的高性能。我们将学习率设置为1.0e-4，避免小学习率导致收敛缓慢或大学习率引发梯度爆炸；批大小为8，梯度累积步数为1，避免显存溢出；采用混合精度训练BF16，相比FP32（32 位浮点数）占用更少显存；热身比例为0.1，使模型在随机初始化阶段先以小步长适应古汉语数据分布，避免因突然使用高学习率导致梯度爆炸；采用余弦调度器，在训练初期保持较高学习率以快速探索参数空间，后期逐渐降低学习率以精细调整。最终，在两张A800GPU (80GB) 上进行了一次18.24天的增量训练。预训练阶段的主要模型参数设定详见表2。

数据来源	数据规模	数据来源	数据规模
chinese-poetry-master	111.90 MB	network	1.04 GB
GuWen-master	8.97 MB	guner2023-master	48.02 MB
text-to-picture-sidamingzhi	6.89 MB	TCM-Ancient-Books-master	234.81 MB
Reservator-master	79.97 MB	PoetrySplider	13.92 MB
chinese-dictionary-main	12.30 MB	Classical-Modern-main	476.74 MB
wenyanguji-directory	1.57 GB	chinese-novel-master-long	245.96 MB
erya	7.72 GB	gushiwenwang	59.40 MB
zh-ancient-texts-master	22.32 MB	core-texts	209.93 MB
Classical-Chinese-master	273.99 MB	chinese-xinhua-master	47.31 MB
daizhigev20	5.04 GB	poems-db-master	498.98 MB
ACCN-INS	1.06 GB	Poetry-master	1.09 GB
core-books-main	747.68 MB	chtxt-main	87.50 MB
chinese-gushiwen-master	22.88 MB	未知	未知

Table 1: 数据来源与规模

参数	值	参数	值
cutoff-len	1024	learning-rate	1.0e-4
finetuning-type	lora	num-train-epochs	1.0
per-device-train-batch-size	8	gradient-accumulation-steps	1
lr-scheduler-type	cosine	warmup-ratio	0.1
lora-rank	64	fp16	True

Table 2: 预训练阶段超参数设置

3.2 微调任务

在我们的实验中，我们考虑了十个古汉语任务，并将其划分为三个类型：在本文中，我们系统地考察了涵盖古汉语处理的十项核心任务，并将这些任务依据其核心目标与处理方式划分为以下三个主要类别。我们按照任务类型给出了各任务的定义、分类、示例指令（输入和输出），如图1 - 3所示。

（1）基础类任务：此类任务聚焦于文本的基础结构与信息抽取。具体包括：句读、词性标注、命名实体识别、事件识别。

（2）翻译类任务：此类任务侧重于古汉语与现代汉语之间的语义阐释。具体包括：翻译、词语解释。

（3）生成类任务：此类任务要求模型基于对古汉语知识的理解进行创造性的信息生成或深度应用。具体包括：反向词典、历史人物知识、诗歌赏析、诗歌生成。

第二十四届中国计算语言学大会论文集，第259页–第272页，济南，中国，2025年8月11日至14日。

句读

定义：正确地使用标点符号将连续的古汉语文本分割为符合语法的句子。
分类：，。《》：？“”！（）；<>

指令：
输入：请为下面的古汉语文本添加标点符号：汝忆否四五年前某夕吾尝语曰与使吾先死也无宁汝先吾而死
输出：汝忆否？四五年前某夕，吾尝语曰：“与使吾先死也，无宁汝先吾而死。”

词性标注

定义：为古汉语文本中的每个词语标注词性（如动词、名词、形容词等）
分类：nr 人名; v 动词; n 名词; r 代词; w 标点符号; c 连词; p 介词; d 副词; t 时间词; y 语气词; u 助词; m 数词; a 形容词; f 方位词; ns 地名; j 缩略语; q 量词。

指令：
输入：请对所给的古汉语语句进行分词并标注词性：十三年，春，宋向魋救其師。
输出：十三年/t ，/w 春/n ，/w 宋向魋/nr 救/v 其/r 師/n 。/w

命名实体识别

定义：识别古汉语文本中的人物、地名、时间、官职名四类实体信息。
分类：人物、地名、时间、官职名

指令：
输入：抽取这段话中的人名、地名、时间和官职名：二十三年，罢宣慰司，立辽阳行省，以亦力撒合为参知政事。
输出：'人物': ['亦力撒合']，'时间': ['二十三年']，'官职': ['宣慰司','参知政事']

事件识别

定义：从古汉语文本中识别历史事件类型及对应的触发词。
分类：采用CHED数据集分类标准，共9大类67小类。

指令：
输入：识别古文内容“赵壮者尽于长平，其孤未壮，举兵击赵。”中的历史事件类型及其触发词。
输出：事件1：事件类型：军事-备战-出兵，触发词：举。 事件2：事件类型：军事-作战-攻击-征伐，触发词：击。

Figure 1: 基础类任务

翻译

定义：将古汉语文本翻译为现代汉语。

指令：
输入：请将“古之学者必有师。师者，所以传道受业解惑也。”翻译成白话文。
输出：古代求学的人必定有老师。老师，是用来传授道理、教授儒家经典著作、解释疑难问题的。

词语解释

定义：为古汉语文本中的词语提供简明准确的解释。

指令：
输入：《花犯·谢黄复庵除夜寄古梅枝》中，“寒泉贮、绀壶渐暖，年事对、青灯惊换了。”的“绀壶”是什么意思？
输出：指插梅枝的天青色水壶。绀（gàn），深青带红的颜色。

Figure 2: 翻译类任务

反向词典

定义：根据定义或描述返回相应的词语。

指令：
输入：哪个成语的含义是“去掉外饰，还其本质。比喻回复原来的自然状态。”?
输出：返璞归真

历史人物知识

定义：回答与历史人物相关的背景信息、事迹、成就以及与其他人物、事件的关系等。

指令：
输入：辛弃疾的字和号分别是什么？有何寓意？
输出：辛弃疾原字坦夫，后改字幼安，别号稼轩。“幼安”有祈求平安之意，“稼轩”则体现其归隐田园后对农耕生活的向往。

诗歌赏析

定义：对给定的古典诗词进行情感与结构分析，提供赏析内容。

指令：
输入：赏析杜甫的《春望》，内容如下：国破山河在，城春草木深。感时花溅泪，恨别鸟惊心。烽火连三月，家书抵万金。白头搔更短，浑欲不胜簪。
输出：此诗前四句写春日长安凄惨破败的景象，饱含着兴衰感慨；后四句写诗人挂念亲人、心系国事的情怀，充溢着凄苦哀思。全诗格律严整……

诗歌生成

定义：根据给定主题或风格，生成符合韵律的古典诗词歌赋。

指令：
输入：请用一首词描写边疆战士在战争中的艰辛与思乡之情。
输出：点绛唇·狼烟
烽火连天，狼烟滚滚遮空际。战尘千里，铁马冰河里。　　戍卒望乡，泪眼盈襟底。何时止？待春风起，共赏桃花水。

Figure 3: 生成类任务

第二十四届中国计算语言学大会论文集，第259页-第272页，济南，中国，2025年8月11日至14日。

3.3 任务训练和评估数据集

数据集原始语料主要来自公开古汉语语料库和权威文本资源。所有数据经过脚本清洗，包括格式化、去重和标准化，然后由四名NLP硕士研究生进行二次校验，确保文本完整性、统一异体字等。对于标注数据，三名NLP硕士研究生进行人工校验，确保标注符合语言规范、任务间标注统一，并修正可能的标注误差。我们按照8:1:1划分了训练集、开发集和测试集。表3总结了我们所使用的数据集的来源和划分情况。

任务	数据集来源	训练集	开发集	测试集
句读	classical-modern	42496	5312	5312
词性标注	EvaHan2022	9952	1242	1247
命名实体识别	春秋三传和二十四史	29923	3740	3741
事件识别	CHED	5650	1218	1218
翻译	classical-modern	41332	6383	6385
词语解释	文言古籍网	31088	3863	3934
反向词典	chinese-dictionary和chinese-xinhua	35703	4462	4461
历史人物知识	古诗文网	5000	1125	1125
诗歌赏析	古诗文网	9526	1162	1252
诗歌生成	Qwen-2.5-14b (Yang et al., 2024)生成	11232	1402	1407

Table 3: 任务训练和评估数据

3.4 多任务学习方案

为充分发挥预训练模型在多类古文任务中的泛化能力，我们设计并实施了一套基于指令微调的多任务学习框架，涵盖古籍文本处理的10个核心任务。训练过程中，我们以单任务微调为基线，并进行系统性的双任务组合训练实验，以探索任务间的互促关系和共享能力。

多任务训练采用统一的指令微调范式。在训练样本中，每个任务样本均包含结构化的“指令+输入+输出”，指令部分（每个任务的指令见图1 - 3）以自然语言明确任务意图，引导模型理解目标操作；输入部分为任务相关的原始文本或片段；输出为任务期望结果。

我们首先对每个任务分别进行单任务微调，作为模型在该任务上的独立性能基线。模型在保持预训练初始化权重不变的前提下，仅在对应任务指令数据上进行训练，评估其单任务表现，为多任务组合效果提供对比依据。为研究不同任务间的知识迁移与协同关系，我们构建了全覆盖的双任务组合实验设计。具体地，对于选定的10个核心任务，构建了其所有不重复的两两组合，共计45组任务对。在每组训练中，模型接收来自两个任务的数据样本进行微调，学习在多样语义场景下的对齐与泛化能力。通过比较双任务训练结果与各自的单任务基线，可以分析任务间的正负迁移情况，识别任务协同的有效组合与潜在冲突结构。为保证实验的可比性与训练稳定性，所有训练过程在一致的超参数设置下进行。微调超参数设置如表4所示。

参数	值	参数	值
cutoff-len	1024	learning-rate	1.0e-4
finetuning-type	lora	num-train-epochs	1.0
per-device-train-batch-size	8	gradient-accumulation-steps	2
lr-scheduler-type	cosine	warmup-ratio	0.1
lora-rank	64	fp16	True

Table 4: 微调阶段超参数设置

4 实验与分析

4.1 评价指标

由于古汉语处理任务在语义层次、句法结构和语言风格上的多样性与复杂性，为了全面评估多任务微调模型在不同任务中的表现，我们根据各任务的输出特性和语义需

求，选取了适当的评价指标，主要包括精确率（Precision）、召回率（Recall）和F1值（F1 Score）、BLEU、BERTScore以及模型打分机制，具体如下。

对于基础类任务（如句读、词性标注、命名实体识别（NER）、事件识别），我们使用精确率（P）、召回率（R）和F1值来评估模型的预测效果。具体计算方式为：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

其中，TP、FP 和FN 分别表示真阳性、假阳性和假阴性的数量。

对于翻译类任务（如翻译、词语解释），我们采用BLEU值作为评估标准。BLEU通过计算模型输出与参考文本之间的n-gram匹配程度来评估生成质量，如用BLEU-4来衡量4-gram匹配情况，并采用平滑策略应对低匹配率的情况。

针对生成类任务（如反向词典、历史人物知识和诗歌赏析），我们使用BERTScore来评估模型输出的语义质量。BERTScore通过计算生成文本与参考文本在预训练语言模型（如BERT）空间中的相似度，尤其关注语义层面的匹配。我们主要使用BERTScore的F1值来衡量综合语义匹配效果。

对于开放式生成类任务（如诗歌生成），传统的评价指标难以全面反映生成结果的质量。为了更好地对生成内容进行评价，我们采用基于提示词的模型评分机制。具体来说，我们设计了一个多维度评分标准，并通过向模型提供特定的提示（prompt）来为生成的诗歌进行评分。

4.2 实验结果及分析

我们用3.3介绍的各任务的训练集微调模型，进行多任务学习，用测试集进行评估。

4.2.1 多任务总体效果

表 5展示了包括句读、词性标注、命名实体识别（NER）、事件识别、翻译、词语解释、反向词典、历史人物知识、诗歌赏析与诗歌生成十个任务的评测结果。表格中主对角线上的数值代表在单任务微调下的性能结果，其余每一单元格表示“行任务+列任务”联合训练后，在行任务上的评测结果。也就是说，我们将行任务作为主要任务，列任务作为辅助任务。

任务	句读	词性标注	NER	事件识别	翻译	词语解释	反向词典	历史人物	诗歌赏析	诗歌生成
句读	**80.39**	80.25	79.69	80.49	80.25	79.94	79.90	80.36	80.33	80.30
词性标注	89.25	**88.38**	89.65	88.13	89.57	88.99	88.99	88.24	89.42	87.73
NER	91.90	92.23	**92.00**	92.27	91.36	91.87	92.07	91.87	91.09	92.24
事件识别	72.02	73.34	73.46	**73.84**	70.59	72.13	73.78	71.72	66.98	72.17
翻译	46.82	47.29	46.74	47.15	**47.29**	47.01	46.61	47.23	47.12	47.32
词语解释	47.37	48.74	50.92	49.57	46.39	**49.00**	49.32	46.92	37.80	47.74
反向词典	73.07	73.13	72.94	73.30	73.58	73.20	**73.31**	73.56	73.24	73.44
历史人物	75.66	75.51	75.50	75.73	75.40	75.69	75.39	**75.70**	75.37	75.79
诗歌赏析	66.81	66.11	66.60	66.30	66.17	66.10	65.66	65.65	**66.22**	65.79
诗歌生成	63.63	64.23	63.67	63.96	63.69	63.58	62.67	63.88	63.58	**64.37**

Table 5: 各任务单任务与双任务微调下的评测得分（主对角为单任务结果）

从表5可见，多任务联合训练整体上对目标任务性能的增益有限，多数情况下效果低于单任务微调。大多数任务在主对角线（即单任务微调）上取得了最优或次优的得分。例如，在句读任务中，单任务微调的得分为80.39，高于与除事件识别外的其他任务的联合训练得分，即使与事件识别联合时获得了80.49的F1值，也仅比单任务微调多了0.1，差异极小。进一步观察其他任务，如事件识别、翻译、历史人物、诗歌生成等，虽然这些任务在个别搭配下略有波动，但总体上单任务微调结果仍能维持在较优水平，多数联合训练无法带来显著提升。例如，翻译在单任务下表现为47.29，已接近与诗歌生成联合时的47.32，联合训练带来的收益微弱甚至可以忽略。事件识别在单任务下的得分为73.84，诗歌生成在单任务下的得分为64.37，高于所有联合训练设定。整体来看，基于LoRA微调策略的多任务联合方式难以稳定提升下游任务性能，甚至

在部分任务中存在性能退化的现象。这说明古汉语任务之间的语义表示与建模需求存在差异，简单的联合训练可能无法有效共享有益信息，反而引入噪声，影响主任务表现。

4.2.2 协同效应与任务干扰

为更直观、公正地衡量多任务联合训练对各任务性能的影响，本文采用增益率（Gain Rate）作为分析指标。与直接计算绝对得分差值相比，增益率通过将联合训练下的得分变化归一化为相对于单任务微调结果的百分比，能够有效消除不同任务评分区间差异的影响，便于跨任务间的横向比较。此外，增益率能够更准确地反映模型性能波动的相对强度，从而揭示多任务学习中的“协同增益”与“负迁移”现象，尤其适用于分析任务间影响的非对称性。例如，某些任务作为辅助任务时能带来较高的增益率，但在作为主任务时却受到显著干扰，这种非对称性在绝对差值下往往被掩盖。综上，增益率作为评估指标在本研究中更具解释力和分析价值。增益率（Gain Rate）定义如下：

$$\text{增益率} = \frac{\text{双任务得分} - \text{单任务得分}}{\text{单任务得分}} \times 100\% \tag{2}$$

增益率结果如表6所示，增益率热力图如附录B所示。

任务	句读	词性标注	NER	事件识别	翻译	词语解释	反向词典	历史人物	诗歌赏析	诗歌生成
句读	—	-0.18	-0.88	+0.12	-0.18	-0.56	-0.61	-0.04	-0.08	-0.12
词性标注	+0.98	—	+1.43	-0.28	+1.35	+0.69	+0.69	-0.16	+1.18	-0.74
NER	-0.11	+0.26	—	+0.30	-0.69	-0.14	+0.08	-0.14	-0.98	+0.26
事件识别	-2.46	-0.67	-0.51	—	-4.40	-2.31	-0.07	-2.86	-9.28	-2.26
翻译	-0.99	-0.00	-1.17	-0.30	—	-0.59	-1.44	-0.12	-0.36	+0.07
词语解释	-3.33	-0.55	+3.91	+1.16	-5.33	—	+0.64	-4.26	-22.87	-2.57
反向词典	-0.33	-0.25	-0.50	-0.01	+0.37	-0.15	—	+0.34	-0.10	+0.18
历史人物	-0.06	-0.25	-0.27	+0.04	-0.40	-0.02	-0.41	—	-0.44	+0.12
诗歌赏析	+0.90	-0.17	+0.57	+0.12	-0.08	-0.18	-0.85	-0.86	—	-0.65
诗歌生成	-1.14	-0.22	-1.09	-0.63	-1.05	-1.23	-2.64	-0.76	-1.23	—

Table 6: 各任务的相对收益和损失（去除百分号）

在90组联合训练中，仅有23组表现出性能提升，其余67组均出现不同程度的性能下降，如表6所示。任务的结构特征、目标差异和梯度干扰是关键因素。NER与词性标注结合时分别获得了0.26%和0.98%的正增益，词语解释与诗歌赏析联合时增益率分别为-22.87% 和-0.18% 。NER的实体标签如“地点”与词性标签如“ns”存在天然关联，模型能通过联合训练更好地捕捉这类信息，从而提升性能。词语解释要求简洁，诗歌赏析需长文本推理，两者联合训练混淆了生成控制机制。短输出任务通常依赖共享的语言特征，能通过多任务学习强化模型在特定能力上的表现并减少任务间干扰。长输出任务依赖复杂语法结构建模和长距离上下文依赖推理，其目标函数与短输出任务的分类损失存在本质差异，导致模型在参数更新时面临多目标优化困境，共享层难以平衡不同任务的梯度方向。此外，在多任务学习中，长输出任务的计算资源和梯度更新需求较大，这可能与短输出任务产生竞争，进而导致梯度干扰，影响模型的整体训练效果，造成性能下降。

实验结果也展示了多任务学习中协同增益的非对称性，即有些任务作为主任务时能够获得较大的提升，而作为辅助任务时则效果较差。特别是词性标注任务，作为主任务时，加入其他任务后，6个任务产生了正向增益，例如，句读（+0.98%）、NER（+1.43%）和翻译（+1.35%）任务都有明显的性能提升。然而，作为辅助任务时，词性标注对其他任务的增益较小，只有命名实体识别任务在加入词性标注时产生了轻微的正向增益（+0.26%），其他大部分任务表现为负向影响。特别是在与事件识别任务联合时，增益率为-0.67%。这表明，词性标注作为主任务时，其他任务能为其提供更多有价值的特征和信息，而在其他任务作为主任务时，词性标注的辅助作用却较为有限，难以发挥协同作用。这些现象进一步证明了任务间协同增益的非对称性，某些任务作为主任务时能够带来显著的正向影响，而作为辅助任务时则未必能够

第二十四届中国计算语言学大会论文集，第259页-第272页，济南，中国，2025年8月11日至14日。

产生同样的效果。这可能与任务的目标函数和语义特征的差异有关，任务间的协同作用需要在适当的条件下才能显现。

4.2.3 任务类型的影响

结合各任务的类型，我们分析后发现基础类任务（如句读、词性标注、NER、事件识别）在联合训练中表现出更强的协同效应。具体来看，基础类任务之间的组合共涉及6组、12个联合结果，其中有5个结果表现为正向增益，占比达41.7%。这表明此类任务在参数共享过程中能够有效利用彼此的语言结构知识，如边界判断、标签序列等，从而实现性能提升。

相比之下，翻译类任务（如翻译、词语解释）之间的组合结果均为负干扰，表现出较差的协同能力。生成类任务（如反向词典、历史人物知识、诗歌赏析与诗歌生成）之间的组合涉及12个联合结果，其中仅有3个为正增益，占比仅为25%。这说明翻译类和生成类任务在多任务训练中的协同能力较弱，且更容易受到语义目标差异的干扰。例如，诗歌生成作为主任务，与其他任务组合时均为负增益，其中与词语解释联合时下降了-2.57%（见表6）。

此外，基础类任务、翻译类任务和生成类任务之间的混合组合亦难以产生稳定的正向增益，反而时常表现出负迁移现象。具体来说：

- 基础类任务作为辅助任务时，针对翻译类任务，共有8个联合结果，其中2个表现为正向增益，占比25%；1个持平；5个为负向干扰，占比62.5%。此外，针对生成类任务，共有16个组合结果，其中4个为正向增益，占比25%，且正向增益占比不到1%（见表6）。这表明，基础类任务作为辅助任务时可以提升其他任务的性能，但提升效果不明显，增益率不超过25%。可能的原因是任务目标的对抗性或参数共享带来的干扰，导致协同效应不明显。

- 翻译类任务作为辅助任务时，针对基础类任务的8个联合结果中，仅2个为正向增益，正向增益占比为25%。这表明翻译类任务作为辅助任务时，虽然部分组合可能带来增益，但总体协同效果较弱，未能有效提升基础类任务的性能。在与生成类任务联合训练时，所有8个组合结果均为负干扰，负干扰占比为100%。这进一步证明了翻译类任务与生成类任务之间的协同效应较差，尤其是在参数共享过程中，任务之间的目标差异和语义冲突可能会抑制模型的整体表现。

- 生成类任务作为辅助任务时，针对基础类任务的16个组合结果中，4个为正向增益，正向增益占比为25%。尽管整体协同效应较弱，但在某些情况下，生成类任务能够为基础类任务带来一定的性能提升。相比之下，在针对翻译类任务的8个联合结果中，2个为正增益，正向增益占比同样为25%。这一结果与基础类任务的表现类似，进一步表明生成类任务作为辅助任务时，虽然能在某些组合中带来增益，但总体上对翻译类任务的帮助有限，且大多数组合仍表现为负向干扰。

综上所述，任务类型的差异对多任务训练的协同效应具有显著影响。基础类任务表现出较强的协同效应，能够通过共享语言结构知识有效提升性能；而翻译类任务和生成类任务则往往表现出较差的协同能力，尤其是翻译类任务往往导致负向迁移。混合任务组合的表现不稳定，尤其是在不同类型任务间的辅助关系中，负迁移现象时有发生。

4.2.4 任务稳定性与贡献性

在本实验中，我们分析了各任务的单任务得分、在双任务设定下的最大值与最小值及其来源、波动范围（绝对值差）以及标准差，以全面评估任务在多任务设定下的稳定性与鲁棒性。实验结果如表7所示。其中，标准差用于衡量每个任务在不同双任务设定下得分的波动程度和多任务训练对各个任务得分稳定性的影响。具公式如下：

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{3}$$

在公式中：σ：表示标准差（Standard Deviation），n：表示数据点的数量，这里指的是双任务训练设置的数量。x_i：表示第i 次双任务训练下任务的得分。μ：表示该任务在单任务训练下的得分，即任务的基准得分。$\sum_{i=1}^{n}$：表示对所有n 次训练结果进行求和。

任务	单任务	最大值	最大值来源	最小值	最小值来源	绝对值差	标准差
句读	80.39	80.49	+事件识别	79.69	+NER	0.80	0.3338
词性标注	88.38	89.65	+NER	87.73	+诗歌生成	1.92	0.8250
NER	92.00	92.27	+事件识别	91.09	+诗歌赏析	1.18	0.4025
事件识别	73.84	73.84	——	66.98	+诗歌赏析	6.86	2.8167
翻译	47.29	47.32	+诗歌生成	46.61	+反向词典	0.71	0.3529
词语解释	49.00	50.92	+NER	37.80	+诗歌赏析	13.12	4.0169
反向词典	73.31	73.58	+翻译	72.94	+NER	0.64	0.2098
历史人物	75.70	75.79	+诗歌生成	75.37	+诗歌赏析	0.42	0.2062
诗歌赏析	66.22	66.81	+句读	65.65	+历史人物	1.16	0.3881
诗歌生成	64.37	64.37	——	62.67	+反向词典	1.70	0.8204

Table 7: 各任务单任务结果、最大最小值及来源、波动范围、标准差

历史人物和反向词典任务的标准差分别为0.2062和0.2098，远低于其他任务，表明这两个任务在多任务训练中的稳定性较强。它们的得分波动较小，表明这些任务能够较好地适应双任务学习，可能是因为它们与其他任务的干扰较小，或者其本身对外部任务的依赖较少。词语解释任务的标准差为4.0169，显著高于其他任务，说明该任务在多任务学习中的得分波动较大，表现出较低的稳定性。词语解释任务的波动范围也较大（13.12），最小值是与诗歌赏析任务共同训练时产生的，可能是由于诗歌赏析任务数据长度较长，倾向于产生全面详尽的内容，而词语解释任务更倾向于产生简洁明确的回答，在两任务的联合训练中，词语解释任务受到干扰较大，导致其性能难以保持稳定。此外，事件识别任务的标准差为2.8167，波动较大，可能与该任务的复杂性和它与其他任务的相互影响有关。其他任务的标准差分布在0.22到0.83之间，表明它们在多任务训练中的稳定性较好，但波动程度较事件识别、历史人物知识、反向词典要大。

任务的最大值和最小值来源反映了任务间的相互影响。从实验结果来看，事件识别任务不仅在自身的单任务训练中获得了最高得分（73.84），还对其他两个任务（句读、命名实体识别）提供了显著帮助，使这些任务在与事件识别联合训练时达到了各自的最高得分。这表明事件识别在多任务框架中具有较强的“贡献性”，其特征或语义信息在共享中能够提升相关任务的表现，尤其是在与基础类任务的结合中效果更为显著。相应的，诗歌生成任务除自身外，也对一个翻译类任务（翻译）和一个生成类任务（历史人物知识）提供了显著帮助，使这两个任务在与其联合训练时取得了最高得分。这表明，尽管诗歌生成在多任务框架中整体稳定性一般（标准差为0.8204，波动范围为1.70），但在与非基础类任务联合训练时，仍可能通过共享语言建模层的表示学习，增强模型对语言生成规律的掌握，从而带来正迁移效应。

生成类任务在多任务学习中呈现出较强的负向贡献性。在十项最小值中，有八项来自于生成类任务。其中诗歌赏析任务成为多任务组合中成为最频繁的最小值来源任务，命名实体识别（NER）、事件识别、词语解释和历史人物知识在与其联合训练时均取得最低得分。这种“负向迁移”现象在与词语解释任务的组合中表现得尤为突出，得分下降了13.12%，为所有任务组合中的最大降幅，这表明诗歌赏析在多任务框架中可能存在较强的干扰性。

5 总结

本文通过对古汉语多任务学习模型的实验分析，深入探讨了多任务学习中的协同效应、负迁移现象、任务间的非对称性增益以及任务的稳定性和贡献性。研究结果表明，在古汉语任务的多任务学习框架中，任务间的协同增益并非总是正向的，且存在明显的负迁移和非对称性协同增益。基础类任务之间的协同效应更加突出，能够稳定提升模型性能，而翻译类任务之间或生成类任务之间或三种类型任务的混合组合，往往会导致负迁移和性能波动。因此，设计多任务学习模型时，需要根据任务的语义层次和结构特性来合理选择任务组合，从而最大化协同效应、减少负迁移、提升模型稳定性，并有效利用任务间的贡献性差异。

未来，我们将进一步扩充数据集，并探索优化方法，包括任务选择、训练策略和模型架构的改进，以实现更加稳健的多任务学习性能。也将尝试引入古今对齐，利用基座模型对现代汉语的相似任务执行能力来增强古汉语的任务能力。同时，我们计划在不进行增量预训练的情况下直接微调模型，并尝试使用不同的基座模型进行实验，以探索泛化性能更优的优化方法。

第二十四届中国计算语言学大会论文集，第259页-第272页，济南，中国，2025年8月11日至14日。

致谢

感谢所有匿名审稿人的宝贵意见。本研究受国家社会科学基金重大项目（22&ZD035）的支持。

参考文献

Long Huang, Yufeng Peng, Haifeng Wang, et al. 2002. Statistical Part-of-Speech Tagging for Classical Chinese. In *Proceedings of the 5th International Conference on Text, Speech and Dialogue (TSD '02)*, pages 115–122.

Haoxi Huang, Changning Sun, Hsin-Hsi Chen. 2010. Classical Chinese Sentence Segmentation. In *CIPS-SIGHAN Joint Conference on Chinese Language Processing*.

Yue Yuan, Dongbo Wang, Shuiqing Huang, et al. 2019. The Comparative Study of Different Tagging Sets on Entity Extraction of Classical Books. *Data Analysis and Knowledge Discovery*, 3(03):57–65.

Na Li. 2018. Automatic Extraction of Alias in Ancient Local Chronicles Based on Conditional Random Fields. *Journal of Chinese Information Processing*, 32(11):41–48+61.

Hongyu Meng, Qingyu Xie, Hong Chang, et al. 2015. Automatic Identification of TCM Terminology in Shanghan Lun Based on Conditional Random Field. *Journal of Beijing University of Traditional Chinese Medicine*, 38(09):587–590.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, et al. 2017. Unsupervised Machine Translation Using Monolingual Corpora Only. *arXiv preprint arXiv:1711.00043*.

Zichao Yang, Pei Cai, Yuan-Fang Feng, et al. 2019. Generating Classical Chinese Poems from Vernacular Chinese. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6155–6164.

Jiaze Cao, Dongxu Peng, Peng Zhang, et al. 2024. TongGu: Mastering Classical Chinese Understanding with Knowledge-Grounded Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4196–4210.

Shen Li, Renfen Hu, Lijun Wang. 2024. Construction and Application of Ancient Chinese Large Language Model. *Chinese Journal of Language Policy and Planning*, 9(05):22–33.

Dongbo Wang, Chang Liu, Zihe Zhao, et al. 2023. GujiBERT and GujiGPT: Construction of Intelligent Information Processing Foundation Language Models for Ancient Texts. *arXiv preprint arXiv:2307.05354*.

Dongbo Wang, Chang Liu, Zihe Zhu, et al. 2022. Construction and Application of Pre-trained Models of Siku Quanshu in Orientation to Digital Humanities. *Library Tribune*, 42(06):31–43.

Chang Liu, Dongbo Wang, Zihe Zhao, et al. 2024. SikuGPT: A Generative Pre-trained Model for Intelligent Information Processing of Ancient Texts from the Perspective of Digital Humanities. *Journal on Computing and Cultural Heritage*, 17(4):17.

Liang Zhang and Dan Moldovan. 2019. Multi-task Learning for Semantic Relatedness and Textual Entailment. *Journal of Software Engineering and Applications*, 12(6):199–214.

Alex Waibel, Hideki Sawai, and Kiyohiro Shikano. 1989. Modularity and Scaling in Large Phonemic Neural Networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(12):1888–1898.

Melih Kandemir, Akos Vetek, Mehmet Gönen, Arto Klami, and Samuel Kaski. 2014. Multi-task and Multi-view Learning of User State. *Neurocomputing*, 139:97–106.

Natasha Jaques, Ognjen Rudovic, Sarah Taylor, Akane Sano, and Rosalind Picard. 2017. Predicting Tomorrow's Mood, Health, and Stress Level Using Personalized Multitask Learning and Domain Adaptation. In *Proceedings of IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, in *Proceedings of Machine Learning Research*, 66:17–33.

Peng Guo, Chiyuan Lee, and Daniel Ulbricht. 2020. Learning to Branch for Multi-Task Learning. In *Proceedings of the 37th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research*, 119:3854–3863.

Guangquan Lu, Xishun Zhao, Jian Yin, Weiwei Yang, and Bo Li. 2020. Multi-task Learning Using Variational Auto-Encoder for Sentiment Classification. *Pattern Recognition Letters*, 132:115–122.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, et al. 2015. Multi-task Sequence to Sequence Learning. *arXiv preprint arXiv:1511.06114*.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask Learning for Mental Health Conditions with Limited Social Media Data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Chloé Braud, Barbara Plank, and Anders Søgaard. 2016. Multi-view and Multi-task Training of RST Discourse Parsers. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1903–1913, Osaka, Japan. The COLING 2016 Organizing Committee.

Héctor Martínez Alonso and Barbara Plank. 2016. When is Multitask Learning Effective? Semantic Sequence Prediction under Varying Data Conditions. *arXiv preprint arXiv:1612.02251*.

Liu Mou, Zhi Meng, Rui Yan, et al. 2016. How Transferable Are Neural Networks in NLP Applications? *arXiv preprint arXiv:1603.06111*.

Joachim Bingel and Anders Søgaard. 2017. Identifying Beneficial Task Relations for Multi-task Learning in Deep Neural Networks. *arXiv preprint arXiv:1702.08303*.

Michael Crawshaw. 2020. Multi-task Learning with Deep Neural Networks: A Survey. *arXiv preprint arXiv:2009.09796*.

Yang Yang, Dilara Muhtar, Yelong Shen, et al. 2025. MTL-LoRA: Low-Rank Adaptation for Multi-task Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20):22010–22018.

Jiaxin Liu, Yiming Chang, and Yichao Wu. 2025. R-LoRA: Random Initialization of Multi-Head LoRA for Multi-Task Learning. *arXiv preprint arXiv:2502.15455*.

Rich Caruana. 1997. Multitask Learning. *Machine Learning*, 28:41–75.

Hao Guo, Rajarshi Pasunuru, and Mohit Bansal. 2019. AutoSEM: Automatic Task Selection and Mixing in Multi-task Learning. *arXiv preprint arXiv:1904.04153*.

Po-Nien Kung, Sheng-Shiang Yin, Yi-Chen Chen, et al. 2021. Efficient Multi-task Auxiliary Learning: Selecting Auxiliary Data by Feature Similarity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 416–428.

Hong Chen, Xiaonan Wang, Chunyuan Guan, Yichong Liu, and Weinan Zhu. 2022. Auxiliary Learning with Joint Task and Data Scheduling. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 3634–3647.

Étienne Grégoire, Muhammad Hammad Chaudhary, and Stijn Verboven. 2024. Sample-level Weighting for Multi-task Learning with Auxiliary Tasks. *Applied Intelligence*, 54(4):3482–3501.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. *Proceedings of the International Conference on Learning Representations (ICLR)*, 1(2):3.

John Achiam, Sam Adler, Sharan Agarwal, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Alessandro Grattafiori, Ankur Dubey, Ashish Jauhri, et al. 2024. The LLaMA 3 Herd of Models. *arXiv e-prints*, arXiv:2407.21783.

Congcong Wei, Zhenbing Feng, Shutan Huang, Wei Li, and Yanqiu Shao. 2023. CHED: A Cross-Historical Dataset with a Logical Event Schema for Classical Chinese Event Detection. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 875–888, Harbin, China. Chinese Information Processing Society of China.

Aoyang Yang, Bo Yang, Bing Zhang, et al. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Yiren Wang, ChengXiang Zhai, and Hany Hassan. 2020. Multi-task Learning for Multilingual Neural Machine Translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1022–1034, Online. Association for Computational Linguistics.

Carlos Escolano, Graciela Ojeda, Christine Basta, and Marta R. Costa-jussa. 2021. Multi-Task Learning for Improving Gender Accuracy in Neural Machine Translation. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 12–17, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Chunhua Ding, Zhewei Lu, Shikun Wang, et al. 2023. Mitigating Task Interference in Multi-Task Learning via Explicit Task Routing with Non-learnable Primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7756–7765.

Jian He, Kai Li, Yicheng Zang, et al. 2024. Not All Tasks Are Equally Difficult: Multi-task Deep Reinforcement Learning with Dynamic Depth Routing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11):12376–12384.

Xiaolong Peng, Yunchao Wei, Aoran Deng, et al. 2022. Balanced Multimodal Learning via On-the-fly Gradient Modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8238–8247.

Zhiwei Zhang, Jie Shen, Chenghao Cao, et al. 2024. Proactive Gradient Conflict Mitigation in Multi-Task Learning: A Sparse Training Perspective. *arXiv preprint arXiv:2411.18615*.

Guangquan Lu, Jiwei Gan, Jun Yin, et al. 2020. Multi-task Learning Using a Hybrid Representation for Text Classification. *Neural Computing and Applications*, 32(11):6467–6480.

Zheng Zhan and Rui Zhang. 2024. Towards Better Multi-task Learning: A Framework for Optimizing Dataset Combinations in Large Language Models. *arXiv preprint arXiv:2412.11455*.

Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2023. A Survey of Multi-task Learning in Natural Language Processing: Regarding Task Relatedness and Training Methods. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 943–956, Dubrovnik, Croatia. Association for Computational Linguistics.

Peng Lu, Ting Bai, and Philippe Langlais. 2019. SC-LSTM: Learning Task-Specific Representations in Multi-Task Learning for Sequence Labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2396–2406, Minneapolis, Minnesota. Association for Computational Linguistics.

Kai Cao and Dániel Yogatama. 2020. Modelling Latent Skills for Multitask Language Generation. *arXiv preprint arXiv:2002.09543*.

Xinyu Yao, Mengdi Wang, Bo Chen, and Xiaobing Zhao. 2025. WenyanGPT: A Large Language Model for Classical Chinese Tasks. In *arXiv preprint arXiv:2504.20609*.

A 十任务联合训练

我们进一步探索了同时引入全部10个任务进行联合训练的效果，采用与双任务相同的超参数设置及微调数据集，十任务联合微调结果如表 8所示。在大部分任务上，十任务训练的性能略有下降，这与双任务训练结果基本一致。其中，词语解释任务的下降幅度非常大（-29.02），这表明其他任务可能会对它产生较大的干扰。结合双任务训练表 6的实验结果，词语解释和诗歌赏析任务的组合带来了最大负增益率（22.87），这也进一步验证了这两个任务间可能存在较大的干扰。在十任务联合训练时，词性标注是唯一一个表现出提升的任务，而在双任务训练时，如表 6所示，词性标注与其他6个任务结合时都有提升，而其他任务最多与4个任务结合时取得了提升，这表明它在多任务学习中具有较强的稳定性和鲁棒性。

任务	句读	词性标注	NER	事件识别	翻译	词语解释	反向词典	历史人物	诗歌赏析	诗歌生成
十任务	78.78	89.22	91.19	70.64	46.38	19.98	72.00	75.29	65.72	62.66
单任务	80.39	88.38	92.00	73.84	47.29	49.00	73.31	75.70	66.22	64.37
Δ	-1.61	0.84	-0.81	-3.20	-0.91	-29.02	-1.31	-0.41	-0.50	-1.71

Table 8: 各任务十任务与单任务微调下的评测得分

B 任务增益率热力图

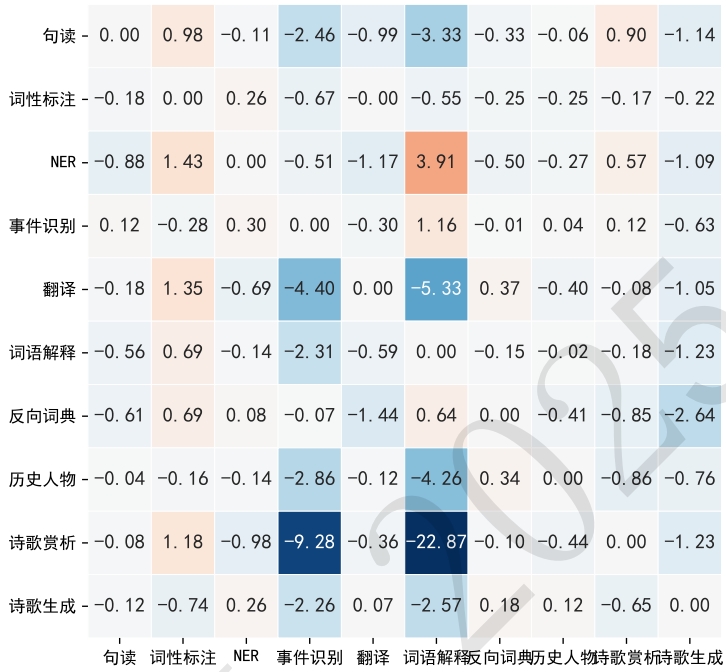

Figure 4: 任务增益率热力图