文言 GPT：用于古代汉语任务的大型语言模型

Xinyu Yao[1], Mengdi Wang[1], Bo Chen[1,2], Xiaobing Zhao[1,2]

[1]Minzu University of China,

[2]National Language Resources Monitoring & Research Center of Languages

chenbomuc@muc.edu.cn

Abstract

作为中华文化的核心载体，古代汉语在古籍传承和研究中发挥了关键作用。然而，现有的自然语言处理模型主要针对现代汉语进行优化，导致对古代汉语的表现不足。本文提出了一种全面的古代汉语语言处理解决方案。通过在 LLaMA3-8B-Chinese 模型上继续预训练和指令微调，我们构建了一个名为 WenyanGPT [1] 的大型语言模型，专为古代汉语任务设计。此外，我们开发了一个评估基准数据集，WenyanBENCH [2]。在 WenyanBENCH 上的实验结果表明，WenyanGPT 在各种古代汉语任务中显著优于当前的先进大语言模型。我们将模型的训练数据、指令微调数据 [3] 和评估基准数据集公开，以促进古代汉语处理领域的进一步研究和发展。

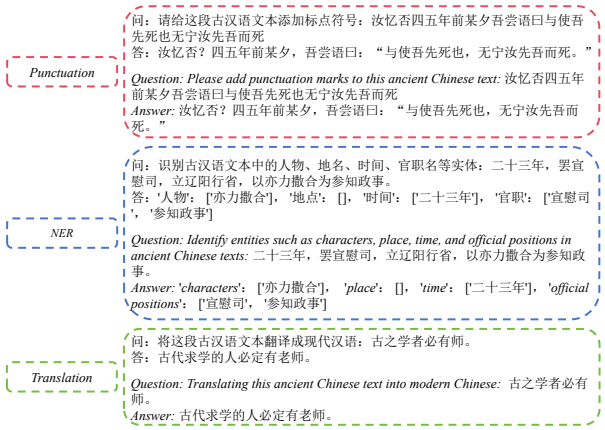

Figure 1: 来自文言 GPT 的任务示例。该模型表现出对文言文的先进知识，并在文言文理解和生成任务中展现出强大的性能。

1 介绍

文言文是中华文化的重要组成部分，具有悠久的历史和深厚的文化底蕴。理解传统中国文化离不开文言文。随着人工智能技术的快速发展，古代汉文本的智能处理为文言文的保存和传承提供了新的解决方案。现代技术，如数字化和自然语言处理，能够高效地保存和传播传统文化，同时促进更深入和创新的学术研究。最大限度地发挥人工智能在处理文言文本中的潜力，已成为文化传承和学术发展的迫切需要。

早期的古代汉语处理研究集中于标点、分词、词性标注、命名实体识别和翻译等任务。这些任务最初依赖于传统的机器学习方法，例如词性标注使用隐马尔科夫模型（HMMs）(Huang et al., 2002)，标点和命名实体识别使用条件随机场（CRFs）(Huang et al., 2010) (Yuan et al., 2019; Li, 2018)。在深度学习领域，RNNs、LSTMs、GRUs 以及注意力机制已经应用于各种任务，包括对联生成和古典诗歌生成 (Yan et al., 2016; Yi et al., 2017)，以及使用 BiLSTM-CRF 模型标点和词性标注 (Wang et al., 2019; Cheng et al., 2020; Zhang et al., 2023b; Chang et al., 2024)。随着 Transformer 架构的兴起 (Vaswani et al., 2017)，研究开始使用大规模平行语料库来训练模型，将古代汉语翻译成现代汉语 (Liu et al., 2018)，并生成古典诗歌 (Huang et al., 2020)。预训练模型的引入，包括 BERT (Kenton and Toutanova, 2019) 和 GPT (Radford and Narasimhan, 2018)，为智能的古代汉语处理提供了新的机会。有些研究将古代汉文本整合到通用预训练模型的训练数据中，提高了古代汉语处理性能，相较于典型预训练模型 (Tian et al., 2020; Wang et al., 2022, 2023a; Liu et al., 2023a)。其他研究则利用古代汉语语料库继续预训练并微调大型语言模型，旨在构建古代汉语的会话模型 (Zhang et al., 2024; Yang et al., 2024b; Cao et al., 2023, 2024)。

然而，文言文处理仍然面临挑战。不同的任务需要训练专门的模型，且没有有效的通用模型。此外，在该领域缺乏标准化的评估基准，现有的评估任务、数据集和指标不一致，使得跨任务比较和系统评估模型性能变得困难。

为了解决这些问题，我们提出了 WenyanGPT，一种古文大规模语言模型。WenyanGPT 的一些示例见图 1。我们还构建了

[1]https://huggingface.co/Wenyanmuc/WenyanGPT
[2]https://github.com/Wenyanmuc/WenyanBENCH
[3]https://github.com/Wenyanmuc/WenyanGPT

可用于持续预训练的最大可用预训练语料库，提高模型的领域适应性。此外，我们提出了一种框架，用于在 WenyanGPT 的开发过程中生成领域特定的指令数据进行监督微调。为了促进古文智能处理的研究，我们构建了 Wenyan-BENCH 评估数据集，并进行了广泛的实验以进行详细分析。主要贡献如下：

- 我们提出了 WenyanGPT，这是一个专注于文言文的大型语言模型。它在标点符号、词性标注、翻译等任务中展示了卓越的性能和广泛的适用性。
- 我们发布了预训练和指令微调的数据集，以及一种用于构建领域特定微调数据的新方法，为未来的研究提供了宝贵的资源。
- 我们引入了 WenyanBENCH，这是一个用于古代汉语任务的评估基准，通过广泛的实验证实 WenyanGPT 在多个任务中具有领先的表现。

2 相关工作

2.1 预训练语言模型

2017 年，谷歌引入了一种新的神经网络架构，Transformer。它利用自注意机制更好地处理远距离的依赖关系，并通过并行计算显著提高了训练效率。在 Transformer 的基础上，各种大型语言模型（LLM）被提出。BERT 采用仅编码器的 Transformer 架构，并通过掩码语言建模和下一个句子预测任务进行预训练。而 GPT 系列则使用仅解码器的 Transformer 架构和自回归语言模型（ALM）。在 GPT 系列的发展过程中，模型规模稳步增长，从最初的 GPT 到随后的迭代，包括 GPT-2 (Radford et al., 2019) 、GPT-3 (Brown et al., 2020) 和 GPT-4 (Achiam et al., 2023) ，性能不断提高。PaLM (Chowdhery et al., 2023) 在仅解码器模型中使用了标准 Transformer 架构，并采用修改的 SwiGLU 激活函数。该模型拥有 5400 亿个参数，在 BIG-bench (Srivastava et al., 2023) 数据集上的一次性学习中达到了人类水平的表现。2023 年，Meta AI 发布了 LLaMA 模型 (Touvron et al., 2023) 。该模型也遵循仅解码器的 Transformer 架构，并在大规模训练后在各种自然语言处理任务中表现出色。2024 年，LLaMA 3 (Dubey et al., 2024) 发布，包括一个拥有 4050 亿参数的预训练版本和一个后训练版本，以及用于输入输出安全的 LLaMA Guard 3 模型。预训练语言模型发展迅速，基于 Transformer 的模型已成为自然语言处理（NLP）中的主流技术。

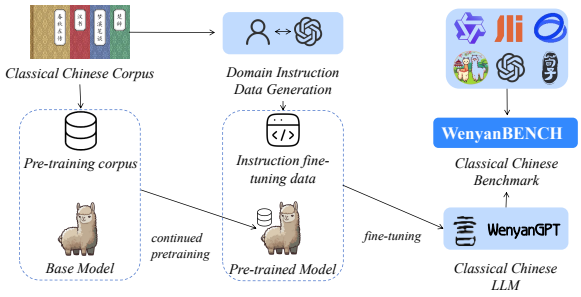

Figure 2: 文言 GPT 的整体训练框架。

2.2 古典中文预训练语言模型

预训练语言模型在自然语言处理领域取得了广泛成功。然而，研究表明，通用领域模型通常缺乏执行特定领域任务的专业知识。使用特定领域数据进行预训练的模型往往在专用任务上表现更好 (Ke et al., 2023; Gupta et al., 2023; Ibrahim et al., 2024; Taylor et al., 2022; Lehman et al., 2023; Liu et al., 2020) 。在古代汉语领域，几项研究通过整合古汉语语料库进行预训练，扩展了如 BERT、RoBERTa 和 GPT 等模型，从而产生了像 AnchiBERT (Tian et al., 2020) 、四库 BERT 和四库 RoBERTa (Wang et al., 2022) 、古籍 BERT 和古籍 GPT 系列 (Wang et al., 2023a) 以及四库 GPT (Liu et al., 2023a) 等专业模型。这些模型在古代汉语任务中较通用预训练模型显示出更好的性能。指令微调是另一种有效策略。使用监督微调 (SFT) 可以激活大型语言模型在特定领域理解并回答问题的能力 (Liu et al., 2023b; Xiong et al., 2023; Wang et al., 2023b; Yue et al., 2023; Huang et al., 2023; Cui et al., 2023; Yang et al., 2023b; Zhang et al., 2023a; Dan et al., 2023) 。古代汉语大型语言模型 (Zhang et al., 2024; Yang et al., 2024b; Cao et al., 2023, 2024) 处于发展的早期阶段。例如，“荀子[4]”古代汉语大语言模型是基于 Qwen2.5 (Yang et al., 2024a) 、百川 2 (Yang et al., 2023a) 和 GLM-4 (Zeng et al., 2024) 等通用模型，以古代汉语相关语料库为训练数据的。它在智能标注、信息提取等任务上表现出色。通过两阶段指令微调，通古 (Cao et al., 2024) 具备古代汉语标点、翻译和鉴赏任务的能力。在本文中，文言 GPT 在质量更高的预训练数据和更大、更具多样性的指令数据集上进行了微调，显示出更优越和更全面的任务处理能力。

为了获得 WenyanGPT 文言文模型，我们首先构建一个文言文预训练语料库，并基于 LLaMA3-8B-中文继续预训练（第 3.1 节）。然

[4] https://github.com/Xunzi-LLM-of-Chinese-classics/XunziALLM.

Source	Scale	Source	Scale
Daizhige	5.2G	Poetry-master	323M
wenyanguji.com	1.6G	PoetrySplider	16M
network resource	1.1G	poems-db	660M
TCM-Ancient-Books	322M	core-texts	232M
chinese-novel	294M	sidamingzhi	6.7M
chinese-gushiwen	23M	chtxt-main	88M
Classical-Chinese	208M	chinese-poetry	115M
Classical-Modern	853M	guner2023	63M
core-books-main	752M	kangxi-master	37M
GuWen-master	2.5M	scripta-sinica	3.7G

Table 1: 古典中文预训练语料库的来源和规模。

后，我们提出了一种构建领域指令数据的方法（第 2.4 节）。在我们的框架中，指令生成是手动构建的，由大语言模型引导，并经过测试以确保微调数据的高质量。完整的训练过程如图 2 所示。

2.3 预训练

Hyper parameter	Value
per device train batch size	16
gradient accumulation steps	1
learning rate	1.0e-4
num train epochs	1
lr scheduler type	cosine
warmup ratio	0.1

Table 2: 继续预训练中的超参数设置。

预训练阶段使用的语料库来源于道藏阁、文言古籍等权威网站，以及从 GitHub 收集和整理的各类文言文相关数据。详细的数据来源和规模如表 1 所示。我们将这些不同来源的数据统一格式化并存储，去除冗余信息、错误、特殊符号以及无效字符。最终，我们获得了一个干净、大规模、高质量的文言文语料库，约 16GB。该语料库涵盖了四书五经，包括儒家经典、历史记录、诸子百家作品、诗歌、散文、戏剧、小说、杂记等其他文学体裁。它还包括了地方志、谱牒、宗教文本、农业、法律、医学、天文学、地理、工艺书籍和军事文本等多领域内容。语料库整合了简体和繁体汉字，时间跨度从先秦时期到中华民国，为古代汉语文本的深度学习和研究提供了丰富的材料。我们选择 LLaMA3-8B-Chinese 作为基础模型，并在训练中使用 bfloat16 数据格式以提高效率。预训练中的超参数设置如表 2 所示。

2.4 监督微调

基于持续的预训练，我们执行监督微调，以更好地使模型适应特定任务和指令。我们使用了一套高质量的指令微调数据，这是我们先前收集和组织的，用来引发模型在预训练过程中获得的知识。构建指令微调数据的详细过程如图 3 所示。

数据选择和初步组织。 我们从古典汉语语料库中选择相关数据，包括三个主要类别：问答对、平行语料和标注语料。平行语料用于翻译和解释任务，而标注语料支持细粒度任务，如标点和词性标注。在语料库缺乏明确问答对的情况下，我们会生成补充数据。在选择阶段，我们优先选择具有清晰内容、标准化语义和与任务相关的高质量数据，以建立初始的高质量输入输出对。

任务指令的手动设计和模型扩展。 我们根据高质量的输入和输出手动设计初始任务指令模板，涵盖的任务包括文言文标点和将文言文翻译成现代汉语。然后，我们使用大型语言模型（LLMs），如 GPT 和 Qwen 系列，来扩展任务指令。一方面，我们从现有指令生成多样化的指令；另一方面，我们允许 LLMs 从现有的高质量输入-输出对进行逆向推理以生成新的指令，确保指令的多样性。在扩展之后，我们对生成的指令进行初步筛选，去除语义不清或不合理的指令，从而形成种子指令集。

测试指令集和优化微调数据。 我们随机选择高质量的输入输出对，并将它们与种子指令集结合，以评估模型在不同任务场景中对指令的遵从性。我们分析测试结果，以优化指令设计并识别出产生高质量输出的任务指令。随后，我们将优化后的指令集与高质量的输入输出对结合，构建特定指令的微调数据，为后续的模型训练提供可靠的数据支持。

生成和补充特定任务数据。 由于语料库中缺乏具体任务指令数据，我们手动设计了多样化的初始指令和输入。然后我们使用包含 Qwen2.5-14B 和 Qwen2.5-72B 在内的大型语言模型生成高质量的答案。我们通过人工选择和自动质量检查移除低质量或不相关的内容，形成另一部分指令数据。

我们整合了从语料库生成的数据和来自大型语言模型的补充数据以形成一个完整的指令数据集。在整合过程中，我们确保指令与输出之间的一致性，涵盖多个文言文任务场景。最后，通过全面的数据验证和优化，我们创建了一个高质量的指令数据集。最终，我们获得了大约 185 万条指令微调数据。详细的数据来源和统

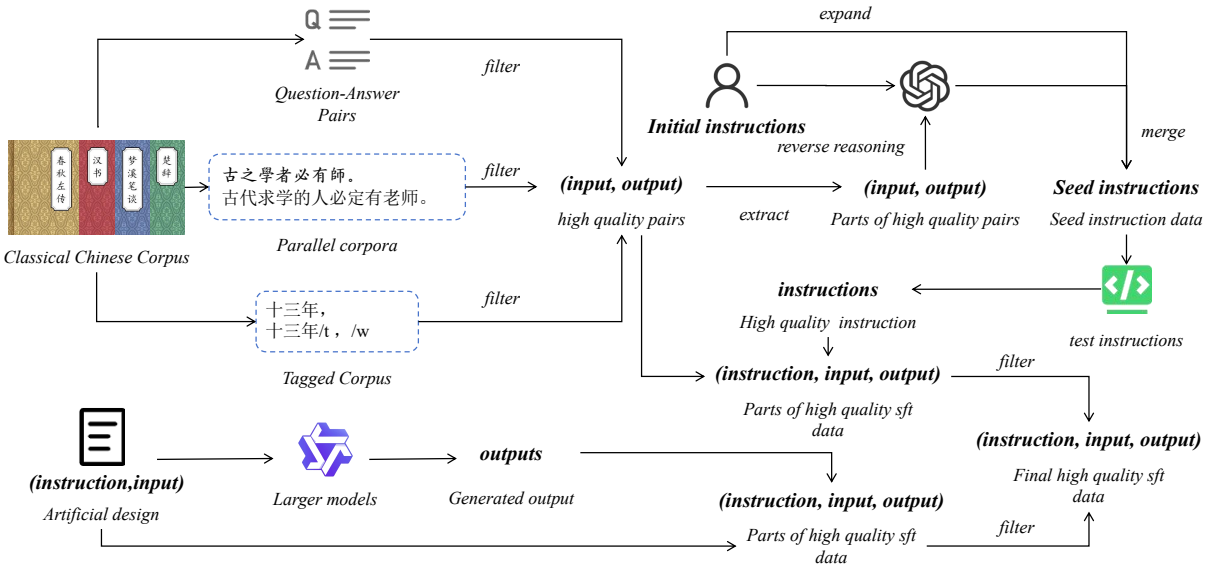

Figure 3: 指令微调数据构建过程。

Task	Data Source	Num
Punctuation	Daizhige	107,3017
Part-of-speech tagging	evahan	9,952
NER	Self-built	29923
Translation	classical-modern	222,700
	wenyanguji.com	302,724
Word explanation	gushiwen.com	31,088
Reverse dictionary	chinese-dictionary	39,708
	chinese-xinhua	138,810
Total		1,847,922

Table 3: 微调数据的来源和规模。

Hyper parameter	Value
per device train batch size	8
gradient accumulation steps	2
learning rate	1.0e-4
num train epochs	1
lr scheduler type	cosine
warmup ratio	0.1

Table 4: 微调中的超参数设置。

计如表 3 所示。我们使用这些数据来微调预训练模型。微调中的超参数设置如表 4 所示。

3 古文任务的基准测试

为了评估模型在古文任务上的表现，我们设计了一个名为 WenyanBench 的基准。WenyanBench 与指令微调数据共享同样的数据来源，并经过了重复数据去除以及人工和 LLM 的验证。为了质量控制，我们对数据的一个子集进行了抽样。WenyanBench 的分布和详细统计信息如表 5 所示。

Task	Num
Punctuation	7,559
Part-of-speech tagging	1,247
NER	3,741
Translation	5,013
Word Explanation	3,931
Reverse Dictionary	4,462
Total	25,953

Table 5: 文言基准的数据来源和详细统计

我们的基准测试包括六项与古典中文相关的任务。其中，在标点符号任务中，我们将 14 种类型的标点符号进行细分；在词性标注任务中，我们将古代汉语的词类划分为 17 个类别；在命名实体识别任务中，我们定义了 4 个类别。

对于 WenyanBench 基准测试，不同类型的任务（理解任务和生成任务）使用不同的评估指标。对于理解任务，评估主要依赖于精确度、召回率和 F1 评分。对于生成任务，BLEU 和 BERT 评分被用作评估指标。BLEU 衡量生成内容与参考答案之间的 N-gram 重叠，而 BERT 评分更好地捕捉生成内容与参考答案之间的语义相似性。

评价方法。 为了有效评估模型性能，我们设计了一套脚本化工具来自动计算 BLEU、BERT-Score 和其他指标。这些工具可以快速准确地量化模型输出，为模型优化提供明确反馈。这种自动化评估方法提高了评估效率，并确保结果的一致性和可比性。

Model	Punctuation			Part-of-speech tagging			NER		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Qwen2.5-7B-Instruct	54.34	53.31	53.82	51.25	48.16	49.65	66.05	46.55	54.61
Baichuan2-7B-Chat	51.05	21.03	29.79	47.11	30.97	37.37	35.10	10.58	16.26
GLM-4-9B-Chat	52.39	55.00	53.66	49.90	54.98	52.32	52.19	45.42	48.57
Meta-Llama-3-8B-Instruct	55.05	22.41	31.85	25.73	17.06	20.52	47.50	57.48	52.01
Llama3-8B-Chinese-Chat	45.76	38.07	41.56	21.34	19.34	20.29	46.85	66.69	55.04
Xunzi-Qwen-1.5-7B-Chat	52.08	47.19	49.51	77.54	78.07	77.81	49.79	51.21	50.49
GPT-4o	52.00	50.70	51.34	82.41	81.11	81.75	61.58	76.97	68.42
Deepseek-V3	56.33	61.94	59.01	79.12	79.18	79.15	56.83	79.75	66.36
WenyanGPT	76.84	74.52	75.66	89.66	88.54	89.1	92.14	90.19	91.16

Table 6: 关于 WenyanBench 的理解任务（标点符号、词性标注、命名实体识别）结果。下划线部分代表第二好的模型的 F1 得分。

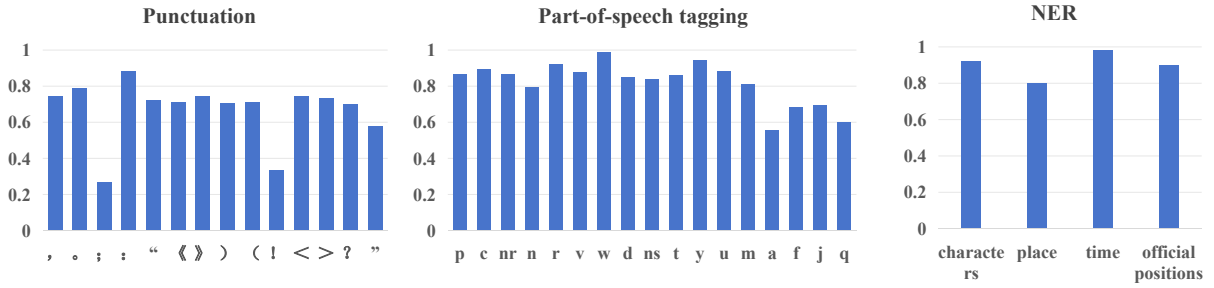

Figure 4: 文言 GPT 在 WenyanBench 上的理解任务子类别（包括标点符号、词性标注和命名实体识别）的 F1 得分。

4 实验

实验评估了 WenyanGPT 在古文理解和生成任务中的表现。

4.1 实验设置

基线。 基线包括通用领域和文言文领域的大型语言模型。通用领域的大型语言模型有 Qwen2.5-7B-Instruct, Baichuan2-7B-Chat, GLM-4-9B-Chat, Meta-Llama-3-8B-Instruct, Llama3-8B-Chinese-Chat, GPT-4o, 和 Deepseek-V3 (DeepSeek-AI et al., 2025) 。文言文领域的大型语言模型是 Xunzi-Qwen1.5-7B-Chat。

数据和评估。 我们使用 WenyanBench 基准进行测试。理解任务包括标点、词性标注和命名实体识别，这些任务通过准确率、召回率和 F1 分数进行评估。生成任务包括词语解释、翻译和反向词典，其中 BLEU 用于词语解释和翻译，而 BERT-Score 用于反向词典。

4.2 实验分析

实验任务理解的结果在表 6 中展示。在命名实体识别任务中，WenyanGPT 的准确率、召回率和 F1 得分都超过了 90 % ，而第二好的模型

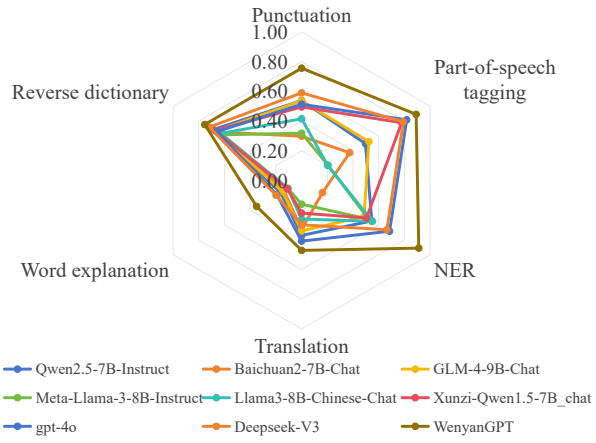

Figure 5: 雷达图显示模型在 WenyanBench 上的性能，值被归一化到 0-1 的尺度。

GPT-4o 在这些指标中均未能超过 80 % 。在标点任务中，WenyanGPT 的 F1 得分比第二好的模型 Deepseek-V3 高 16.65 % ，达到 75.66 % 。此外，在词性标注任务中，WenyanGPT 的 F1 得分比第二好的模型 GPT-4o 高 7.35 % 。这些结果突出了 WenyanGPT 在理解任务中的明显优势，特别是在命名实体识别和标点方面。这种表现归因于模型在古文数据上的广泛预训练，使其能够处理复杂的语言现象，确保在基

Model	Translation				Word explanation				Reverse dictionary		
	Bleu1	Bleu2	Bleu3	Bleu4	Bleu1	Bleu2	Bleu3	Bleu4	P(%)	R(%)	F1(%)
Qwen2.5-7B-Instruct	0.37	0.23	0.17	0.14	0.16	0.09	0.07	0.05	68.43	68.99	68.66
Baichuan2-7B-Chat	0.33	0.20	0.14	0.11	0.14	0.08	0.05	0.04	64.81	66.21	65.42
GLM-4-9B-Chat	0.34	0.21	0.15	0.12	0.15	0.09	0.06	0.05	65.58	68.04	66.69
Meta-Llama-3-8B-Instruct	0.16	0.09	0.06	0.05	0.11	0.06	0.05	0.04	59.41	64.13	61.48
Llama3-8B-Chinese-Chat	0.26	0.15	0.10	0.08	0.11	0.06	0.04	0.03	61.8	65.18	63.28
Xunzi-Qwen1.5-7B-Chat	0.22	0.15	0.11	0.09	0.11	0.08	0.06	0.05	66.47	68.45	67.35
GPT-4o	0.41	0.27	0.19	0.14	0.19	0.13	0.09	0.07	64.96	66.76	65.81
Deepseek-V3	0.30	0.19	0.13	0.10	0.20	0.14	0.11	0.08	71.93	71.84	71.88
WenyanGPT	0.47	0.33	0.24	0.19	0.35	0.31	0.27	0.23	75.51	75.31	75.39

Table 7: WenyanBench 上的生成任务（翻译、词汇解释、反向词典）结果。下划线标注的结果代表第二好的模型的 BLEU1 分数和 BERT-Score-F1 分数。

本语言理解任务中具有更高的准确性和稳定性，例如词性标注和命名实体识别。

如图 4 所示，我们模型在三个任务的子类别中的 F1 分数总体上是稳定且较高的。具体而言，在命名实体识别任务中，WenyanGPT 的 F1 分数保持在 80 % 以上，显示出其在正确识别古代汉语文本中的历史人物、地点和专有名词等实体方面的强大能力。这一表现表明它在处理古代汉语文本时具有较强的准确性和稳健性，并能够有效地捕捉复杂的上下文关系和词义变化。总体而言，WenyanGPT 在古代汉语理解任务中的高 F1 分数不仅反映了它在基础任务中的高效性，还展示了它在处理古代汉语中的细粒度任务时的优势和潜力。

WenyanGPT 可以显著提高生成任务中内容的质量。 实验结果显示在表格 7 中。WenyanGPT 在语义保持和上下文一致性方面表现出色，并在翻译和单词解释任务中表现出优越和更稳定的性能，且保持一致的高 BLEU 分数（BLEU1-BLEU4）。在古文翻译中，模型的 BLEU1 分数为 0.47，比第二好的模型高 0.06。此外，WenyanGPT 在反向词典任务中的 F1 分数比第二好的模型高出 3.47 %。通过缜密的预训练和多任务训练，WenyanGPT 发展出强大的上下文一致性，能够生成准确反映预期意义和上下文的内容。这一能力确保了在长文本生成和复杂任务（如单词解释）中高质量的语义传递和内容连贯性。

WenyanGPT 在古文任务中可以表现出比之前的开源大型语言模型更好的性能。 WenyanGPT 通过大规模预训练和多任务联合优化，在古文处理任务中展现出了显著的能力，明显领先于现有的主流大规模语言模型。这验证了所提出方法的有效性。图 5 显示了各种大规模语言模型在 WenyanBench 上的表现。可以看出，WenyanGPT 在六项古文任务中取得了最高分。WenyanGPT 的多任务训练策略整合了各种古文处理任务，增强了模型在任务间的学习能力。任务间共享的语言特征和语义信息的相互强化显著提升了模型的泛化能力。这种任务间的协同作用不仅提高了单项任务的表现，还使 WenyanGPT 能够有效地同时处理多项任务，保持稳定表现，尤其在复杂任务中，展示了任务间的卓越适应能力。

4.3 案例研究

我们提供了来自五个大语言模型在词性标注上的响应示例：WenyanGPT, Deepseek-V3, GPT-4o (Hurst et al., 2024) , Qwen2.5-7B-Instruct，以及 Xunzi-Qwen1.5-7B-Chat，如表 8 所示。通过分析这些示例，我们可以看到不同大语言模型在访问任务中的性能差异，尤其是在古代汉语理解任务中的表现。

表 8 展示了词性标注任务中的一些典型错误。具体来说，GPT-4o 在标注时间词和专有名词时出现错误。例如，它错误地将时间表达 “UTF8gbsn 四年 ”（第四年）标注为普通名词，同时也难以识别专有名词 “UTF8gbsn 🄵 州吁 ”（魏周许）。Qwen2.5-7B-Instruct 在标注词性 “UTF8gbsn 春 ”（春天）时出错，有时会用不合适的替代字符替换原文中的字符。荀子-Qwen1.5-7B-Chat 主要是无法区分时间词和名词，错误地使用了简化字 “UTF8gbsn 杀 ” 代替正确的 “UTF8gbsn 弒 ”。这些错误突显了该模型在精确区分相似词汇方面的困难。

文言 GPT 在处理文言文任务时表现出强大的语义理解和记忆能力。其准确标注词性和生成简洁而富有诗意的文字解释的能力展示了对文言文细微差别的深刻掌握。相比之下，其他大型语言模型在理解文言文的细微区别和生成忠实于原文内容和意义的回应方面表现不佳。

Case 1	Please segment the following Classical Chinese content and accurately tag the parts of speech: UTF8gbsn 四年春，🄵州吁弒桓公而立。
Ground Truth	UTF8gbsn 四年/t 春/n ，/w 🄵州吁/nr 弒/v 桓公/nr 而/c 立/v 。/w
WenyanGPT	UTF8gbsn 四年/t 春/n ，/w 🄵州吁/nr 弒/v 桓公/nr 而/c 立/v 。/w
Deepseek-V3	UTF8gbsn四/m 年/t 春/t ，/w 卫 州吁/nr 弒/v 桓公/nr 而/c 立/v 。/w
GPT-4o	UTF8gbsnfour years/n 春/n ，/w 卫州/n 吁/v 弒/v 桓公/n 而/c 立/v 。/w
Qwen2.5-7B-Instruct	UTF8gbsn 四年/t 春/w ，/w 卫州吁/nr 射/v 桓公/nr 而/c 立/v 。/w
Xunzi-Qwen1.5-7B-Chat	UTF8gbsnfour-year spring/t ，/w 🄵州吁/nr 杀/v 桓公/nr 而/c 立/v 。/w

Table 8: 来自不同大型语言模型的词性标注任务的响应示例。词性标注中的错误用红色标记，而文本错误用蓝色突出显示。

这使得文言 GPT 成为处理文言文领域复杂任务的强大工具，其在准确性和文学性方面远远超过其他大型语言模型。

5 结论

我们提出了一个针对文言文语言处理挑战的全面解决方案，包括开发了专注于文言文领域的大型语言模型 WenyanGPT，以及用于文言文任务的评估基准数据集 WenyanBENCH。我们发布了预训练和指令微调数据集，并描述了构建指令微调数据集的方法。通过系统实验和分析，我们展示了领域特定预训练和多任务指令微调对提高文言文处理能力的显著影响。我们的模型在各种下游任务中优于现有主流大型语言模型。未来，我们计划通过结合文言文文本与图像数据（如铭文和手稿）来增强处理能力，从而探索多模态模型的潜力。

6

限制

虽然 WenyanGPT 在文言文任务中取得了一些进展，但仍然存在一些局限性。首先，由于篇幅限制和评估的主观性（人工评估），例如诗歌生成等任务未被纳入本文。其次，模型主要依赖大规模、高质量的指令数据集。最后，在处理长篇文言文文本和复杂句法时，仍有改进的空间。

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jiahuan Cao, Dezhi Peng, Yongxin Shi, Zongyuan Jiang, and Lianwen Jin. 2023. Translating ancient chinese to modern chinese at scale: A large language model-based approach. In *International Conference on Algorithmic Learning Theory*.

Jiahuan Cao, Dezhi Peng, Peirong Zhang, Yongxin Shi, and 1 others. 2024. Tonggu: Mastering classical chinese understanding with knowledge-grounded large language models. In *Conference on Empirical Methods in Natural Language Processing*.

Bolin Chang, Yiguo Yuan, Bin Li, Zhixing Xu, and 1 others. 2024. Automatic word segmentation and part-of-speech tagging for classical chinese based on radicals. *Data Analysis and Knowledge Discovery*, 8(11):102–113.

Ning Cheng, Bin Li, Liming Xiao, Changwei Xu, and 1 others. 2020. Integration of automatic sentence segmentation and lexical analysis of ancient chinese based on bilstm-crf model. In *Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 52–58.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*.

Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, and 1 others. 2023. Educhat: A large-scale language model-based chatbot system for intelligent education. *Preprint*, arXiv:2308.02773.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kshitij Gupta, Benjamin Th'erien, Adam Ibrahim, Mats L. Richter, Quentin G. Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re)warm your model? *ArXiv*, abs/2308.04014.

Chuen-Min Huang, Kuo-Lin Lu, Yi-Ying Cheng, and Yu-Chen Peng. 2020. Generating chinese classical poetry with quatrain generation model (qgm) using encoder-decoder lstm. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 5700–5702. IEEE.

Hen-Hsen Huang, Chuen-Tsai Sun, and Hsin-Hsi Chen. 2010. Classical chinese sentence segmentation. In *CIPS-SIGHAN Joint Conference on Chinese Language Processing, Beijing, China, August 28-29, 2010*.

Liang Huang, Yinan Peng, Huan Wang, and Zhenyu Wu. 2002. Statistical part-of-speech tagging for classical chinese. In *Proceedings of the 5th International Conference on Text, Speech and Dialogue*, TSD '02, page 115–122, Berlin, Heidelberg. Springer-Verlag.

Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Adam Ibrahim, Benjamin Th'erien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models. *ArXiv*, abs/2403.08763.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bin Liu. 2023. Continual pre-training of language models. In *International Conference on Learning Representations*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models? *ArXiv*, abs/2302.08091.

N Li. 2018. Automatic extraction of alias in ancient local chronicles based on conditional random fields. *J. Chin. Inf. Process*, 32:41.

Chang Liu, Dongbo Wang, Zhixiao Zhao, Die Hu, Mengcheng Wu, Litao Lin, Si Shen, Bin Li, Jiangfeng Liu, Hai Zhang, and Lianzheng Zhao. 2023a. Sikugpt: A generative pre-trained model for intelligent information processing of ancient texts from the perspective of digital humanities. *ArXiv*, abs/2304.07778.

Dayiheng Liu, Jiancheng Lv, Kexin Yang, and Qian Qu. 2018. Ancient–modern chinese translation with a new large training dataset. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19:1 – 13.

June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023b. Chatcounselor: A large language models for mental health support. *ArXiv*, abs/2309.15461.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *International Joint Conference on Artificial Intelligence*.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, and 1 others. 2019. Language models are unsupervised multitask learners.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Trans. Mach. Learn. Res.*, 2023.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, and 1 others. 2022. Galactica: A large language model for science. *ArXiv*, abs/2211.09085.

Huishuang Tian, Kexin Yang, Dayiheng Liu, and Jiancheng Lv. 2020. Anchibert: A pre-trained model for ancient chinese language understanding and generation. *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, and 1 others. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Dongbo Wang, Chang Liu, Zhixiao Zhao, Si Shen, and 1 others. 2023a. 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, Jiangfeng Liu, and 1 others. 2022. Sikubert and sikuroberta: Construction and application of pre-trained models for the siku quanshu in the field of digital humanities. *Library Tribune*, 42(06):31–43.

Hao Wang, Chi-Liang Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023b. Huatuo: Tuning llama model with chinese medical knowledge. *ArXiv*, abs/2304.06975.

Hongbin Wang, Haibing Wei, Jianyi Guo, and Liang Cheng. 2019. Ancient chinese sentence segmentation based on bidirectional lstm+ crf model. *Journal of advanced computational intelligence and intelligent informatics*, 23(4):719–725.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *ArXiv*, abs/2304.01097.

Rui Yan, Cheng-Te Li, Xiaohua Hu, and Ming Zhang. 2016. Chinese couplet generation with neural network structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2347–2357.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, and 1 others. 2023a. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Baosong Yang, Beichen Zhang, and 1 others. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, and 1 others. 2024b. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19368–19376.

Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023b. Investlm: A large language model for investment using financial domain instruction tuning. *Preprint*, arXiv:2309.13064.

Xiaoyuan Yi, Ruoyu Li, and Maosong Sun. 2017. Generating chinese classical poems with rnn encoder-decoder. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 16th China National Conference, CCL 2017, and 5th International Symposium, NLP-NABD 2017, Nanjing, China, October 13-15, 2017, Proceedings 16*, pages 211–223. Springer.

Y Yuan, D Wang, S Huang, and B Li. 2019. The comparative study of different tagging sets on entity extraction of classical books. *Data Analysis and Knowledge Discovery*, 3(03):57–65.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, and 1 others. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *ArXiv*, abs/2309.11325.

Team Glm Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *ArXiv*, abs/2406.12793.

Jundong Zhang, Songhua Yang, Jiangfeng Liu, and Qi Huang. 2024. Aigc empowering the revitalization of ancient books on traditional chinese medicine:building the huang-di large language model. *Library Tribune*, 44(10):103–112.

Xuanyu Zhang, Qing Yang, and Dongliang Xu. 2023a. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*.

Yiqin Zhang, Sanhong Deng, Qi Zhang, Dongbo Wang, and Hongcun Gong. 2023b. Comparative analysis of language models for linguistic examination of ancient chinese classics: A case study of zuozhuan corpus. In *2023 International Conference on Asian Language Processing (IALP)*, pages 154–161. IEEE.

A 词性标注类别

词性标注包括以下 17 个类别：nr –专有名词（人名），v –动词，n –名词，r –代词，w –标点符号，c –连接词，p –介词，d –副词，t –时间

词，y –语气助词，u –助词，m –数词，a –形容词，f –方位词，ns –专有名词（地名），j –缩写，q –量词。

B 指令构建示例

(1) 初始化指令模板：

instruction = " 从下面的文言文段落中提取实体，并列出人物、地点、时间和官职。以以下格式返回结果：' 人物': [...] ,' 地点': [...] ,' 时间': [...] ,' 官职': [...]."

（2）任务扩展说明：

提示：根据以下示例指令，为命名实体识别 (NER) 任务生成附加指令。这些指令应要求模型从给定的古代汉语段落中提取四类实体，包括人物、地点、时间和官职，并以结构化格式返回结果。生成的指令应多样化、自然且符合汉语语言习惯，并明确指定任务目标和输出格式。示例指令：{ instruction }

（3）逆向推理：

提示：根据给定的输入内容和对应的输出，生成一些清晰而合理的任务指令。指令应有效指导模型从古文中提取四类命名实体，包括人物、地点、时间和官方职位，并以与提供的输出一致的结构化格式返回。请在保持语义准确的同时确保语言多样性。

参考示例：

输入：从以下古代文本中提取角色、地点、时间和官职：UTF8gbsn 侍御史赵承庆见之，叹曰：御史言及此，天下福也。

输出：' 角色': [' UTF8gbsn 赵承庆 '],' 地点': [],' 时间': [],' 官方职位': [' UTF8gbsn 御史 ']"

答案：

请对以下古文进行实体分析，以识别出所有人物的全名、地名、历史时间节点和官职信息，并以 JSON 格式对其进行分类。如果某一类别没有对应的实体，则将列表留空。

2. 您需要从以下古代资料中找出四种类型的信息：人名、具体地点、朝代/年份和官职。请根据示例格式安排结果：' 人物': [],' 地点': [],' 时间': [],' 官方职位': [] "

...

(4) 筛选生成的指令：

接受示例：

输入：对给定的古代汉语文本进行命名实体识别。提取人物、地点、时间和官职，并以如下格式返回：' 人物': [...],' 地点': [...],' 时间': [...],' 官职': [...]。古代汉语文本：UTF8gbsn 天成初，移镇许州，加同平章事。

输出：' 角色': [],' 地点': [' UTF8gbsn 许州 '],' 时间': [' UTF8gbsn 天成初 '],' 官方职位': [' UTF8gbsn 同平章事 ']

拒绝示例：

输入：对给定的古文文本 UTF8gbsn 天成初，移镇许州，加同平章事。进行命名实体识别

输出：UTF8gbsn 天成初：时间，许州：地点，同平章事：官职。

C 词语解释任务案例

Case 2	UTF8gbsnExplain the meaning of " 晕红" in the line " 一片晕红才著雨，几丝柔绿乍和烟" from " 浣溪沙 · 谁道飘零不可怜".
Ground Truth	UTF8gbsn 形容海棠花的色泽。Describes the color of begonia flowers.
WenyanGPT	UTF8gbsn 形容海棠花色泽鲜艳而浓丽。 Describes the bright and rich color of the begonia flower.
Deepseek-V3	UTF8gbsn 指雨后花朵泛起的朦胧红晕。 It refers to the hazy red glow that appears on flowers after the rain.
GPT-4o	UTF8gbsn 指花瓣在雨水滋润下呈现出的一种柔和而略带湿润光泽的红色。Refers to the soft and slightly moist red color that the petals appear to have when nourished by rain.
Qwen2.5-7B-Instruct	UTF8gbsn 指花瓣上晕染开的淡红色，适合颜色柔和、渐变，类似于涂抹或被雨水滋润后散开的红晕。 Refers to the light red color that spreads across the petals. It is suitable for soft and gradual colors, similar to the blush that spreads after being smeared or moistened by rain.
Xunzi-Qwen1.5-7B-Chat	UTF8gbsn 指花瓣上的红色晕染，因为花瓣在雨中被淋湿，所以红色晕染更加明显。Refers to the red smudge on the flower petals. Because the petals got wet in the rain, the red smudge is more obvious.

Table 9: 来自不同大规模语言模型的词语解释任务的响应示例。

在单词解释任务中，WenyanGPT 的回应最接近真实情况。虽然 WenyanGPT 添加了一些修饰词，但它在描述颜色时保持了简单和精炼，这与诗歌的语气相符。Deepseek-V3、GPT-4o、Qwen2.5-7B-Instruct 和 Xunzi-Qwen1.5-7B-Chat 没有将主体指定为 “UTF8gbsn 海棠花 ”(海棠花), 而是过度解释了花瓣的湿润和颜色的扩散过程, 偏离了 “UTF8gbsn 晕红 ”(面色红润）的标准简洁描述。具体来说，Qwen2.5-7B-Instruct 和 Xunzi-Qwen1.5-7B-Chat 的回应扩展了外部环境的影响，这与原文的表达并不完全一致。这种过度扩展表现出这些大型语言模型未能抓住原始描述的简洁性和深度。