基于强化学习的大语言模型古文释义选择研究

徐维潞
南京大学
软件新技术国家重点实验室
weiluxu@smail.nju.edu.cn

黄书剑*
南京大学
软件新技术国家重点实验室
huangsj@nju.edu.cn

摘要

古文释义选择任务对语言模型的语义理解与语境匹配能力提出了较高挑战。本文提出一种基于强化学习的训练框架，通过结果导向的奖励设计，引导大语言模型优化古文释义判断策略。实验表明，相比监督微调（Supervised Fine-tuning, SFT），强化学习方法在准确率指标上表现更优。进一步分析发现，强化学习仅在释义选择任务上的训练不仅提升了模型的古文翻译能力，还在古汉语通用能力评估基准（ACLUE）上展现出更优的跨任务迁移性。相较之下，SFT训练后的模型在翻译与其他古文任务中的表现出现明显下降。本研究为古文处理任务提供了新的训练范式，验证了强化学习在非推理类语言任务中的有效性与泛化潜力。

关键词： 古文翻译；大语言模型；强化学习

A Reinforcement Learning-Based Approach to Ancient Chinese Interpretation Selection with Large Language Models

Weilu Xu
National Key Laboratory for
Novel Software Technology
Nanjing University
weiluxu@smail.nju.edu.cn

Shujian Huang*
National Key Laboratory for
Novel Software Technology
Nanjing University
huangsj@nju.edu.cn

Abstract

The task of selecting appropriate interpretations for ancient Chinese texts presents significant challenges in semantic understanding and contextual alignment for language models. This paper proposes a reinforcement learning-based training framework that guides large language models to optimize their interpretation selection strategies through outcome-driven reward design. Experimental results demonstrate that compared to supervised fine-tuning (SFT), the reinforcement learning approach yields notable improvements in accuracy. Further analysis reveals that reinforcement learning training focused solely on interpretation selection not only enhances the model's ancient Chinese translation capabilities but also shows superior cross-task generalization on the ACLUE benchmark. In contrast, models trained via SFT exhibit a performance decline in translation and other ancient Chinese tasks. This study introduces a novel

*通讯作者
基金项目：国家自然科学基金（No. 62176120, 62376116），中央高校基本科研业务费专项资金（No. 2024300507）

training paradigm for ancient Chinese processing and validates the effectiveness and generalizability of reinforcement learning in non-reasoning language tasks.

Keywords: Ancient Chinese Translation , Large Language Models , Reinforcement Learning

1 引言

古文翻译作为中华文化传播与传承的关键环节，近年来在自然语言处理（NLP）领域受到广泛关注。由于古文语言结构独特、词义多变、文法灵活，并深受历史语境与文化因素影响，古文翻译始终面临着巨大的技术挑战(Zhou, 2023; Wu et al., 2024)。传统的翻译方法难以充分捕捉古文中的语义细节与上下文关系，因而研究者尝试引入词典释义增强技术，以弥补现代模型在语义匹配方面的不足(Li et al., 2023)。通过构建面向古文的释义词典，并将其合理融合到翻译过程中，提升翻译的准确性与可读性。

随着大型语言模型（Large Language Models, LLMs）在机器翻译任务中的快速发展，基于Transformer架构的模型展现出了强大的语言理解与生成能力(Hadi et al., 2023; Zhao et al., 2020)。然而，现有研究发现，LLM在古文翻译任务中的表现仍然不尽如人意，常出现词义偏差、上下文理解不足等问题(Zhu et al., 2024)，这表明其在处理历史语言与语境推理方面仍有待提升。

近年来，强化学习（Reinforcement Learning, RL）被广泛应用于自然语言生成任务，在复杂推理场景中显示出强大的训练效率与性能优势尽管如此，强化学习在非代码或数学推理类任务（如古文翻译中的释义选择）上的应用仍处于探索阶段。如何构建适用于该类任务的强化学习训练流程，并验证其在实际应用中的有效性，仍是当前亟待深入探索的研究方向。

本文尝试探索在古文释义选择任务中结合强化学习与大型语言模型的训练策略。主要贡献包括：

- 本文基于大语言模型构建了一套适用于古文释义选择任务的强化学习训练流程，验证了强化学习在该任务中的有效性。
- 本文系统对比了强化学习与监督微调（Supervised Fine-tuning, SFT）两种训练方式在该任务中的表现差异。
- 本文还分析了不同训练策略对模型在古文翻译与理解等相关任务中的泛化能力影响，并通过系列实验证实强化学习在多任务场景下具备更强的泛化能力与训练稳定性。

2 相关工作

2.1 古文翻译研究现状

古文机器翻译的发展大致可分为基于规则、实例、统计和神经网络四个阶段(李亚超et al., 2018)。早期的规则方法依赖大量人工构建的语言规则，难以适应复杂语境；实例方法通过相似句匹配实现翻译，但覆盖能力有限；统计方法在短语对齐和词汇转换方面取得一定进展，仍受限于语料稀缺；神经机器翻译（NMT）利用编码器-解码器结构提升了语义建模和句法调整的能力，现已成为主流。

词典资源的引入被证明有助于提升模型对古汉语词义的识别与转换能力，如通过双语词典构造伪平行语料或引入外部记忆模块增强翻译模型的知识获取能力。此外，近年来有研究关注古文释义选择任务(Zhao et al., 2020; Feng et al., 2017)，帮助模型在上下文中准确辨析多义词义和语句含义，对提高翻译准确性具有积极意义。本文正是从这一方向出发，着重提升大模型对古文释义的判别能力，并进一步探讨其对翻译质量的具体影响。

2.2 大型语言模型在翻译中的应用

随着大型语言模型（LLM）的快速发展，其在机器翻译中的表现已超越传统统计模型，成为当前主流方法之一。基于Transformer架构的LLM通过堆叠的自注意力机制捕捉文本中的长距离依赖，有效建模句法与语义信息(Hadi et al., 2023; Zhao et al., 2020)。

此外，Prompting、Few-shot Learning 以及QLoRA 等轻量化调优方法的兴起，显著提升了LLM在低资源场景下的适应能力与泛化性能(Zhang et al., 2023)。尽管如此，已有研究发现LLM在古文翻译中仍存在缺陷，如词义模糊、译文不连贯等问题，表明其对历史语言和文化语境的理解能力仍有待提高(Zhu et al., 2024)。

本文不同于传统直接建模翻译的方法，提出将古文释义选择作为独立子任务，引入结构化的输入输出格式，并评估其对古文翻译等任务性能的提升效果。

2.3 强化学习与语言模型的结合

近年来，强化学习（Reinforcement Learning, RL）被广泛应用于自然语言生成任务，尤其是结合人类反馈的强化学习（RLHF）在对话生成中取得了卓越成果(Xu et al., 2025; Dai et al., 2023)。诸如群组相对策略优化（Group Relative Policy Optimization, GRPO）等新方法进一步提升了训练效率与输出质量，在复杂推理任务中表现出色(Guo et al., 2025)。然而，强化学习在非数学或代码推理类任务中的应用仍处于探索阶段，尤其在古文翻译任务中，如何用强化学习训练引导模型进行更优的释义选择，仍是一个亟待研究的问题。

本文将强化学习范式引入释义选择任务，探索在非数学或代码推理类任务中强化学习的可行性。通过与监督微调路径的对比，系统分析两种训练方式在古文理解与泛化能力方面的差异，为古文语言处理模型的优化与训练策略提供实证依据与方法指导。

3 方法

本节介绍本文在古文释义选择任务上的数据构建流程、训练框架与方法设置。具体内容包括：基于古文字典构建的释义选择数据集、两种训练方式（监督微调与强化学习）的实现细节，以及在LLM 基础上构建的任务适配方法。

3.1 数据构建

为支持古文释义选择任务，本文基于互联网上公开的古文字典资源构建了一个字义标注语料库。该数据集的构建过程主要包括以下步骤:

3.1.1 释义和例句抽取

本文抽取词典中汉字对应的所有释义，并收集每条释义所附的多个古文例句，每个例句均标明了该字在该上下文中对应的义项。例如，对于“拔”字，词典中可能列出“拔起”、“选拔”、“突出”等多种释义，每条释义都附带若干例句，如“力拔山兮气盖世”中“拔”表示“拔起；抽出”。

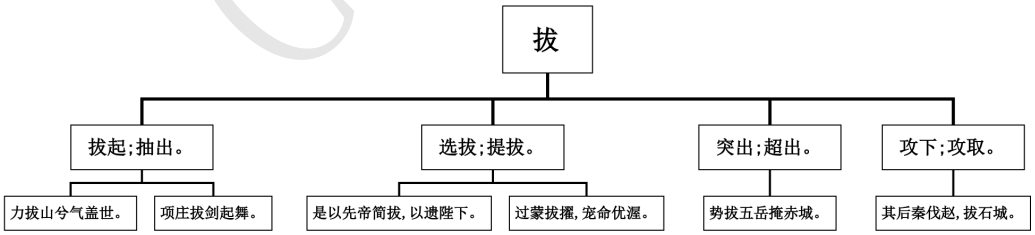

Figure 1: 释义词典结构图

3.1.2 数据格式设计

本文将每一条数据构造成一个释义选择任务：模型输入包括目标汉字、上下文例句以及一组候选释义列表，输出为最符合该语境的释义。例如，模型接收如2所示:

最终构建的数据集包含近万条标注样本，覆盖3000多个常用古文高频字，并覆盖广泛的语境与义项分布。

第二十四届中国计算语言学大会论文集，第640页–第650页，济南，中国，2025年8月11日至14日。

请根据以下释义选择句子中‘拔’的正确含义，注意以下释义中可能没有该字的解释：
句子：力拔山兮气盖世。
拔：1. 拔起；抽出。2. 选拔；提拔。3. 突出；超出。4. 攻下；攻取。
请只回答对应的序号，例如：1, 2, 3 或4，或者“无匹配释义”
目标输出为：1

Figure 2: 古文释义选择示例

3.2 模型训练框架

本文探索强化学习方法在提升大语言模型（LLM）古文释义选择能力方面的可行性与有效性。由于古文释义选择任务不同于典型的逻辑推理任务，其更依赖于上下文理解与语义匹配，因此当前强化学习（RL）方法在此类任务上的效果尚未充分验证。为此，本文构建了一个完整的训练流程，基于强化学习与监督微调两种方式分别对LLM 进行训练与对比，以评估其在古文理解任务中的表现差异。

在模型架构方面，本文采用Qwen2.5-1.5B-Instruct 作为基础语言模型，并分别在LLaMA-Factory 框架(Zheng et al., 2024)与VeRL 框架(Sheng et al., 2024)中开展训练实验。

3.2.1 监督微调(Supervised Fine-tuning)

监督微调部分基于LLaMA Factory 框架完成。LLaMA Factory 是一个易用且高效的大语言模型训练平台，支持主流模型的微调与训练，具备模块化设计、可复现性强、易于本地部署等优势。该平台支持指令微调（SFT）、奖励模型训练、强化学习预训练等多种训练范式，适配多种加速与量化机制，为本研究提供了良好的实验基础。

本研究选取Qwen2.5-1.5B-Instruct 作为基础模型，采用指令监督微调方式（Instruction Supervised Fine-tuning），使用古文释义选择任务构造的数据集进行训练。输入格式采用指令问答形式，具体格式见3.1.2节。

本文分别尝试了以下两种微调策略：

LoRA（Hu et al., 2022)微调：使用参数高效微调策略LoRA（Low-Rank Adaptation），在不改变基础模型参数的前提下，仅引入少量可训练参数以完成微调过程，显著降低显存占用和计算成本。

全参数微调(Full-parameter Fine-tuning)：在全参数微调配置下，本文采用DeepSpeed 提供的分布式训练框架，对Qwen 模型进行完整的梯度更新。尽管该方法在计算资源和时间开销上更为昂贵，但其允许模型在参数层面进行深入调整，有助于充分挖掘模型的潜在能力。

两种方式均使用混合精度（FP16）进行训练，并在训练过程中监控损失函数收敛情况与验证准确率，确保训练过程稳定可靠。

3.2.2 强化学习训练（Reinforcement Learning）

为探讨强化学习在古文释义选择这一非推理类语言任务中的应用效果，本文采用VeRL 框架对大语言模型进行训练。VeRL 是一个面向大语言模型的强化学习训练库，兼容主流训练基础设施，支持包括PPO(Proximal Policy Optimization)(Schulman et al., 2017) 与GRPO (Group Relative Policy Optimization)(Shao et al., 2024)在内的多种强化学习策略，具备良好的扩展性与训练效率。

数据处理与格式转换 训练数据格式与监督微调阶段一致，采用指令式问答格式（如3.1.2所述）。为了提升大规模训练的效率，本研究进一步将数据转为`Parquet` 格式，用于优化I/O 吞吐并减少加载延迟，这对VeRL 的分布式训练过程尤为关键。

算法说明 本研究使用了以下两种强化学习算法：

PPO（Proximal Policy Optimization）(Schulman et al., 2017)：该算法通过限制策略更新幅度以提升训练稳定性。其优化目标如下：

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right],$$

其中$r_t(\theta)$ 为新旧策略的概率比，$\hat{A}_t$ 是优势估计项。

优势函数采用广义优势估计（GAE）：

$$A_t = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}, \quad \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$$

本研究在PPO 框架下使用了三个模块：Actor、Critic 与Reference 模型。其中奖励信号并非由训练得到的奖励模型（Reward Model, RM）提供，而是使用基于规则的评分函数（详见“奖励函数设计”部分）直接赋予离散奖励。PPO 通常对同一批交互数据进行多轮更新，引入剪裁机制防止策略剧烈变动：

$$\text{actor\_loss} = -\min(\text{Adv}_t \cdot r_t, \text{Adv}_t \cdot \text{clip}(r_t, 0.8, 1.2))$$

GRPO（Group Relative Policy Optimization）(Shao et al., 2024)：该算法通过组内奖励机制替代值函数估计，适用于仅对最终输出token 提供奖励的LLM 场景。其训练目标如下：

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \left(\min\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}, \right.\right.\right.$$
$$\left.\left.\left. \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}, 1-\varepsilon, 1+\varepsilon \right) \hat{A}_{i,t} \right) - \beta \text{D}_{\text{KL}}\left[\pi_\theta \| \pi_{\text{ref}}\right] \right) \right]$$

其中，KL 散度项用于约束策略分布相对于参考模型的偏移：

$$\text{D}_{\text{KL}}\left[\pi_\theta \| \pi_{\text{ref}}\right] = \frac{\pi_{\text{ref}}(o_{i,t}|q,o_{i,<t})}{\pi_\theta(o_{i,t}|q,o_{i,<t})} - \log\frac{\pi_{\text{ref}}(o_{i,t}|q,o_{i,<t})}{\pi_\theta(o_{i,t}|q,o_{i,<t})} - 1$$

GRPO 不估算全局基线，通过组内归一化奖励简化优势估计，有效降低了训练成本。

奖励函数设计 本研究未使用训练得到的奖励模型（Reward Model, RM），而是设计了基于规则的离散奖励函数，用于直接赋予每条样本以标注为准的奖励值，简化训练流程并提升稳定性。这种方式避免了额外的RM训练带来的噪声与资源开销，便于专注于强化学习框架本身在古文释义选择任务中的效果验证。

具体地，奖励函数定义如下：

- 从模型输出中提取所选释义编号$a \in \{A, B, C, D\}$；
- 若$a = a^*$，其中a^* 为标准答案，则给予奖励$r = 1$；
- 若提取失败或$a \neq a^*$，则奖励为$r = 0$。

该离散奖励直接用于优势估计$\hat{A}_t$，构成PPO 或GRPO 算法中的训练信号，无需额外回归网络或RM参与。该设计以结果为导向，鼓励模型自主探索并对正确结果提供奖励，为后续复杂奖励机制提供了简洁有效的基线。

4 实验与分析

本实验系统评估不同训练策略（包括监督微调与强化学习）对语言模型在古文释义选择等任务中的性能影响。首先，本研究在构建的释义选择任务上测试各模型对于释义选择的准确率，以衡量其字词层面语义辨析能力。随后，本研究进一步考察各模型在对应释义选择训练集和测试集上的整句翻译表现，从句子生成质量角度进行分析。最后，为全面评估模型的语言理解与知识迁移能力，本研究在古汉语通用能力评估基准集ACLUE (Zhang and Li, 2023)上进行了测试，涵盖词汇、句法、语义、推理与知识五大类别的15 个任务，从多维度反映模型的综合能力与泛化能力。

第二十四届中国计算语言学大会论文集，第640页-第650页，济南，中国，2025年8月11日至14日。

4.1 释义选择任务结果

表 1 显示了在释义选择任务上的表现。基线模型`Qwen2.5-1.5B-Instruct` 的正确率为22.3%，而通过监督微调与强化学习训练后均获得显著提升。其中，PPO 策略训练后的模型达到75.10% 的最高正确率；使用GRPO 的模型紧随其后为73.61%，显示出替代值函数策略在资源效率与性能上的有效性。

模型	正确率
Qwen2.5-1.5B-Instruct	22.30
SFT (LoRA)	39.34
SFT (Full)	70.93
PPO	**75.10**
GRPO	73.61

Table 1: 释义选择任务各模型正确率（%）,基线模型均为Qwen2.5-1.5B-Instruct

本研究进一步从词义分布的角度分析了经PPO 训练后模型的泛化能力。测试集被划分为三类：训练集中未出现该字的释义组合（diff_word_diff_def）、出现该字但未出现该义项（same_word_diff_def）、以及出现相同字义但例句不同的样本（same_word_same_def_diff_ex）。在最具挑战性的diff_word_diff_def 情况下，模型依然取得了82.35% 的准确率；在same_word_diff_def 上准确率为74.03%，虽相对略低，但仍显著优于SFT 基线。这一结果表明，PPO 训练不仅提升了模型对已见样本的拟合能力，更增强了其对字义结构组合的抽象理解与泛化能力。

类别	正确率
diff_word_diff_def	82.35
same_word_diff_def	74.03
same_word_same_def_diff_ex	77.97

Table 2: 不同释义分布条件下PPO 模型表现

为进一步理解见过相同字不同释义正确率偏低的原因，对相关错误样本进行了分析。观察发现，当句中语义关系较为隐晦、存在如词类活用等典型古文现象时，模型倾向于选择训练中更常见的释义项。例如（见3）在句子“非口不能味也”中，“味”作动词使用，意为“辨别味道”；而模型却选择了更常见的名词义项“滋味、味道”，与其在训练集中见过的句子“口能辩味”中释义一致。这表明，在语境不甚明确或语法结构复杂的情况下，模型可能优先依赖其在训练中形成的释义记忆。尽管这种策略在一定程度上体现了模型对词项频率的敏感性，但也提示其仍需提升对复杂语境中意义细微差别的感知与判断能力。该趋势提示，模型性能的提升可能依赖于减少对训练记忆的错误重用。一种可能的改进方向是，在训练过程中引入机制，对过度依赖高频释义的选择行为施加一定的惩罚，以促使模型更加重视语境信息。

测试集相同字不同释义错误样例
“sentence”：“非口不能味也。”
“dictionary”：“味：1. 滋味；味道。2. 辨别味道。”
“answer”：“2”
“model_answer”：“1”

见过的训练样本
“sentence”：“耳能辩声，口能辩味。”
“dictionary”：“味：1. 滋味；味道。2. 辨别味道。”
“answer”：“1”

Figure 3: same_word_diff_def错误样例

4.2 翻译任务实验结果

在现代文翻译任务中，本文比较了不同策略训练下模型在seen（训练集中出现过相关释义选择）与unseen（未见过的释义选择）样本上的BLEU 分数，如表3 所示。任务使用的翻译数据来自互联网上公开的古汉语-现代汉语平行句对语料，主要来源于《二十四史》等古籍。结果表明，PPO 训练不仅在释义选择表现优异，在翻译任务上也取得了提升。而监督微调对模型的翻译能力产生了负面影响，特别是SFT-full 微调虽然在释义选择任务中表现良好，但其对翻译能力产生明显负面影响，尤其在seen 数据中，BLEU 降至9.27。

模型	unseen	seen
Qwen2.5-1.5B-Instruct	16.07	16.12
SFT (LoRA)	16.53	15.76
SFT (Full)	15.71	**9.27**
PPO	**17.88**	16.90

Table 3: 翻译任务BLEU 分数比较

Case Study 本例中，原文中的关键词“但”意为“只；仅；唯独”，而“匿”意为“隐藏；躲藏”，均在释义词典中有清晰定义。在多个模型输出中，**仅PPO 模型准确翻译出“只”与“躲藏”两个关键语义**，体现了对原文细粒度词义的把握能力。相比之下，基础模型和SFT 模型虽句式通顺，但在细节上存在语义遗漏或偏差。这表明PPO 强化学习不仅提升了释义选择准确性，也帮助模型更好地将词义融入翻译任务中，实现结构性泛化。

古文句：何但远走，亡匿于幕北寒苦无水草之地为?
参考译文：何必只是一味地向北逃跑，躲藏在大漠北边寒冷凄苦、没有水草的地方呢?
Qwen1.5B-Instruct：难道要远远离去，躲藏在北边的寒冷干旱之地吗?
SFT (LoRA)：难道会逃到北方的寒冷艰苦而没有水草的地方隐居吗?
SFT (Full)：何至于远走而逃入到幕北的寒冷荒凉之地，那里没有水草，以致无法生存呢?
PPO：你只要远远地逃跑，躲藏在北方苦寒无水草的地方吧?

Figure 4: 翻译任务Case Study

本研究进一步统计了不同训练策略下模型在翻译任务中对已见释义信息的利用情况。验证流程如图 5所示。首先，在参考译文中检索是否包含目标释义项，若包含，则认为该释义在当前语境中对翻译具有贡献，并进入下一步分析。接着，依次检查基线模型（Qwen2.5-1.5B-Instruct）的翻译结果是否包含该释义，以及不同训练策略下模型的翻译结果是否包含。最终统计在基线模型未包含而训练后模型包含目标释义的情况，以评估训练对释义信息引入的效果。

结果显示，在参考译文中包含目标释义项（即该释义在当前语境下对翻译具有实际贡献）的样本中，基线模型仅有16.9%能在翻译中体现相应释义。而经过PPO训练的模型在此类样本中有约**41.5%** 成功地在翻译中使用了目标释义，相较之下，经过全参数微调(SFT Full)的模型仅为23.5%。这一结果表明，经过PPO训练后的模型在释义选择能力上具备更强的泛化性，能够更有效地将该能力迁移至翻译任务中。

4.3 释义选择和翻译任务联合测试

为了进一步评估模型在复杂任务组合下的能力，本文设计了包含“释义选择+整句翻译”两项子任务的联合任务。本研究将测试集的例句在上述平行句对语料库中匹配出标准释义。

具体的Prompt 格式如6所示:

在双任务综合测试中，所有模型在释义选择任务上的准确率均较单任务时有所下降，可能由于输入指令变得更复杂，增加了模型解析任务的难度。然而，PPO 模型在该情境下的准确率仍保持在**68.06%**，下降幅度最小，显示出其在处理多目标任务时更强的稳健性和泛化能力。相比之下，LoRA 微调模型略有下降（**44.44%**），而全参数微调模型表现显著退化至**17.36%**。

在古文翻译任务中，PPO 模型同样展现出最优性能，BLEU 分数达到**20.33**，优于基础模型（**19.40**）与LoRA 微调模型（**19.65**）。而全参数微调模型的BLEU 分数仅为**1.03**，表明其

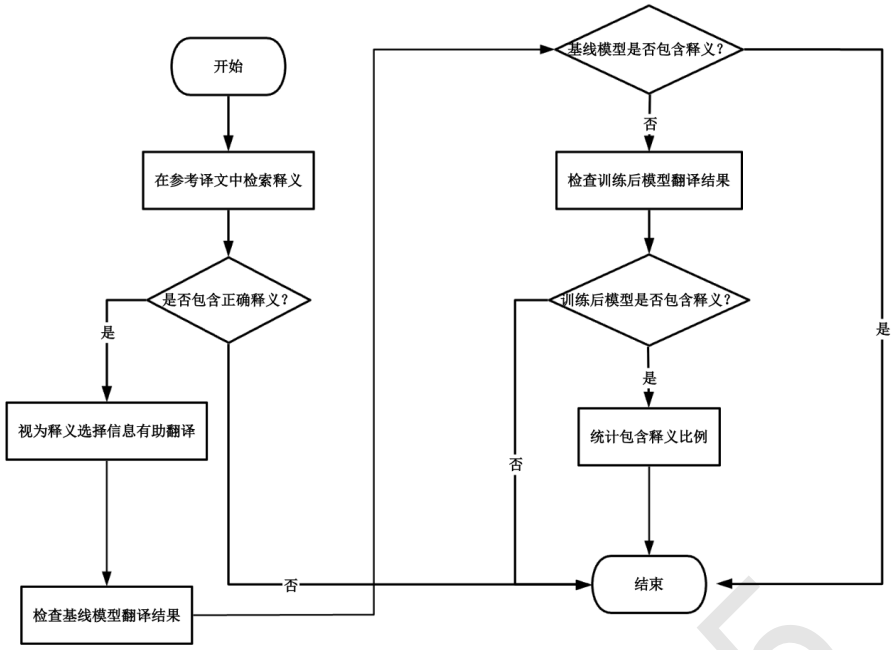

Figure 5: 分析释义选择信息对翻译影响的实验流程

```
请完成以下两个任务，缺一不可：
任务一：选择句子中“{{word}}”的正确含义（如无合适选项，请写“无匹配释义”）。
任务二：将整句翻译为现代文，语言准确通顺，不能遗漏！
请参考以下示范：
—
input:
句子：何但远走，亡匿于幕北寒苦无水草之地为?
释义：但：1. 只；仅；唯独。2. 只管；尽管。3. 不过；只是。4. 徒然；白白地。5. 只要。

output:
释义选择：4
翻译：何必白白地向北逃跑，躲藏在大漠北边寒冷凄苦、没有水草的地方呢?

input:
句子：烈士暮年，壮心不已。
释义：暮：1. 傍晚；日落时分。2. 迟；晚。

output:
释义选择：2
翻译：烈士虽到暮年，壮志雄心不已。

input:
句子：{sentence}
释义：{dictionary}

output:
释义选择：
翻译：
```

Figure 6: 联合测试提示词

第二十四届中国计算语言学大会论文集，第640页-第650页，济南，中国，2025年8月11日至14日。

Model	释义选择准确度(%)	BLEU 分数
Qwen2.5-1.5B-Instruct	43.06	19.40
PPO	**68.06**	**20.33**
SFT (LoRA)	44.44	19.65
SFT (Full)	17.36	1.03

Table 4: 联合测试结果

在面对复合任务指令时的泛化能力不足，可能因过拟合单一任务而导致复合任务中性能崩溃。

进一步观察发现，双任务综合测试中，第一个任务（释义选择）给第二个任务（翻译）带来了上下文(inconext)的帮助信息，除了全参数微调模型外，其余模型的翻译得分均高于第4.2节中单独翻译任务的测试结果。说明模型能够有效利用前序任务提供的上下文信息，从而提升翻译质量。而全参数微调模型翻译性能的下降，可能与其指令跟随能力减弱以及在第一阶段错误选择释义有关。

综合来看，PPO 强化学习不仅提升了释义选择准确率，也在多任务场景中保持了更好的稳定性与迁移能力，体现出相较于传统监督微调(SFT) 更优的任务稳定性。

4.4 通用能力测试：ACLUE评估

为了全面评估模型在古文理解上的泛化能力，本文采用**ACLUE** (Zhang and Li, 2023)测试集作为基准。ACLUE 是一个面向古代汉语的大型语言模型评估基准，包含15 个多项选择题任务，涵盖古代汉语中的词汇、句法、语义、推理、知识等多个层面，是目前最系统的古汉语理解评估集合之一。

模型	总得分
Qwen2.5-1.5B-Instruct	44.40
SFT (LoRA)	43.45
SFT (Full)	45.33
PPO	**49.07**

Table 5: ACLUE Benchmark 各模型总得分

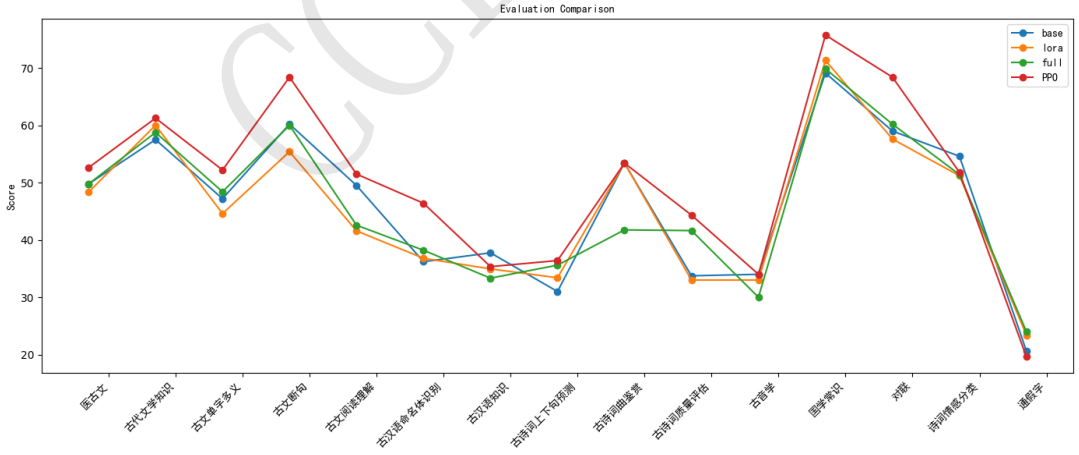

Figure 7: 在ACLUE上各任务的表现

本研究评估了基础模型（Qwen-1.5B-Instruct）、监督微调模型（包括LoRA 和全参数微调），以及经PPO 强化学习训练后的模型在ACLUE 基准测试集上的表现，结果如表 5 所示。PPO 模型在整体得分上达到**49.07**，优于基础模型（44.40）和两种SFT 微调模型（LoRA: 43.45，Full: 45.33）。进一步分析表明，PPO 模型在“释义选择”相关任务（如单字多义识别与

命名实体识别）上表现提升显著，而在其他类型任务中的性能变化较小，整体表现稳定。相比之下，SFT 微调虽在部分知识类任务中带来增益，但在古音学、诗词鉴赏等任务上存在较为明显的性能下降，反映出其对非目标任务可能产生一定的干扰效应。

PPO 强化学习中引入的KL 散度约束机制，有助于保持模型在训练后与原始模型之间的相似性，从而有效控制对无关任务的影响。此外，“释义选择”任务的训练本身增强了模型对文义细节的理解能力，对古文类任务也产生了一定的正向迁移效应。

综上所述，PPO 强化学习不仅在特定任务上实现了性能提升，同时在古文理解等通用能力上展现出更强的稳健性与迁移能力。

5 结论

本文系统探索了强化学习在古文释义选择任务中的应用价值及其对古文翻译与理解的泛化影响。通过构建基于PPO 的训练流程，并系统比较其与监督微调在释义选择、翻译任务及ACLUE 基准评测中的表现，实验结果表明，PPO 强化学习策略在多个层面均展现出显著优势：不仅在释义选择任务中取得最高准确率，具备更强的语义辨析与泛化能力，也能将该能力有效迁移至整句翻译任务中，实现更高质量的语言生成。在联合任务中，PPO 模型展现出更好的任务稳定性与指令理解能力，能够充分利用前置任务提供的上下文信息；在ACLUE 的综合评测中亦取得最高得分，表现出对多样化古文任务的稳健适应能力。相比之下，SFT（尤其是全参数微调）虽然在部分任务中有所提升，但存在过拟合和泛化能力不足的问题。综上所述，本研究验证了基于PPO 的强化学习策略在提升大语言模型古文处理能力方面的有效性，不仅增强了模型的任务特定能力，也提升了其对复杂语境和多义词判断的整体理解与迁移能力。未来的工作将进一步探索更细致的奖励设计与多任务融合机制，以推动古文语言模型在理解上的持续进步。

参考文献

Chengbin Zhou. 2023. 基于深度学习的古文翻译方法研究. Doctoral dissertation, 中北大学. DOI: 10.27470/d.cnki.ghbgc.2023.000078.

Mengcheng Wu, Litao Lin, Die Hu, et al. 2024. 我国古代典籍时代特征视角下的机器翻译研究. 图书馆论坛, 44(10):93–102.

Jiahuan Li, Ruochun Wu, Weilu Xu, Shujian Huang, et al. 2023. 词典释义增强的古文机器翻译研究. In 全国机器翻译大会*(CCMT)*.

Muhammad U. Hadi, Rameez Qureshi, Arsalan Shah, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*, 3.

Hang Zhao, Jitendra Jain, and Vladlen Koltun. 2020. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10076–10085.

Xiaoyu Zhang, Narges Rajabi, Kevin Duh, et al. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481.

Danhao Zhu, Zhixiao Zhao, Yiping Zhang, et al. 2024. 面向古文自然语言处理生成任务的大语言模型评测研究. 信息资源管理学报, 14(05):45–58. DOI: 10.13365/j.jirm.2024.05.045.

Feng Xu, Qiang Hao, Ziyan Zong, et al. 2025. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. *arXiv preprint arXiv:2501.09686*.

Jinchao Dai, Xiaosen Pan, Ruize Sun, et al. 2023. Safe RLHF: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Dongxu Guo, Duyu Yang, Haoyu Zhang, et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

第二十四届中国计算语言学大会论文集，第640页–第650页，济南，中国，2025年8月11日至14日。

李亚超, 熊德意, 张民. 2018. 神经机器翻译综述. 计算机学报, 41(12): 2734–2755.

Yang Zhao, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2020. Knowledge graphs enhanced neural machine translation. In *International Joint Conference on Artificial Intelligence*.

Yang Feng, Shiyue Zhang, Andi Zhang, Dong Wang, and Andrew Abel. 2017. Memory-augmented neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1390–1399, Copenhagen, Denmark. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. *LlamaFactory: Unified efficient fine-tuning of 100+ language models*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics. arXiv:2403.13372.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. *HybridFlow: A flexible and efficient RLHF framework. arXiv preprint arXiv:2409.19256*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Yixuan Zhang and Haonan Li. 2023. Can large language model comprehend Ancient Chinese? A preliminary test on ACLUE. In *Proceedings of the Ancient Language Processing Workshop*, pages 80–87, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.