

武汉大学

本科毕业论文（设计）

基于大模型的跨模态古诗创作

姓 名 : 张志东

学 号 : 2021302111480

专 业 : 软件工程

学 院 : 计算机学院

指 导 教 师 : 朱卫平 副教授

二〇二五年四月

## 原创性声明

本人郑重声明：所呈交的论文（设计），是本人在指导教师的指导下，严格按照学校和学院有关规定完成的。除文中已经标明引用的内容外，本论文（设计）不包含任何其他个人或集体已发表及撰写的研究成果。对本论文（设计）做出贡献的个人和集体，均已在文中以明确方式标明。本人承诺在论文（设计）工作过程中没有伪造数据等行为。若在本论文（设计）中有侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

作者签名：张志东 指导教师签名：李平  
日期：2025年5月14日

## 版权使用授权书

本人完全了解武汉大学有权保留并向有关部门或机构送交本论文（设计）的复印件和电子版，允许本论文（设计）被查阅和借阅。本人授权武汉大学将本论文的全部或部分内容编入有关数据进行检索和传播，可以采用影印、缩印或扫描等复制手段保存和汇编本论文（设计）。

作者签名：张志东 指导教师签名：李平  
日期：2025年5月14日

## 摘 要

中国古代诗歌作为中华传统文化的重要载体，拥有高度凝练的语言风格、丰富多变的意象系统以及严格的格律结构。伴随着自然语言处理技术的发展，古诗生成作为生成式人工智能的研究热点，逐渐引起广泛关注。如何让机器模型深入理解并创作出具有文化内涵和艺术价值的古诗，继而为中华传统文化的再创造赋能，是一个富有挑战的有趣问题。

然而，现有方法普遍存在三方面不足：一是仅支持图像或文本的单一模态输入，难以捕捉用户完整的创作意图；二是缺乏对生成过程与结果的可解释机制，对古诗的评价也缺乏可解释性；三是生成的古诗停留在形式语义，无法触及典故、意象等更深层的文化内涵，缺乏文化深度与艺术性。

为此，本文旨在利用大模型的强大语义理解与生成能力，开发一个支持文图双模态的古诗生成系统，在生成时提供用户友好的解释性内容，并支持对古诗的分析与多轮优化。该系统采用 DeepSeek-VL2 模型提取图像中的场景、意象与情感要素，生成具有文化联想价值的描述文本。之后结合用户输入的文本提示，参考 CRISPE 框架设计提示词，调用 DeepSeek-R1 生成高质量的格律古诗，并提供赏析、注释和白话文翻译以提高结果可解释性。为量化评估古诗质量，本文设计了一个包含格律规范、意象意境、主题思想、语言锤炼和创新性五个维度的评分体系，结合 Few-shot 提示框架引导大模型进行客观评分与改进建议生成，同时借助 ROUGE 等自动度量方法作为辅助参考。最后，系统支持多轮优化，在结合改进建议的同时不偏离用户的原始输入，结合评分体系进行针对性质量提升，实现“生成—分析—反馈—优化”闭环。

实验表明，基于强化学习训练的 DeepSeek-R1 相比于 ERNIE-4.0 等大语言模型，在典故意象的理解与运用能力上更具优势，能够给出充分合理的典故注释。而基于中文语境训练的 DeepSeek-VL2 相比于英文模型 CLIP 与 Minigpt4，更能捕捉图像中的文化符号和情感。通过对质量分层的古诗作品集的评分测试，证明系统的评分功能具有良好的区分度与可信度，并进一步检验了自动度量方法的有效性。此外，本文还尝试对比 DeepSeek-R1 与 ERNIE-4.0 生成古诗的能力，但由于测试数据《唐诗三百首》与大模型训练语料重叠而失效。通过消融实验对文图模态进行测试，并未发现其对古诗生成质量的显著作用，但图像模态的加入能够帮助用户更好地表达创作意图，提升系统的可用性与交互体验。

本文通过设计与实现一个基于大模型的跨模态古诗生成系统，结合文图两张模态来强化生成古诗与用户需求的契合度，让模型输出用户友好的解释性文本，提高了系统输出结果的可解释性；并设计一套完备有效的古诗评分体系，探索利用大模型进行古诗评分的可行性，并辅以自动度量方法进行验证。本文还测试探索了不同大模型在图像描述和古诗生成上的表现，尤其是对诗意图象典故的把握，并设计实验检验系统有效性，批判地分析了文图模态对古诗生成的影响。本文的研究工作为古诗生成系统的设计与实现提供了新的思路与方法，推动了大模型在中华传统文化传承与再创造中的应用。

**关键词：**古诗生成；大模型；跨模态；可解释性

## ABSTRACT

Ancient Chinese poetry, as a vital carrier of traditional Chinese culture, is characterized by its highly condensed language, rich and varied imagery systems, and strict metrical structures. With the advancement of natural language processing technologies, poetry generation has emerged as a key research area in generative artificial intelligence, attracting increasing attention. Enabling machine models to deeply understand and compose culturally rich and artistically valuable ancient poems —thereby empowering the reinterpretation and recreation of traditional Chinese culture —poses a challenging yet fascinating problem.

However, existing methods generally suffer from three major limitations: (1) reliance on single-modal inputs (either text or images), making it difficult to fully capture user intent; (2) lack of explainability in both the generation process and evaluation, hindering interpretability; and (3) generated poems often remain superficial in form and semantics, failing to incorporate deeper cultural elements such as allusions and imagery, thus lacking cultural depth and artistic merit.

To address these issues, this study leverages the powerful semantic understanding and generation capabilities of large models (LMs) to develop a multimodal poetry generation system that supports both text and image inputs. The system provides user-friendly explanatory content during generation and enables poem analysis and iterative refinement. Specifically, the DeepSeek-VL2 model extracts scenes, imagery, and emotional elements from images to generate culturally evocative textual descriptions. These descriptions, combined with user-provided text prompts, are then processed using the CRISPE framework to guide DeepSeek-R1 in generating high-quality metrically structured poems. To enhance interpretability, the system provides annotations, appreciations, and vernacular translations.

For quantitative evaluation, a five-dimensional scoring system is designed, assessing metrics such as metrical correctness, imagery and artistic conception, thematic coherence, linguistic refinement, and creativity. A few-shot prompting framework guides LLMs in generating objective scores and improvement suggestions, supplemented by automated metrics like ROUGE for validation. Additionally, the

system supports iterative optimization, refining poems based on feedback while preserving user intent, forming a closed-loop workflow of “generation – analysis – feedback – optimization”.

Experiments demonstrate that DeepSeek-R1, trained via reinforcement learning, outperforms models like ERNIE-4.0 in understanding and applying allusions and imagery, providing more reasonable annotations. Meanwhile, DeepSeek-VL2, optimized for Chinese cultural contexts, surpasses English-based models like CLIP and MiniGPT-4 in recognizing culturally significant symbols and emotions in images. Evaluation on a stratified poetry dataset confirms the scoring system’s discriminative power and reliability, while validating the effectiveness of automated metrics.

Ablation studies on multimodal inputs reveal that while images do not significantly improve generation quality, they enhance user intent expression and system usability. Additionally, comparative tests between DeepSeek-R1 and ERNIE-4.0 were inconclusive due to overlap between the test dataset (*Three Hundred Tang Poems*) and the models’ training corpora.

This study contributes a novel multimodal poetry generation system that strengthens alignment between generated poems and user intent while improving interpretability through explanatory outputs. It also proposes a comprehensive scoring framework, exploring the feasibility of LLM-based evaluation supplemented by automated metrics. Furthermore, it critically examines the role of multimodal inputs in poetry generation and evaluates different LLMs’ capabilities in capturing poetic aesthetics and cultural depth. The findings provide new insights into AI-driven poetry generation, advancing the application of large models in the preservation and creative reinterpretation of traditional Chinese culture.

**Keywords:** Poetry Generation; Large Models; Cross-modal; Interpretability

# 目 录

|                          |     |
|--------------------------|-----|
| <b>摘要</b> .....          | I   |
| <b>ABSTRACT</b> .....    | III |
| <b>1 绪论</b> .....        | 1   |
| 1.1 研究背景和意义 .....        | 1   |
| 1.2 研究现状 .....           | 3   |
| 1.2.1 中文古诗生成 .....       | 3   |
| 1.2.2 古诗质量评价 .....       | 4   |
| 1.2.3 大模型技术 .....        | 5   |
| 1.2.4 提示工程 .....         | 6   |
| 1.3 研究思路和主要贡献 .....      | 6   |
| 1.4 论文组织结构 .....         | 8   |
| <b>2 现有技术介绍</b> .....    | 9   |
| 2.1 古诗生成任务概述 .....       | 9   |
| 2.2 古诗质量评估 .....         | 10  |
| 2.2.1 BLEU .....         | 10  |
| 2.2.2 ROUGE .....        | 11  |
| 2.2.3 Distinct .....     | 11  |
| 2.2.4 Similarity .....   | 11  |
| 2.2.5 人工评估 .....         | 12  |
| 2.3 DeepSeek 大模型 .....   | 12  |
| 2.3.1 DeepSeek-VL2 ..... | 12  |
| 2.3.2 DeepSeek-R1 .....  | 13  |
| 2.4 本章小结 .....           | 14  |
| <b>3 古诗生成与优化</b> .....   | 16  |
| 3.1 系统架构概述 .....         | 16  |

|                        |           |
|------------------------|-----------|
| 3.2 图像分析 .....         | 17        |
| 3.3 古诗生成 .....         | 20        |
| 3.4 古诗评价 .....         | 24        |
| 3.4.1 系统评分 .....       | 24        |
| 3.4.2 自动度量方法 .....     | 33        |
| 3.5 古诗优化 .....         | 35        |
| 3.6 本章小结 .....         | 39        |
| <b>4 实验及结果分析 .....</b> | <b>41</b> |
| 4.1 基于白话文的古诗生成实验 ..... | 41        |
| 4.2 评分功能实验 .....       | 42        |
| 4.3 文图结合的古诗生成实验 .....  | 47        |
| 4.4 本章小结 .....         | 49        |
| <b>5 结语 .....</b>      | <b>50</b> |
| <b>参考文献 .....</b>      | <b>53</b> |
| <b>致谢 .....</b>        | <b>59</b> |
| <b>附录 A 成果 .....</b>   | <b>60</b> |

# 1 绪论

本章主要介绍研究背景与意义、中文古诗生成相关的研究现状，阐述本论文的研究思路与主要贡献，并介绍论文的组织架构。

## 1.1 研究背景和意义

文本生成任务（Text Generation）是自然语言处理（Natural Language Processing, NLP）的一个重要研究方向，需要在给定输入或上下文的条件下，输出符合要求的文本，涵盖机器翻译、文本摘要、对话生成、作品创作等多个应用方向。而在文本生成领域中，中国古代诗歌的生成更是一个困难的任务。

古诗是中华传统文化的瑰宝，作为最早形成的中国古代文学作品体裁之一，其措辞简洁、内涵丰富且韵律整齐。中国古诗的典故意象运用极具文化特色、用词凝练优雅，这些独特的艺术特色都为古诗的机器创作带来巨大的挑战。如何让机器模型深入理解并创作出具有文化内涵和艺术价值的古诗，继而为中华传统文化的再创造赋能，是一个富有挑战的有趣问题。

早期的古诗生成主要依赖于其他子领域的研究思路。例如，基于规则的方法利用现有古诗作为模板，根据既定规则替换字词来生成新的古诗，这样生成的古诗虽在形式上合格，但表达力欠佳<sup>[1]</sup>；基于摘要的方法将其看作是一个摘要生成任务，只是输入是作者的表达意图，且需要考虑中文古诗独特的韵律形式约束<sup>[2]</sup>；基于统计机器翻译的方法则将其看作一个机器翻译任务，将古诗的上下句子分别看作是翻译的源语言和目标语言，利用统计机器翻译（Statistical Machine Translation, SMT）的方法来生成古诗<sup>[3]</sup>。

随着深度学习技术的发展，近年的中文古诗生成大多将古诗生成视作是“从序列到序列”的预测任务（Sequence-to-Sequence），并由此出发训练循环神经网络（Recurrent Neuro Networks, RNNs），如编码器-解码器（Encoder-Decoder）模型<sup>[4]</sup>，并以此为基础设置额外机制来增强语义表现<sup>[5]</sup> 或韵律格式约束<sup>[6, 7]</sup>。然而，这些方法对古诗内涵的掌握往往停留在上下文语义或是韵律对仗，无法进一步触达诸如典故、意象和全诗连贯性等更复杂的方面。所幸，大模型（Large Model, LM）展现出强大的语义理解与创作能力，而在中文领域也出现了诸如ERNIE<sup>[8]</sup>、DeepSeek-R1<sup>[9]</sup>这样的大语言模型（Large Language Model, LLM）和

DeepSeek-VL<sup>[10]</sup> 等跨模态大模型，为这一领域注入全新的活力<sup>[11]</sup>。

除了文本输入外，中国古诗往往蕴含着丰富的场景意象，其对应的视觉信息难以通过用户输入来精确描述。目前也有方案直接使用图像作为输入，如在循环神经网络外增加卷积神经网络（Convolutional Neuro Networks, CNNs）以处理图像信息，捕捉图像关键主体并把握整体氛围，最终生成古诗<sup>[12, 13]</sup>。但相应地，这些方案放弃了文本输入能够具体描述要求的优势，输入图像所含信息的繁杂也导致生成古诗主题、内涵乃至风格的波动。现有的方案大都局限于或文本或图像的单一模态输入，要么局限于上下文语义或韵律的形式规律而无法触达更高的艺术层次，要么受制于图像信息的多变而无法实现更精细的输出控制。这种单一模态的输入限制了用户描述需求的可能，也使生成的古诗缺乏层次，这暗示着文图跨模态输入的研究方向，也是本选题希望探讨的内容（如图 1.1）。

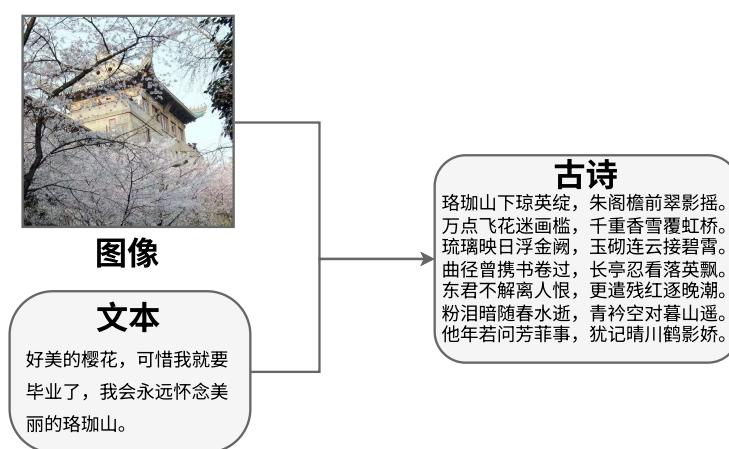


图 1.1 文图跨模态的古诗生成

另一个重要的问题是生成结果的“可解释性”（Interpretability），其指系统以人类可理解的术语解释或呈现模型行为的能力<sup>[14]</sup>。深度学习技术在带来更高表现的同时，也因“黑盒”的特性使得可解释性问题愈发突出。在古诗生成任务中，可解释性可体现为三个方面：

- (1) 过程可解释性，即系统是如何从输入的文本和图像中提取信息，并进一步并生成古诗的，其中包含哪些中间步骤；
- (2) 结果可解释性，即系统为何生成这样的古诗结果，其中遵守怎样的韵

律形式、又运用了哪些典故意象；

- (3) 反馈可解释性，即系统如何分析生成古诗的质量，如何让用户理解古诗“好坏在哪里”。

现有的古诗生成系统大多缺乏对生成过程的可解释性，用户难以理解系统是如何从输入信息中提取出关键信息并生成古诗的。而在结果可解释性方面，现有的系统也往往只提供了生成古诗的文本，而没有进一步解释其韵律、意象等方面内涵。在反馈可解释性方面，现有的系统也缺乏对古诗质量评价维度的详细说明，用户需要具备较高的文学素养以理解和甄别对古诗质量的评价。因此，提升古诗生成系统的可解释性，将有助于增强用户对系统的信任感和使用体验，使得用户能够更好地理解和利用系统输出的古诗和相应的质量判断。

本文旨在探讨文图跨模态的中文古诗生成，开发了一个基于大模型的古诗创作系统。在给定用户两种模态输入的条件下，其能够通过图像的分析描述来充分提取图像信息，结合用户输入的文本信息，生成符合古诗韵律和意象的古诗。此外，系统还将提供对古诗的分析文本、量化评分以及改进意见，涵盖韵律对仗、典故意象、主题思想、语言用词等多个赏析方面，并支持多轮迭代优化。

## 1.2 研究现状

下面从与本选题相关的研究现状出发，介绍中文古诗生成、古诗质量评价、大模型技术和提示工程等四个方面的发展状况，为本文工作做铺垫。

### 1.2.1 中文古诗生成

近年来，中文古诗生成领域引起了广泛的研究兴趣。2016年，为解决古诗生成过程中主题漂移的问题，有研究使用修改的注意力结构的编码器-解码器模型，但限制关键词的数量和顺序，降低了系统的灵活性<sup>[15]</sup>。这一问题在2018年通过记忆网络（增加记忆组块的RNNs）基于图像生成古诗解决<sup>[13]</sup>。2020年，有研究将注意力机制引入了Seq2Seq模型，实现了基于关键字的自定义古诗生成<sup>[16]</sup>。

古诗的生成可能会出现多方面的质量波动，包括主题、语言风格、字数格式、韵律对仗等等，需要在研究中纳入考量。2020年，有研究设计了一个基于Transformer的自回归模型，改进注意力机制并进一步收紧了包括中文古诗、歌

词、英文十四行诗等特殊文体生成的格式要求<sup>[6]</sup>。2021 年，有研究尝试从图像中提取物体关键词、情感和风格，以生成古诗<sup>[17]</sup>。还有研究将数十万首古诗按照风格、情感、格式与主要关键词分类，并利用掩蔽自注意力机制来建立标签到诗句的关联，以此来生成情感与风格均可控的古诗<sup>[18]</sup>。到 2022 年，有研究将 GAN 中的判别器与生成器结构加入到 CVAE 中，实现对风格和情感的控制<sup>[19]</sup>。2023 年，有研究构建了一个古诗图像数据集，精确标注了其中的诗歌元素，并利用 GRU 网络来增强生成古诗的上下文关联<sup>[20]</sup>。2024 年，扩散模型首次被使用来生成古诗，以实现对语义与韵律的同时控制<sup>[7]</sup>。

后来逐渐出现使用大语言模型生成古诗的方案。2024 年，有研究通过强化学习算法 PPO 对 GLM 模型进行训练，提升其在古诗生成方面的表现<sup>[21]</sup>。“CharPoet”这一研究则修正了 LLM 以 token 计数会导致输出格式错误的问题，改以字符计数，达到了极高的格式精度<sup>[11]</sup>。还有研究提出了一种图片输入的三段式绝句生成方法，将短语特征纳入考量，并构建了一个图像主题数据集。其将输入的图片映射到一个主题词，再随机选择与该主题词相关的短语，再通过一正一反两种方向的 LLM 来依次生成古诗的首行、标题和其他主体内容<sup>[22]</sup>。

值得注意的是，鲜有研究关注文图跨模态的古诗生成。据调研，目前只有一项研究同时包含文本和图像两种模态的输入，其通过 Clarifai 来将图像映射到两个具体的主题词，根据这两个主题词在已有短语库中进行检索拓展，再进一步将得到的短语通过一个自注意网络来生成古诗<sup>[23]</sup>。这一系统允许用户来限制古诗的诗句前缀词，因此可广义地认为实现了图文的跨模态输入。

### 1.2.2 古诗质量评价

如何评价生成古诗的质量是一个难题，考虑到古诗体裁的艺术性，其质量评价往往依赖于人工评估。除此之外，也可使用以往文本生成的自动度量方法，如源自机器翻译领域基于  $n$ -gram 的翻译文本评估方法 BLEU<sup>[24]</sup> 和 ROUGE<sup>[25]</sup> ( $n$ -gram 表示文本中连续的  $n$  个词或字符，如“我爱你”中基于字符的 2-gram 集合为[“我爱”、“爱你”])。其中，BLEU 指标计算生成文本中有多少  $n$ -gram 出现在参考文本中，即使用精确率（Precision）来评估生成文本有多接近参考文本；相反，ROUGE 指标计算参考文本中有多少  $n$ -gram 能够被生成文本包含，即使用召回率（Recall）来评估生成文本能否完整地覆盖参考文本。可见，BLEU 和 ROUGE

指标均依赖于与高质量参考文本的对比。此外，Distinct<sup>[26]</sup> 基于  $n$ -gram 的多样性来评估生成古诗的多样性，即诗句里有多少不同的  $n$ -gram。而为了评估上下文语义的一致性，Similarity<sup>[27]</sup> 使用词向量来计算句子之间的相似度。这些方法都能脱离参考文本独立地评价文本质量。

除了自动度量方法外，也有一些研究通过深度学习模型实现了对古诗的质量优化。2020 年，有研究提出一个质量感知的掩蔽语言模型，实现一个可迭代的古诗优化框架，可用于判断古诗是否需要优化，并在润色时定位不恰当部分<sup>[28]</sup>。2023 年，有研究提出一种人机协作的古诗创作系统，能够在不同的约束条件下对古诗进行润色<sup>[29]</sup>。2024 年，又有研究提出了一个可迭代的古诗优化框架，基于 BiLSTM 和 CRF 构建用于检测低质量用词的检测网络、基于 BERT 模型构建用于修正用词的校正网络<sup>[30]</sup>。这些研究均实现了“评价 + 优化”的流程功能，且效果良好。美中不足的是未提供对“评价”本身的解释，即“被选中的词为何是低质量的，而优化又依据着什么”的问题。

### 1.2.3 大模型技术

近年来，大模型技术取得了显著的进步，其中以大语言模型最为典型，例如 OpenAI 开发的生成预训练转换器（GPT）系列模型、百度开发的知识整合增强表示（ERNIE）系列模型<sup>[8]</sup>。DeepSeek 开发了首个通过强化学习训练的大语言模型 DeepSeek-R1<sup>[9]</sup>，其采用混合专家模型（Mixture-of-Experts, MoE）架构，实现 在相同计算成本下大幅提升模型的参数规模与推理能力。

而与 LLM 擅长自然语言类似，还有许多跨模态大模型适用于不同模态间数据的信息表征处理，如文本和图像。由 OpenAI 开发的 CLIP 模型<sup>[31]</sup> 包含一个文本解码器和一个图像解码器，在大量图像及其对应的文本描述的数据集上进行预训练，因而能够把握视觉表征与文本之间的关联。MiniGPT-4<sup>[32]</sup> 是 GPT-4 模型的缩小版，它将一个与 BLIP-2 架构相同的视觉编码器与语言模型 Vicuna 通过一个单一的投影层链接起来，在图像描述生成方面展现出卓越的能力。DeepSeek 开发的 DeepSeek-VL<sup>[10]</sup> 的训练数据来自广泛的现实场景，其采用了一个混合视觉编码器，可在固定的 token 预算内有效处理高分辨率图像，同时保持相对较低的计算开销，这一设计选择确保了模型在各种视觉任务中捕捉关键语义和详细信息的能力。有趣的是，DeepSeek-VL 在一开始便整合了 LLM 的能力训练，促

进视觉和语言两种模态能力的平衡整合，使其在能够捕捉视觉语义信息的同时，仍然具有强大的语言能力。在此基础上，DeepSeek-VL2<sup>[33]</sup> 进一步改进了视觉与文本的能力整合。视觉上融入动态分块的编码策略以处理不同比例尺寸的图像，文本上则引入了 MoE 架构，可动态选择专家模型完成不同的任务。

#### 1.2.4 提示工程

提示工程逐渐作为一种通过自然语言来调整和控制 LLM 行为的技术。目前人们已提出多种提示词的设计原则与策略：Few-shot 框架<sup>[34]</sup> 为模型提供少量示例样本，以有效地指示模型完成指定任务。思维链（CoT）框架<sup>[35]</sup> 指导模型将任务分解为若干子任务来逐步完成，使模型能在无其他修改的情况下完成复杂的因果推理任务。自洽性（Self-Consistency）<sup>[36]</sup> 旨在小样本 CoT 中对多种推理路径进行采样，并在尝试生成之后选择最一致的答案，其有助于提高 CoT 提示在涉及算术推理和常识推理的任务中的性能。此外，开发者们也提出了十分多样的提示词框架，例如 CRISPE<sup>[37]</sup>、ICIO、BROKE，均经过了开源社区的检验，令人目不暇接。

### 1.3 研究思路和主要贡献

本选题旨在利用大模型的通用能力，设计并实现一个文图跨模态的中文古诗创作系统，使其能够：1) 分析输入的图像，生成易使用的描述文本，兼顾关键物体与整体氛围；2) 生成高质量的古诗，同时给出赏析、典故注释和白话文翻译，提高结果的可解释性；3) 对生成古诗进行评分，从多个维度解析古诗的质量，并给出改进建议；4) 对生成的古诗进行改进，在不脱离用户输入需求的前提下进行优化。研究内容如图 1.2 所示。

为处理输入的图像模态，需要使用跨模态能力优良的视觉与文本编码模型，能够为图像生成描述文本，对图像中的物体特征、空间信息和整体氛围进行充分描述。而由于这段图像描述会用于后续的古诗生成，模型还需要能识别出图片中的文化符号和隐含的情感信息，能够与中国传统文化要素产生勾连，输出的描述文本也应当体现出这些信息。

为输出符合格律要求的古诗，需要设计提示词来指导和约束模型的输出。而为了让输出的古诗不仅符合韵律要求，还能展现出古诗的意境和典故，并且在

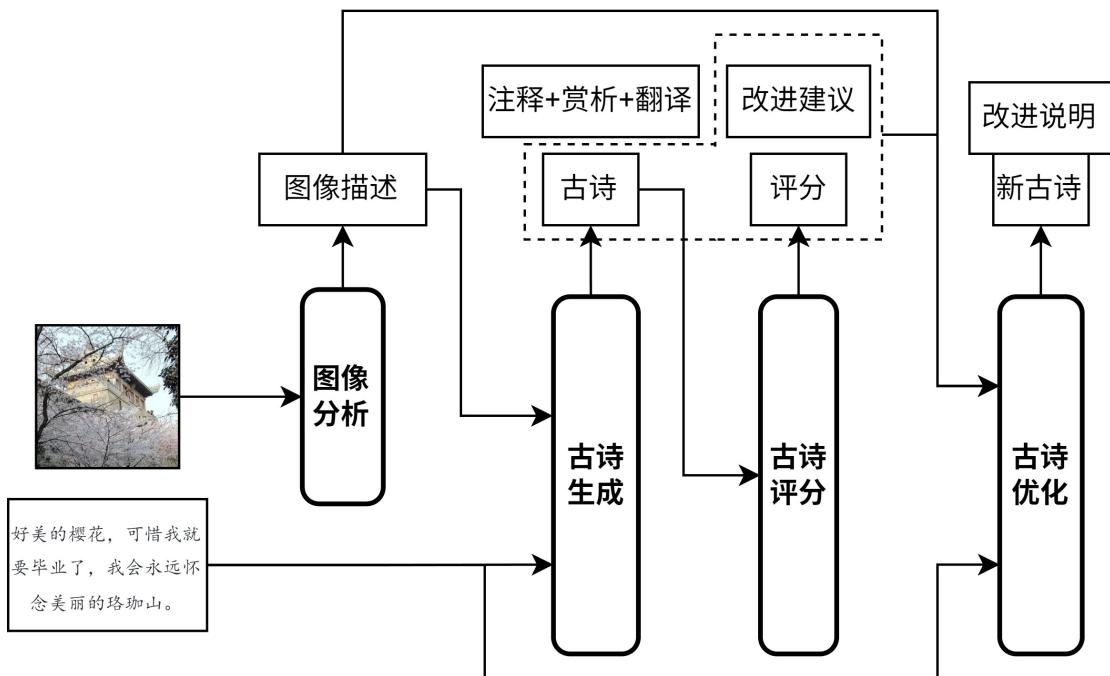


图 1.2 研究内容

注释、赏析、白话文翻译等附加文本中进行用户友好的解释，需要使用对典故意象方面理解与运用能力较强的模型。

为有效地评估生成古诗的质量，需要设计合理科学的评分体系，将古诗的质量划分为不同的维度，对不同维度下的不同给分区间进行举例描述。根据体系获得的评分应当能给出具有区分度的分数，以有效反映古诗的质量，还应针对表现欠佳的维度给出针对性的修改意见，以供后续优化使用。

为有效改进生成的古诗，需要参考改进意见来对古诗进行针对性修改。但同时，为了避免优化时忽略其他方面的考虑、导致其他维度的分数降低，需要同时参考原古诗的评分。同理，为了避免优化过程不偏离用户的原始意图，还需要将文本输入与图像描述纳入考量。如此，系统对原古诗的改进将兼顾作品质量和用户需求。

本文使用 DeepSeek-VL 模型来进行图像分析，相比于英文模型 CLIP 与 MiniGPT4 更能捕捉图像中潜在的文化意味与情感，生成的图像描述文本更具诗意美感，且在古诗生成的质量上也更具优势。古诗生成、评分与优化则使用基于强化学习训练的 DeepSeek-R1 模型，其链式的推理过程包含对多种候选推理路径的动态评估和筛选，能够在反复验证中确保生成古诗的韵律要求，并且也

展现了比 ERNIE-4.0 模型更强的典故意象运用能力。为有效地评估生成古诗的质量，本文设计了一个包含“格律规范”、“意象意境”、“主题思想”、“语言锤炼”、“创新性”等五个维度的评分体系，并附上各分数段的具体例子，有效提升了评分结果的区分度和合理性，也为后续的优化提供了更清晰的方向。而作为补充，本文也将利用 BLEU、ROUGE、Distinct、Similarity 等指标作为辅助评估。

此外，本文还设计了一系列实验来检验系统的功能，包括不同模型生成古诗的能力对比，古诗评分功能的有效性检验，以及文图跨模态输入对生成古诗的影响。通过这些实验，本文验证了系统功能的有效性，并为后续研究提供了参考。

本文的主要贡献在于：1) 结合文图两种模态来强化生成古诗与用户需求的契合度，基于 DeepSeek-R1 输出既符合格律要求、又富有典故意象的高质量古诗；2) 在生成古诗时，输出注释等用户友好的解释性文本，提高了系统输出结果的可解释性；3) 设计维度合理、标准明确的古诗评分体系，实现了可解释性强的古诗评分，且相比于自动度量方法更稳定；4) 结合用户输入和评价结果，实现了兼顾作品质量与用户意图的古诗优化。此外，本文还测试探索不同大模型在图像描述和古诗生成上的表现，尤其是对诗意图美学和意象典故的把握，并讨论了文图跨模态输入对生成古诗的影响。

## 1.4 论文组织结构

第一章是概论部分。

第二章介绍古诗生成任务的现有技术，对古诗的基本要求、自动度量方法、DeepSeek 大模型等进行概述。

第三章介绍本系统的设计与实现，包括系统整体架构、各个模块的功能和实现方法。

第四章介绍开展实验的设计与结果分析，基于已有古诗数据集，结合自动度量方法来评估系统，并展开分析论述。

第五章是结语，对本文的工作内容进行总结，并探讨局限性与改进空间。

## 2 现有技术介绍

本章主要介绍古诗生成任务的基本要求和古诗质量的度量方法，并简要介绍 DeepSeek-R1 与 VL2 两个大模型，为后续的系统设计方案作铺垫。

### 2.1 古诗生成任务概述

古诗生成任务要求模型能够根据需要生成符合要求的中国传统诗词，这要求模型的输出不仅能达到古诗的韵律、对仗等形式要求的建筑美，还要在内容上符合古诗的意境、情感等内涵要求。

古诗体裁多变，产生于汉朝前的古体诗形式自由、不拘泥于严格的格律，唐朝的近体诗（如律诗、绝句）与词则都格律严格。而在古诗生成任务中，往往会选择格律严格的近体诗或词作为生成目标，以便评估模型约束输出的能力。格律要求大概如下三点：

1. 押韵：指诗中某些句子的末尾字使用相同或相近的韵母，形成韵脚和谐的音韵效果。律诗和绝句均要求在偶数句的句末押韵，首句可押可不押，且通常押平声韵。在唐诗中，往往要求一韵到底，即整首诗只能使用同一个韵部的字来押韵，中途不可换韵。
2. 平仄：指汉字声调的高低，分为平声和仄声，在现代汉语中大致对应为一二声和三四声。在唐诗中，五言与七言各有不同的格律格式，如五言诗有四种基本的平仄类型，分别为“平平平仄仄”“仄仄平平仄”“仄仄仄平平”“平平仄仄平”，不同类型可在同一首诗中交替使用。此外，律诗的偶数句中的第二字须与前一句的第二字平仄相同，称为“粘连”<sup>[38]</sup>。
3. 对仗：指律诗中颔联（第三、四句）与颈联（第五、六句）在词性、语义、平仄等结构上形成对称关系，这也是律诗重要的特征之一。如“大漠孤烟直，长河落日圆”，前后两句的词性、语义、平仄均形成对称关系，朗朗上口。

排律是格律的变体，其延续律诗严格的格律要求，但篇幅较长，通常在十句以上，不过押韵要求较为宽松，中途可换韵。排律的篇幅和结构更为复杂，十分考验创作者的文学功底和创作技巧，因而在传统诗歌中，排律的作品相对较少。

## 2.2 古诗质量评估

古诗质量评估有多种方法，此处仅介绍与本文相关的自动度量方法以及人工评估方法，这也是相关研究中常用的方法。

### 2.2.1 BLEU

BLEU (Bilingual Evaluation Understudy)，又名双语替换测评，是一种用于评估机器翻译质量的指标，核心是通过比较机器翻译结果与参考翻译之间的  $n$ -gram 匹配情况来评估翻译质量，并以精确率 (Precision) 作为衡量指标<sup>[24]</sup>。

具体而言，其计算候选句子中在参考句子中出现的  $n$ -gram 的次数  $Count$ 。而为了避免导向无意义的  $n$ -gram 重复，BLEU 对候选句子中的  $n$ -gram 计数进行截断 (clip)，使同一个  $n$ -gram 的计数不超过参考句子中该  $n$ -gram 的最大次数，即  $Count_{clip} = \min\{Count, Max\_Ref\_Count\}$ 。

而在含有多个句子的长文本段落中，BLEU 的处理单元依旧是其中的句子。对候选段落中的每个句子  $C$ ，计算其所有的  $n$ -gram 的截断后计数  $Count_{clip}$ ，再按句子累加在一起，整体除以同理得来的非截断计数  $Count$ ，于是得到一个归一化的分数，以适用于不同长度的文本比较。修正后的精确率如式 (2.1)：

$$p_n = \frac{\sum_{C \in \{Candidate\}} \sum_{n\text{-gram} \in C} Count_{clip}(n\text{-gram})}{\sum_{C \in \{Candidate\}} \sum_{n\text{-gram} \in C} Count(n\text{-gram})} \quad (2.1)$$

为了结合不同  $n$  取值下  $n$ -gram 的指标分数，对不同的  $n$ -gram 分数进行加权几何平均，其中权重  $\sum \omega_i = 1$ ，如式 (2.2)：

$$\exp\left(\sum_{n=1}^N \omega_n \log p_n\right) \quad (2.2)$$

进一步，为了避免机器翻译生成过短的句子来提高匹配的精确率，引入了简洁性惩罚 (Brevity Penalty, BP)，对候选文本  $C$  的长度  $c$  与参考文本  $R$  的长度  $r$  进行比较，计算方式如式 (2.3)。

$$BP = \begin{cases} 1, & c > r \\ e^{1-\frac{r}{c}}, & c \leq r \end{cases} \quad (2.3)$$

最终可得 BLEU 的完整公式如式 (2.4)：

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^N \omega_n \log p_n\right) \quad (2.4)$$

BLEU 的值范围在  $[0, 1]$  之间，值越大表示翻译质量越高，但绝对的数值意

义不大，因此不需要追求接近于 1 的分数。

在实际使用中，权重  $\omega_i$  通常设置为  $\frac{1}{N}$ ，即  $n$ -gram 的加权平均。例如，BLEU-2 的权重设置为  $\omega_1 = \omega_2 = 0.5$ 。古诗中的词大多是一到两个字，因此古诗生成任务中通常使用 BLEU-1 和 BLEU-2。

## 2.2.2 ROUGE

与 BLEU 类似，ROUGE（Recall-Oriented Understudy for Gisting Evaluation）同样通过比较候选文本与参考文本之间的重叠  $n$ -gram 来衡量文本翻译的准确率，但又更适用于文本摘要等注重信息提取和保留的任务<sup>[25]</sup>。

区别于 BLEU 适用准确率，ROUGE 使用召回率（Recall）作为衡量指标，计算候选文本  $C$  中重叠的  $n$ -gram 的数量与参考文本  $R$  中  $n$ -gram 的数量之比。如式 (2.5)

$$\text{ROUGE-n} = \frac{\sum_{C \in \{\text{Candidate}\}} \sum_{n\text{-gram} \in C} \text{Count}_{clip}(n\text{-gram})}{\sum_{R \in \{\text{Reference}\}} \sum_{n\text{-gram} \in R} \text{Count}(n\text{-gram})} \quad (2.5)$$

此外，ROUGE 也支持多种变体，如 ROUGE-L 基于参考文本和候选文本之间的最长公共子序列（LCS）来计算召回率，如式 (2.6)。

$$\text{ROUGE-L} = \frac{\sum_{C \in \{\text{Candidate}\}} \text{LCS}(C, R)}{\sum_{R \in \{\text{Reference}\}} \text{Length}(R)} \quad (2.6)$$

## 2.2.3 Distinct

除了 BLEU 和 ROUGE 等基于参考文本对比的度量方法外，也有方法尝试独立地衡量文本自身的质量。其中一个例子是 Distinct 指标，其计算文本中独特的  $n$ -gram 的数量与文本中所有  $n$ -gram 的数量之比，以衡量文本中用词的多样性。而同理于 ROUGE，常常选取 1-gram 和 2-gram 来计算 Distinct 指标<sup>[26, 39]</sup>。

## 2.2.4 Similarity

为了衡量古诗中前后句子间的语义联系和一致性，也有研究尝试使用词向量，并基于余弦相似度来计算句子间的语义相似度。这些研究由于输出固定为绝句（共四句诗），因此可固定计算头两句相似度 Sim12、后两句相似度 Sim34、以及二者的相似度 Sim2L<sup>[28, 30]</sup>。

### 2.2.5 人工评估

由于古诗体裁的高度的艺术性，过去的研究中往往会邀请人类评审来评估古诗的质量。评估往往基于单独设计的分析角度进行，如“流畅性”、“艺术性”、“连贯性”等等，依赖于分析维度的先验设计。此外，人类评审的结果往往依赖于评审员自身的文化素养、个人品味与喜好，结果难有一致性。此外，招募具有高文学素养的人类评审员也是一个难题。

由此，本文认为可利用大模型的语言能力行使人类评审员的功能，设计一套严格的质量评估体系，作为提示词指导大模型，便可得到具有高解释性、高一致性的古诗质量评估结果。已有部分研究尝试过使用大模型来进行古诗评判，但并问深入探索大模型的评判能力<sup>[40]</sup>。

## 2.3 DeepSeek 大模型

下面概述 DeepSeek-VL2 和 DeepSeek-R1 两个大模型的技术特点。

### 2.3.1 DeepSeek-VL2

DeepSeek-VL2<sup>[33]</sup> 采用了三阶段混合架构，包括视觉编码器、视觉-语言适配器和混合专家语言模型三大模块。

1. 视觉编码器（Vision Encoder）：为了支持更大分辨率、不同比例尺寸的图像，在先前 VL1 模型 SigLIP 框架的基础上引入动态切片策略（dynamic tiling strategy），将高分辨率图像自适应分割为  $384 \times 384$  的子图块，再与全局缩略图块组合。该策略在保证宽高比不变的前提下，通过填充面积最小化算法选择最优分割方案，解决了传统固定分辨率编码导致的细节丢失问题，使模型支持最高  $1152 \times 1152$  的分辨率输入。如图 2.1 所示。
2. 视觉-语言适配器（VL Adaptor）：承接上一步中加入的<换行Token>（用户区分局部图块的结束）和<视觉分隔 Token>（用于区分全局缩略图和局部图块），采用双层感知机与  $2 \times 2$  像素混洗（pixel shuffle）操作，将视觉特征维度压缩映射到文本嵌入空间。该设计在保留局部细节特征的同时，实现了跨模态特征的高效对齐。

图 2.1 DeepSeek-VL2 动态切片策略<sup>[33]</sup>

3. 混合专家语言模型：采用稀疏激活的专家混合架构，每个输入 token 动态激活 Top-2 专家网络。结合多头潜在注意力 (Multi-head Latent Attention, MLA) 机制，通过奇异值分解键值缓存压缩为潜在向量，使 4096 长度序列显存占用降低至传统架构的 6.7%

DeepSeek-VL2 的训练采用了三阶段训练范式，即先使用 120 万个图文对来建立跨模态关联，再使用混合 70% 的图文数据和 30% 纯文本数据来进行预训练，最后再专注于 OCR 增强和文档理解方面的监督微调。值得一提的是，DeepSeek-VL2 的 MoE 架构包含有 64 个专家分组，每组都能处理特定的模态组合，并能在训练中通过负载均衡损失函数来优化专家激活分布，实现稀疏激活，从而大大提高推理效率（在总体 176B 的参数量下仅激活 4.5B 参数）。

### 2.3.2 DeepSeek-R1

为了训练模型的推理能力，过去的方法通常是在监督微调 (supervised fine-tuning, SFT) 后加入大量的思维链 (CoT) 范例数据，引导模型学会链式的推理思考。但 DeepSeek-R1<sup>[9]</sup> 则采用了强化学习 (RL) 的路径来训练模型推理能力。

在正式构建 DeepSeek-R1 前，DeepSeek 团队先尝试验证强化学习方向的可行性——直接在基底模型上应用强化学习，而不使用任何 SFT 的数据，通过准确性奖励和格式奖励来训练出 DeepSeek-R1-Zero。此外，为节省训练开销，采用相对策略优化 (Group Relative Policy Optimization, LGRPO) 算法，通过计算组内输出结果的得分均值来获得整个损失函数的期望值。

而 DeepSeek-R1 的训练过程分为四个阶段（如图 2.2）。

1. 冷启动 SFT：为了解决 RL 训练早期的不稳定性和语言混杂的问题，先采用较小规模的 CoT 数据集，对基底模型 DeepSeek-V3-Base 进行冷启动 SFT，作为 RL 训练的初始模型。这里使用数据是带有反思和验证的详细思考答案，由 DeepSeek-R1-Zero 生成。
2. 推理导向 RL：流程与训练 DeepSeek-R1-Zero 一致。此外，为了减少语言混杂的问题，引入了语言一致性奖励，以性能略微下降为代价，提高模型输出的可读性。
3. 拒绝采样 SFT：利用上一阶段 RL 训练过程中的 checkpoint 进行拒绝采样，生成多个候选的推理轨迹，再利用 DeepSeek-V3 充当奖励模型来进行打分，仅保留评分最高的样本，以此获得 60 万推理数据。此外，还纳入了 DeepSeek-V3 的部分训练数据，并对部分任务进行提示生成，获得 20 万非推理数据。于是模型在大小为 80 万的数据集上进行两轮微调。
4. 全场景 RL：为进一步实现人类偏好对齐，再次进行 RL 训练，使用基于规则的奖励来训练推理方面的学习，使用奖励模型来训练通用领域中较隐晦的人类偏好。

此外，可直接使用“拒绝采样 SFT”中使用的 80 万数据集对其他较小模型（如 Qwen2.5-14B 和 Llama3.1-8B）进行 SFT，得到蒸馏模型 DeepSeek-R1-Distill。需要注意，这一过程并未包含强化学习训练。

## 2.4 本章小结

本章介绍了古诗生成任务的基本要求，包括押韵、平仄、对仗等格律规范，以及 BLEU、ROUGE 等自动评估指标和人工评估方法。针对评估的主观性问题，提出可利用大模型构建更稳定的评判体系。同时概述了 DeepSeek-VL2 和 DeepSeek-R1 两大模型的技术特点：VL2 采用混合专家架构实现多模态理解，R1 通过强化学习优化推理能力，采用分阶段训练策略提升生成质量。这些内容为后续系统设计奠定了基础。

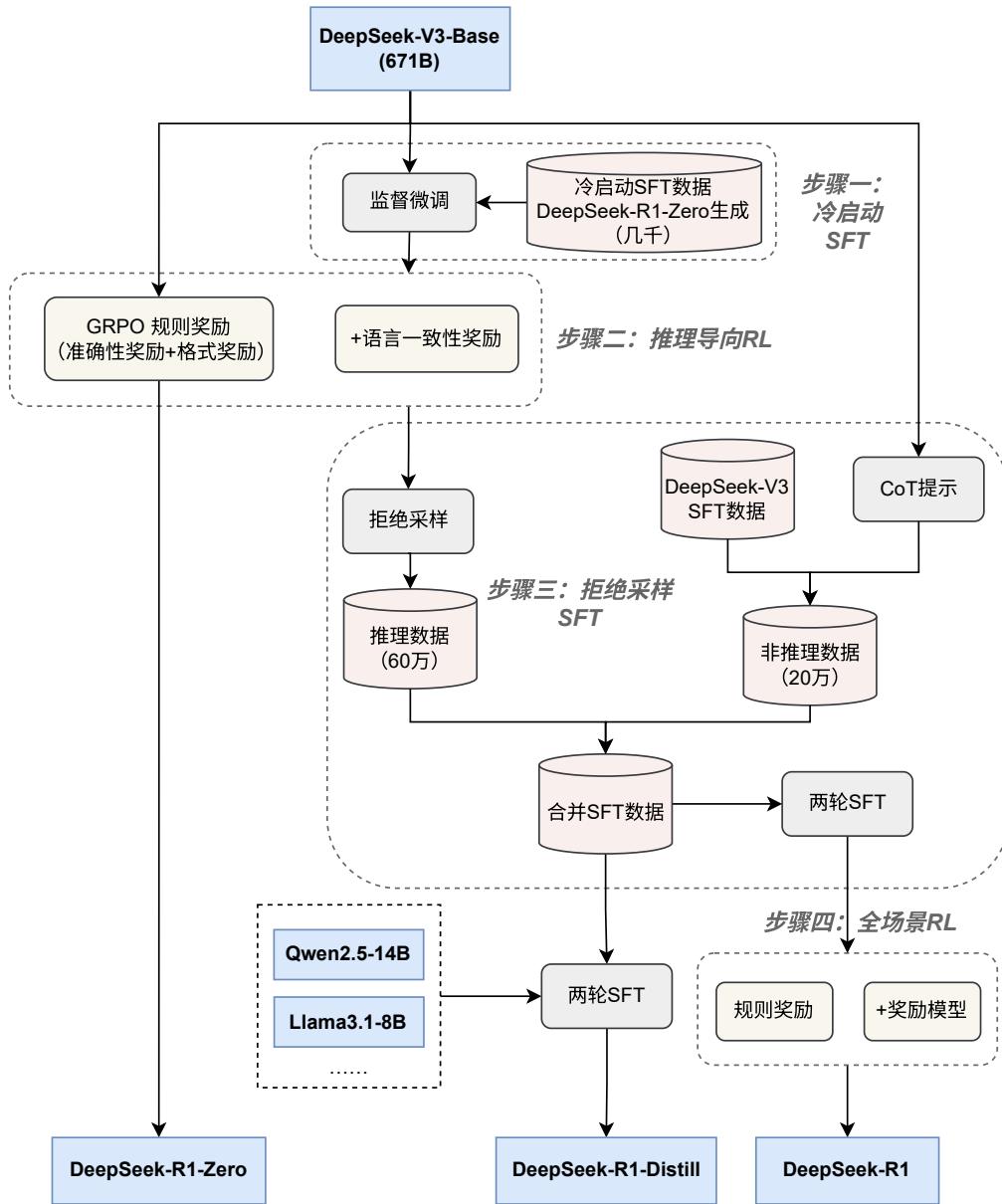


图 2.2 DeepSeek-R1 训练过程

### 3 古诗生成与优化

本章主要介绍本文对古诗生成优化的功能设计，以及系统的模块设计，包括系统的整体架构、各个模块的功能以及设计的考量。

#### 3.1 系统架构概述

本系统包含图像分析、古诗生成、古诗评价和古诗优化四个模块，基于 Python 语言开发，使用百度智能云提供的 API 接口来调用模型，整体架构如图 3.1 所示。

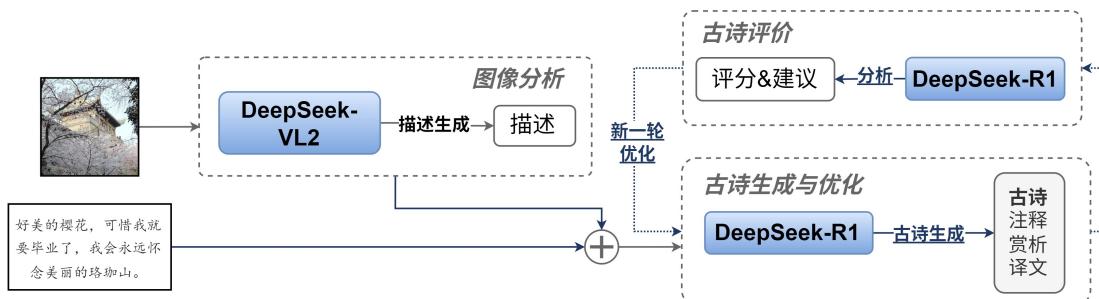


图 3.1 系统架构

在调用 DeepSeek-R1 模型时，需要配置指定的参数（见表 3.1），其中需要基于 API Key 鉴权，将 ‘Bearer’ 与 API KEY 拼接起来作为鉴权字符串（V2 版调用策略，编码时也尝试了 V1 版的安全认证 AK 和 SK 鉴权，但不如 API Key 便捷）。而请求体的 ‘messages’ 中以列表形式存储对话的上下文记录，每份记录为字典形式，包含 ‘role’ 与 ‘content’ 两部分，分别表示消息主体（如系统 ‘system’ 或用户 ‘user’）和消息内容。完成参数设置后，通过 POST 方法对百度智能云的 URL 提交访问请求。

表 3.1 百度智能云 API 请求参数

| 参数类型    | 参数名                           | 参数说明              | 参数取值                                     |
|---------|-------------------------------|-------------------|--|
| Headers | Content-Type<br>Authorization | 请求体的数据格式<br>鉴权字符串 | ‘application/json’<br>‘Bearer <API_KEY>’ |
| Body    | model<br>messages             | 调用模型名<br>聊天上下文信息  | ‘deepseek-r1’<br>Array 类型                |

对 DeepSeek-VL2 模型的调用略有不同，由于其需要对图像进行分析，其中 ‘messages’ 的 ‘content’ 换成了元素为字典的列表，其中分别包含文本输入 ‘text’ 和图像输入 ‘image\_url’。需要注意，DeepSeek-VL2 的调用并不支持直接传输图

像数据，而是需要提供一个公开访问的图像 URL，还需要将用户提供的图像上传到云端。

本文使用阿里云的对象存储 OSS 服务实现图像的云存储，在控制台创建新的存储单元 Bucket 后，利用 OSS Python SDK 进行编码，实现图像的上传、下载和 URL 生成。具体而言，在将用户图像上传到云端后，编码生成带有过期时间（默认值 600 秒）的 GET 方法预签名 URL。将其作为请求体的图像 URL 参数，供后续 API 调用使用。

## 3.2 图像分析

为处理输入的图像模态，需要使用跨模态能力优良的视觉与文本编码模型，使其能够为图像生成涵盖物体特征、空间信息和整体氛围的描述文本。同时，还应当能识别出图片中的文化符号和隐含的情感信息，与中国传统文化要素产生勾连，便于后续古诗生成。

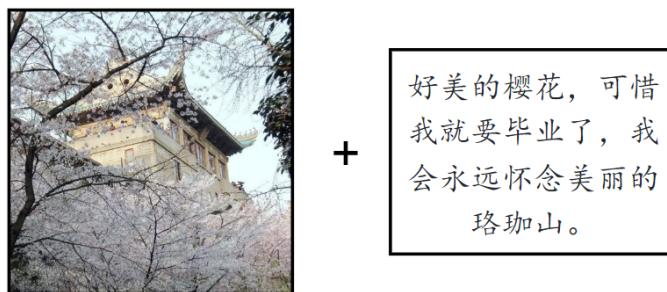


图 3.2 示例输入

为此，本文选用 DeepSeek-VL2 模型<sup>[33]</sup>进行图像分析。举例而言，对图 3.2 中的图像输入，使用图 3.3 的提示词，DeepSeek-VL2 模型的输出如图 3.4 所示。

请描述这张图片，注意要明确提及图像中的物体，描述清楚物体的色彩、大小、相对位置等基本信息，并兼顾整体的情感色彩，确保读者能够根据描述在心里构建出一个清晰的画面。请确保使用中文，不超过 7 句话，并使用一个段落完成。

图 3.3 提示词（图像分析）

在选定 DeepSeek-VL2 模型之前，本文曾尝试使用 CLIP 和 MiniGPT-4 两个

这张图片展示了一座传统中式建筑，其屋顶装饰着精美的绿色琉璃瓦，显得庄重典雅。建筑主体为浅黄色，四周有栏杆围绕，似乎有游客站在阳台上欣赏风景。前景是繁茂的樱花树，粉白色的花朵覆盖了大部分画面，营造出浪漫而宁静的氛围。天空清澈，与花朵的柔和色彩形成鲜明对比，增添了画面的层次感。整幅画面让人感受到春天生机勃勃的气息，同时也透露出一种悠闲和诗意的情感。

图 3.4 图像描述文本（示例）

英文模型组合，分别关注图像中的关键物体和整体描述，并最后交由 ERNIE-4.0 模型总结为中文描述。尽管生成的描述较精确、但无法捕捉有助于古诗创作的中国文化联想素材与情感色彩。举例而言，对图 3.2 中的图像输入，先前的工作流输出结果如下

**CLIP:** Xinyang, Yunfu, Zhoukou, Blossom, Pagoda, Teahouse

**MiniGPT4:** The image shows a large, ornate building with a tall, pointed roof and a large, open balcony on the second floor. The building is surrounded by a large, flowering cherry tree, which is in full bloom and covered in pink blossoms. There are several people standing on the sidewalk in front of the building, looking up at the cherry tree. The sky is clear and blue, with a few clouds in the distance. The building appears to be made of stone or brick, with intricate carvings and decorations on the facade. There are several other buildings in the background, some of which are also covered in blooming cherry trees. The overall scene is one of beauty and serenity.

**ERNIE-4.0:** 图中展示了一座装饰繁复的大型建筑，其拥有高耸的尖顶 和二楼宽敞的开放式阳台。建筑被一棵盛开的樱花树环绕，满树繁花，粉白交错。建筑前的人行道上，数人驻足，仰头欣赏樱花的美丽。晴空万里，远处的几朵云彩点缀在蓝天之中。这座看似由砖石砌成的建筑，立面雕刻精美，装饰繁复。背景中还有其他几座建筑，其中一些同样被盛开的樱花所簇拥。整幅画面洋溢着美丽与宁静的气息。

可以发现，虽然最终的文字准确描述了画面内容，但依旧无法生成符合古诗创作需求的文化意象：

- CLIP 误读文化符号：CLIP 输出的“Teahouse”（茶室）与“Pagoda”（塔）形成文化

错位，这一点源于 CLIP 的训练数据分布偏差——在西方语料库中，“pagoda”常与日本茶道场景关联，而中文语境下塔与茶室并无必然联系，况且画面中的建筑并不是塔。更严重的是，模型甚至输出了“Xinyang”（信阳）、“Yunfu”（云浮）等意义难明的拼音，暴露了其对中文地理文化符号的认知缺失。

- MiniGPT4 直译式描述：MiniGPT-4 虽然能识别“樱花树”“尖顶建筑”等元素，但描述方式太平直白。例如将中式建筑的“飞檐”简单说成“pointed roof”（尖顶），而将樱花描述为“pink blossoms”（粉色花朵），这种直译式描述丢失了文化内涵。

于是，当 CLIP 和 MiniGPT4 的输出交由 ERNIE-4.0 来总结时，最后的结果也难免停留在缺失文化内涵的直译文本，“只识其形，不知其意”，无法进一步提供与中国传统文化要素的关联。

因此，本文尝试直接使用中文语境下的图像大模型替代之前的多模型组合方案，一步到位地为图像生成兼顾关键物体识别、整体场景信息和情感色彩的描述。

使用相同的图像输入，结合简单的提示词（见图 3.3）对 DeepSeek-VL2 和 Qwen2.5-VL-3B 这两个中文图像模型进行初步测试，发现前者的表现要好于后者。DeepSeek-VL2 的输出如前文图 3.4 所示，而 Qwen2.5-VL-3B 的输出如下：

**Qwen2.5-VL-3B:** 这张图片展示了一座传统的中国建筑，可能是寺庙或宫殿。建筑的屋顶是典型的翘角设计，颜色为黄色和绿色。建筑周围有许多盛开的樱花树，花朵呈现出粉红色，与建筑的色彩形成了美丽的对比。树木的枝叶繁茂，覆盖了大部分画面。在建筑的阳台上，可以看到一些人在欣赏风景。整体氛围宁静而美丽，给人一种春天的感觉。

可以看到，DeepSeek-VL2 的输出描述了更具象的色彩意象，而 Qwen 的输出缺只提供视觉要素的平面罗列。以色彩为例，前者的输出包含有“绿色琉璃瓦”、“浅黄建筑”、“粉白樱花”等精确的色彩描述，而后者只模糊地提及“黄色和绿色”。而在空间层次上，前者通过“前景樱花-中景建筑-远景天空”的立体构图清晰地描述出了图像中的立体信息，而后者却丢失了这份空间关系的描述。

因此，本文使用 DeepSeek-VL2 作为图像描述模型，一步到位为用户输入的图像生成充分体现空间信息、具有中式美感的描述文本，为进一步的古诗创作

提供充分的视觉感官描述。

### 3.3 古诗生成

为了优化用户体验、增强系统的可解释性，需要生成意象丰富的高质量古诗，并给出注释、赏析、白话文翻译等解释性文本。以帮助用户理解古诗的创作内涵，增强结果的可解释性。

为此，本文选用基于强化学习训练的 DeepSeek-R1<sup>[9]</sup> 模型，生成不仅满足韵律规则、还具有丰富意象的高质量古诗。对图 3.2 中的输入，系统输出的古诗、白话文翻译、注释和赏析分别如图 3.5、图 3.6、图 3.7 和图 3.8 所示。

珞珈山下琼英绽，朱阁檐前翠影摇。  
万点飞花迷画槛，千重香雪覆虹桥。  
琉璃映日浮金阙，玉砌连云接碧霄。  
曲径曾携书卷过，长亭忍看落英飘。  
东君不解离人恨，更遣残红逐晚潮。  
粉泪暗随春水逝，青衿空对暮山遥。  
他年若问芳菲事，犹记晴川鹤影娇。

图 3.5 生成古诗（示例）

珞珈山下樱花绽如琼瑶，朱红楼阁前翠影正轻摇。  
万点飞花迷蒙了雕栏画柱，千层香雪压弯了求学虹桥。  
琉璃瓦映日似浮动金殿，白石阶入云直通向碧霄。  
曾捧书卷走过这林荫曲径，长亭外怎忍看落英飘摇。  
春神不懂人间别离苦楚，偏让残红追着晚潮远漂。  
粉泪默默随春水流逝去，青衫空对暮色里的山遥。  
若问当年花开烂漫旧事，定记得晴川上鹤影正娇。

图 3.6 古诗白话文翻译（示例）

由于任务要求较多，且系统对模型输出内容的格式规范有严格要求，需要精心设计提示词来确保模型按照预期的格式输出想要的内容。而 CRISPE 框架<sup>[37]</sup>通过模块化的设计将复杂的提示工程分解为不同的组件，覆盖了提示词设计的多个方面，使得提示词的设计更加清晰、易于理解和维护。具体而言，CRISPE 框架包括六个部分：

1. 能力与角色（Capacity and Role）：大模型应当具备的角色与能力。

1. 琼英：取自《诗经·齐风》“尚之以琼英乎而”，既指樱花如玉，又喻学子才俊
2. 朱阁：对应图片中浅黄色建筑，化用苏轼“转朱阁”意象，暗喻学术殿堂
3. 虹桥：既实指校园拱桥，又暗用《东京梦华录》虹桥典故，喻求学之路
4. 琉璃：特指图片中绿色琉璃瓦，呼应白居易“琉璃波面月”的华美意象
5. 玉砌：化用李煜“雕栏玉砌应犹在”，将现代建筑古典诗化
6. 东君：司春之神，此处双关东风与毕业季的时令特征
7. 青衿：源自《诗经》“青青子衿”，既指学子装束，又喻求知初心
8. 晴川鹤影：化用崔颢“晴川历历汉阳树”，“鹤”取《诗经》“鹤鸣于九皋”意象，喻志向高远
9. 香雪：借李商隐“匝路亭亭艳”咏樱花之意，特指白色樱花如雪
10. 画槛：对应图片中栏杆意象，暗用“画栋朝飞南浦云”的滕王阁典故

图 3.7 古诗注释（示例）

这首七言排律以武大樱花为情感载体，通过工丽的对仗与典雅的意象，构建出毕业离别的双重意境。首联以“琼英”“翠影”点题，既暗合《诗经》“尚之以琼英乎而”的典故，又呼应图片中琉璃瓦与樱花的色彩映照。颔联“万点飞花”与“千重香雪”形成空间纵深感，化用岑参“千树万树梨花开”的意境而翻出新意。颈联转写建筑，“琉璃映日”直取图片意象，“玉砌连云”暗用李煜“雕栏玉砌”之典，将现代学府与传统意象完美融合。尾联“晴川鹤影”既化用崔颢“晴川历历汉阳树”，又暗喻学子如鹤高飞，留下余韵悠长的想象空间。全诗严守平水韵，中二联“迷画槛”“对”“覆虹桥”，“浮金阙”“对”“接碧霄”，工稳中见灵动，以物象变迁写时光流逝，最终在“鹤影”的意象中完成对母校记忆的诗意图定格。

图 3.8 古诗赏析（示例）

2. 背景信息（Insight）：为完成任务，大模型应当知晓的背景知识信息，以及用户需求的上下文语境。
3. 指令（Statement）：大模型需要完成的任务。
4. 输出风格（Personality）：大模型输出回复的风格、特色以及规范。
5. 实验（Experiment）：尝试让大模型提供一些例子，以便更好地调试提示词。

其中，最后投入使用的只包括前五个部分，最后一个部分“实验（Experiment）”只是作调试用，方便提示词的设计迭代。

参考这样的框架，本文为古诗生成设计的提示词可见图 3.9，其中系统接收用户的文本输入user\_text，与用户输入图像的描述description结合，同时指定古诗的体裁poem\_type（如五言绝句、七言律诗、不少于十句的排律等）。之后在“输

出风格”部分，系统指定模型必须以 JSON 格式输出内容，包含古诗的标题、内容、赏析、注释以及白话文翻译，并且规定必须不输出其他的冗余内容，以便于直接转化为 JSON 对象进行后续处理。

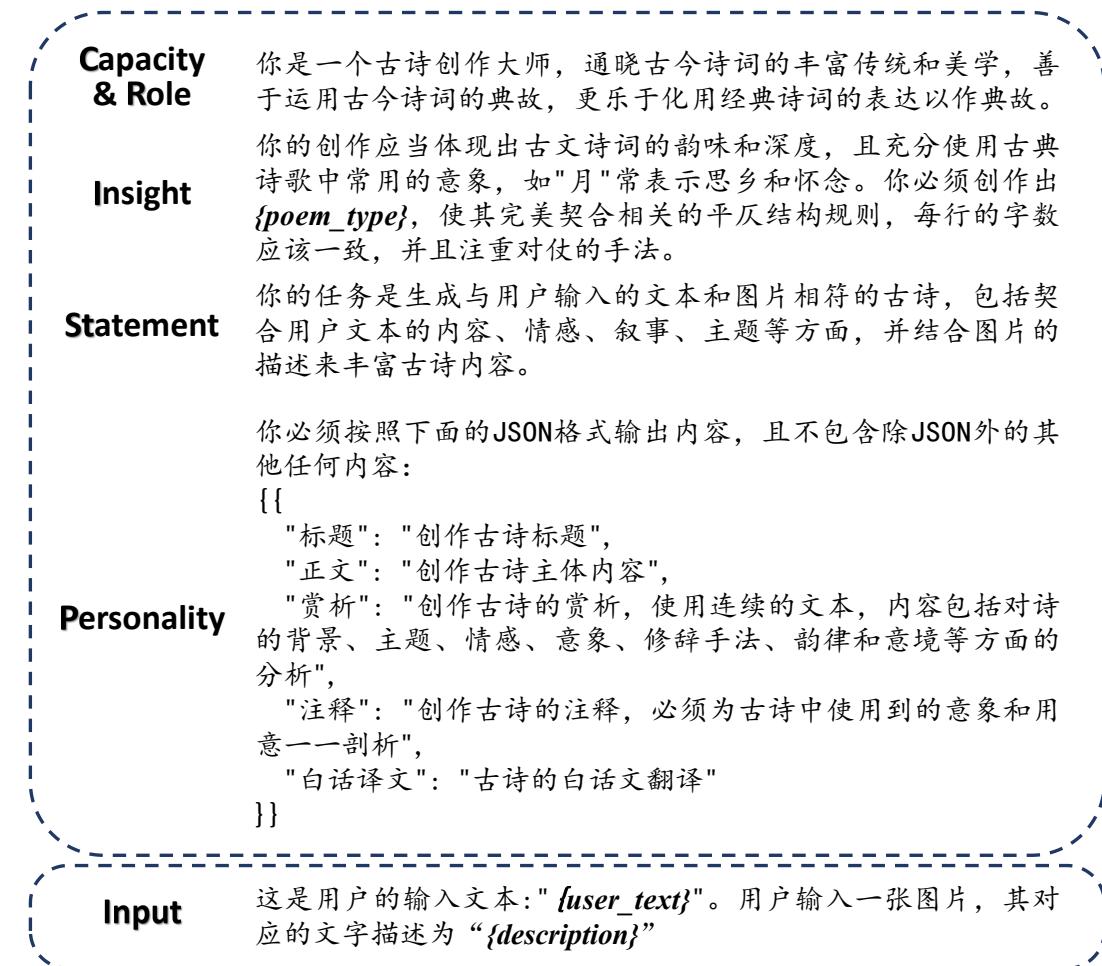


图 3.9 提示词（古诗生成）

修改提示词后，本文尝试了多种大语言模型，包括 DeepSeek-V3<sup>[9]</sup> 和先前工作尝试的 ERNIE-4.0，在控制输入条件相同的条件下，分别进行古诗生成。输入文本与图 3.2 中一致，图像描述使用图 3.4 的内容，古诗体裁为“不少于十句的排律”，以更长的输出长度来考察模型输出结果的参差。

经过初步对比测试，发现 ERNIE-4.0 和 DeepSeek-V3 在意象理解和运用上的表现都显著弱于 DeepSeek-R1，这一点尤其体现在“注释”的输出上（“注释”能十分直观地反映模型输出的古诗在典故运用方面的表现，相比之下“赏析”和“白话文翻译”两个部分难以轻易看出差异）。这两种模型输出的“注释”内容部分如下：

#### **ERNIE-4.0:**

珞珈山：指用户所怀念的校园所在地，此处借指校园。

樱如雪：形容樱花盛开时如雪般洁白美丽。

绿瓦黄墙：描述传统中式建筑的屋顶绿色琉璃瓦和浅黄色墙体。

春华：春天的花，此处泛指春天的美景。

花满袖：形容花香扑鼻，似乎连衣袖都沾满了花香。

.....

#### **DeepSeek-V3:**

珞珈山：位于中国湖北省武汉市，武汉大学所在地，以樱花闻名。

樱雪：形容樱花盛开如雪。

翠微：山色青翠。

琉璃瓦：传统中式建筑屋顶的装饰材料，绿色琉璃瓦显得庄重典雅。

黄墙：传统中式建筑的主体颜色，浅黄色。

.....

可见，图 3.7 中 DeepSeek-R1 的注释内容清晰地揭示了其在原文创作中对典故意象的巧妙运用，这种能力明显优于另外两个模型。例如，“琼英”一词既描绘樱花美感、也隐喻学子的才情与品德，这种双重意象的运用极大地丰富了诗句的文化内涵。同样，“朱阁”一词不仅对应了图片中的建筑特征，还化用了苏轼的诗句，将现代校园建筑与古典意象相融合，展现出学术殿堂的庄重与典雅。这些典故意象的运用并非生硬堆砌，而是自然流畅地融入诗句之中，使得整首古诗在形式上符合排律的规范，内容上则充满了古典韵味与文化底蕴。

相比之下，ERNIE-4.0 和 DeepSeek-V3 的输出虽然也符合要求，但在典故意象的运用上明显逊色。从注释内容上看，它们更多是对景色的直观描述，缺乏对文化典故的深度挖掘和巧妙运用，无法触及更深层次的文化意蕴和诗意图。

因此，本文使用 DeepSeek-R1 作为古诗生成的模型，其在古诗创作中展现出的典故意象运用能力，不仅体现了其对传统文化的深刻理解，也彰显了其在语言生成和文化传承方面的强大优势。

### 3.4 古诗评价

为了评估生成古诗的质量，本系统设计了一套合理详细的古诗评分体系，调用 DeepSeek-R1 模型来进行评分，同时结合 BLEU 等自动度量方法来进行辅助评估。

#### 3.4.1 系统评分

为利用大模型来评估古诗质量，本文结合格律诗的韵律规则与基本的鉴赏方法，设计了一套包含格律规范、意象意境、语言锤炼、文化引用和创新性五个维度的评分体系，每个维度又各自细分若干子维度，并给出相应的分值和评分标准，每个分数段都附有具体的例子，以帮助模型更好地理解评分要求。评分体系如表 3.2 所示。

表 3.2 古诗评分体系

| 维度   | 分值 | 子维度  | 小分 | 备注  |
|------|----|------|----|---|
| 格律规范 | 25 | 平仄音韵 | 10 | 9-10: 完全符合唐体格律(例:杜甫《登高》“风急天高猿啸哀，渚清沙白鸟飞回”平仄严谨) |
|      |    |      |    | 7-8: 个别拗句但有救(例:王维《终南别业》“行到水穷处”第三字拗,第四字救)      |
|      |    |      |    | 5-6: 三平尾/三仄尾不超过两处(例:韦应物《滁州西涧》“独怜幽草涧边生”三平尾)    |
|      |    |      |    | 0-4: 严重失律(例:打油诗体)                             |
|      | 25 | 对仗工稳 | 10 | 9-10: 工对 + 借对精妙(例:李商隐《锦瑟》“庄生晓梦迷蝴蝶, 望帝春心托杜鹃”)  |
|      |    |      |    | 7-8: 宽对但结构平衡(例:王勃《送杜少府》“海内存知己, 天涯若比邻”)        |
|      |    |      |    | 5-6: 词性不对称(例:拙劣仿作“青山对绿水, 饮酒对弹琴”名词对动词)         |
|      |    |      |    | 0-4: 无对仗意识                                    |
|      | 30 | 押韵协调 | 5  | 5: 严格遵循平水韵(例:李白《静夜思》“床前明月光”押下平七阳韵)            |
|      |    |      |    | 3-4: 邻韵通押(例:杜牧《清明》“纷”属文韵, “魂”属元韵通押)           |
|      |    |      |    | 1-2: 出韵超过两处                                   |
|      |    |      |    | 0: 完全无押韵                                      |
|      |    |      |    | 18-20: 传统意象出新境(例:王维《使至塞上》“大漠孤烟直”重构“孤烟”意象)     |
|      |    | 古典运用 | 20 | 14-17: 精准使用经典意象(例:柳宗元《江雪》“孤舟蓑笠翁”的渔父符号)        |
|      |    |      |    | 转下一页  |

表 3.2 古诗评分体系

| 维度   | 分值 | 子维度  | 小分 | 备注  |
|------|----|------|----|---|
| 主题思想 | 20 | 意境层次 | 10 | 10-13: 意象堆砌无深意（例：劣作“残阳古道瘦马，西风落叶昏鸦”）<br>0-9: 意象误用（例：用“东篱”指代监狱） |
|      |    |      |    | 9-10: 多层意境交织（例：李商隐《夜雨寄北》时空折叠技法）                               |
|      |    |      |    | 7-8: 单一意境完整（例：孟浩然《春晓》的晨醒意境）                                   |
|      |    |      |    | 5-6: 意境破碎（例：拼贴“明月松间照，股票涨停板”）                                  |
|      |    |      |    | 0-4: 无意境构建  |
|      |    | 情感真挚 | 12 | 11-12: 情志合一（例：杜甫《月夜》“遥怜小儿女，未解忆长安”的家国之痛）                       |
|      |    |      |    | 9-10: 情感明确但稍显直露（例：高适《别董大》“莫愁前路无知己”）                           |
|      |    |      |    | 6-8: 情感造作（例：仿古风“朕与将军解战袍”）                                     |
|      |    |      |    | 0-5: 情感空洞   |
|      |    |      |    | 7-8: 接通传统文脉（例：苏轼《题西林壁》对禅理的化用）                                 |
| 语言锤炼 | 15 | 凝练度  | 8  | 5-6: 简单模仿前人（例：仿写“采菊东篱下”无新解）                                   |
|      |    |      |    | 3-4: 曲解经典（例：将“仁者乐山”解为爱好登山）                                    |
|      |    |      |    | 0-2: 思想谬误   |
|      |    | 典雅度  | 7  | 7-8: 字字珠玑（例：贾岛《题李凝幽居》“鸟宿池边树，僧敲月下门”）                           |
|      |    |      |    | 4-6: 可删减 1-2 字（例：初稿“推”改为“敲”的炼字过程）                             |
|      |    |      |    | 1-3: 冗余明显（例：劣作“我看到青山高又高，绿水长流流不停”）                             |
| 创新性  | 10 | \    | \  | 6-7: 文白交融自然（例：李清照《声声慢》“寻寻觅觅”的白话感）                             |
|      |    |      |    | 4-5: 文言生硬（例：强行用“之乎者也”凑韵）                                      |
|      |    |      |    | 1-3: 语体混乱（例：夹杂“OK”“Hi”等外来词）                                   |
|      |    |      |    | 9-10: 传统技法新用（例：王安石《泊船瓜洲》“绿”字形容词动用）                            |
|      |    |      |    | 7-8: 有限度创新（例：崔颢《黄鹤楼》前半打破律诗常规）                                 |
|      |    |      |    | 5-6: 为变而变（例：强行改写五绝为六言）  |

基于这套评分体系，系统将逐一分析古诗的每个维度，给出分数和评语，并进一步地根据体系指导，对得分较低的维度逐一地给出修改意见，以提高古诗的质量。（本文对这一评分体系有效性的检验实验见 4.2）

作为对比，之前设计的评分体系包含“结构与形式”(10 分)、“语言与风格”(20 分)、“意象与主题”(30 分)、“协调与一致”(10 分)、“历史语境”(20 分)、“创新

性”（10分）等六个维度，每个方面下又细分若干子维度，给定分值并描述相应维度所考察的具体方面。原先的评分体系如下：

1. 结构与形式（Structural and Formal Aspects） - 总分 10 分

诗歌类型：最高 5 分。评估是否符合指定类型（如律诗、绝句）的基本结构和规则。

韵律规则：最高 5 分。分析诗歌的韵律是否规整，是否符合传统韵律规则。

2. 语言与风格（Language and Style） - 总分 20 分

用字选词：最高 10 分。评估字词的适当性、丰富性、创新性。

修辞手法：最高 10 分。评估修辞的恰当使用和创造性。

3. 意象与主题（Imagery and Themes） - 总分 30 分

意象运用：最高 15 分。评估意象的原创性、适当性和表达效果。

主题深度：最高 15 分。评估主题的深度和情感表达。

4. 整体协调性（Cohesion and Coherence） - 总分 10 分

内在逻辑：最高 5 分。评估诗句之间的逻辑连贯性。

情感连贯性：最高 5 分。评估情感表达的一致性和流畅性。

5. 文化与历史背景（Cultural and Historical Context） - 总分 20 分

文化引用：最高 10 分。评估诗中的文化、历史元素的准确性和适当性。

历史背景适应性：最高 10 分。评估内容与时代背景的契合度。

6. 创新性与原创性（Originality and Innovation） - 总分 10 分

独特视角：最高 5 分。评估提供的新颖观点或表达。

创新手法：最高 5 分。评估在结构、语言或主题上的创新。

可见，原评分体系的叙述已较为成形，有进一步的子维度划分和各自的分值设置，为模型提供评估指导。但这一设计也存在一些问题：

1. 评分标准模糊：原评分体系并未给出具体的评分标准，如不同的表现应当如何对应不同的分数段，这使得评分缺乏清晰的标准，给分也难有区分。以之前的工作为例，ERNIE 模型在对“经典唐诗”、“现代人创作的古诗”和“打油诗”三种不同的数据集进行打分测试时，给出的平均得分约为 0.84、0.82 和 0.67，整体偏高且前两者的区分度很小。

2. 评判角度解释不足：原评分体系中的描述十分模糊，如诗韵律规则中仅提

及“符合传统韵律规则”，却并未提供任何对韵律规则的说明；而“创新性”更是抽象，究竟怎样的创新是出气而不是出糗、是精妙而不是作秀，并未被详细描述。这使得模型在评分时缺乏明确的指导，评分结果也尤其依赖模型自身的古诗知识储备，而非评分系统自身的设计。

3. 维度设置欠妥：原评分体系中的各个维度存在内容重叠，如“结构与形式”中对格律规则的要求同样是“整体协调性”的要求，而“意象与主题”与“文化与历史背景”两个维度均涉及到文化元素的运用，更抽象的“创新性与原创性”也与其他维度存在交叉，这使得评分时难以明确区分各个维度的具体考察内容，给出的分数也难以反映出古诗的真实质量。

为了使其能够给出合理、细致的评分，本文设计的新评分体系针对这三个问题进行了改进：

1. 细化评分标准：每个子维度都设置了明确的分数段对应具体表现。如格律规范中的平仄音韵，9-10分对应“完全符合唐体格律”，并附杜甫《登高》为例；5-6分则明确限定“三平尾/三仄尾不超过两处”，避免了旧体系中“符合传统韵律规则”的模糊表述。这种量化标准使模型在评估古诗作品时，能通过平仄失误数量（如三平尾出现次数）给出更具区分度的评分。
2. 增强解释性：对每个评分层级都配有典型诗例说明。以“对仗工稳”为例，7-8分的“宽对但结构平衡”引用王勃《送杜少府》的“海内存知己，天涯若比邻”。相比于旧体系仅标注“评估修辞恰当性”，这种具象化指引降低了模型对自身知识储备的依赖，即使面对“借对”等专业技法，也能通过“李商隐《锦瑟》‘庄生晓梦迷蝴蝶’的示例进行对标判断<sup>[41]</sup>。
3. 重构维度逻辑：将旧体系重叠的“结构与形式”“整体协调性”整合为“格律规范”单一维度，涵盖平仄、对仗、押韵三项核心要素；把分散在“文化引用”“意象运用”中的古典意象评估，集中到“意象意境”维度下的“古典运用”指标。同时新增“语言锤炼”维度，专门评估旧体系未明确涉及的凝练度与典雅度问题。这种重构使“创新性”能更纯粹地聚焦于“守正出新”的诗学突破，避免与修辞手法等基础要求混淆。

由于表3.2中的评分体系较复杂、所占文本较多，容易干扰提示词中的其他需求描述，需要通过范例来达成保证输出内容和格式的作用。于是提示词的设

计参考 Few-shot 框架<sup>[34]</sup>，在完成需求说明后，给定两个输入与输出的范例，确保模型按照预期的格式输出内容。其中提示词框架见图 3.10，示例见图 3.11 和图 3.12。

下面是对创作古诗的详细评分规则，我将提供古诗文本给你，请你严格基于以下进行评价和打分。在完成评分后，我还需要你提供对古诗文本的具体改进建议，同样依照这些角度来一一说明，特别是得分较低的角度。注意输出只包含“评分”与“改进建议”两部分，且须严格遵循格式要求。

<评分体系>

<输入输出示例 1>

<输入输出示例 2>

图 3.10 提示词（古诗评分）

冬日寒风吹雪霜，心驰神往樱花乡。花开花落春又至，不见花影空自伤。昨夜梦中花开满，今朝空余泪沾裳。唯愿春风吹花醒，再现人间美景光。

[评分]

### 1. 格律规范 (17/25)

1.1 平仄音韵 (7/10)：颔联“花开花落春又至”出现三平尾（春又至），像连续敲了三个重鼓（三个平声），传统认为这样太单调。颈联“今朝空余泪沾裳”存在孤平现象，像独唱时突然少了个伴奏（缺少平仄交替），听着不够流畅。总体符合近体诗框架但存在局部失律

1.2 对仗工稳 (5/10)：领颈两联对仗失衡，“花开花落”与“不见花影”结构失对，就像用“高山流水”对“大海捞针”——前两句讲变化，后两句讲消失，结构不匹配。“昨夜梦中”与“今朝空余”时空对比尚可但词性未完全对应，但就像用“红苹果”对“吃香蕉”。

1.3 押韵协调 (5/5)：严格押平水韵七阳部，所有结尾字（霜、乡、伤、裳、光）都押着唐代的标准韵脚。

### 2. 意象意境 (21/30)

2.1 古典运用 (14/20)：“樱花”意象在汉诗传统中较晚出现（始见晚唐），与传统咏春意象稍显隔阂，建议替换为“桃李”等经典意象。

2.2 意境层次 (7/10)：构建了冬春时空转换的对比结构“冬天盼春→春天来了却失落→梦中花开→现实流泪”，但结尾“美景光”略显直白，就像四格漫画最后忽然直接加文字解说，削弱意境深度，少了韵味。

### 3. 主题思想 (15/20)

3.1 情感真挚 (10/12)：思春情怀真挚可感，符合古典伤春传统

3.2 思想传承 (5/8)：未突破传统伤春模式，可强化个体生命体验维度

### 4. 语言锤炼 (10/15)

4.1 凝练度 (5/8)：“心驰神往”稍显冗余，而且像现代广告词，建议化用谢朓“驰晖不可接”典故

4.2 典雅度 (5/7)：“美景光”属现代汉语组合，建议改用“锦绣光”、“旧年

妆”等传统语汇

### 5. 创新性 (7/10)

时空折叠技法值得肯定（冬-春-梦-现实的转换），像电影蒙太奇，但意象系统创新不足

总分：70/100

[改进建议]

#### 1. 格律规范

1.1 平仄音韵：原诗存在两处关键格律问题

三平尾问题：“春又至”三字全为平声（普通话中“春”为阴平，“又”为去声但古音为平声，“至”为去声但在此处需按韵书判定），形成“平平仄”的失衡结构。可调整为“春复至”（平仄仄），既保持语义又修正平仄。

孤平现象：“今朝空余泪沾裳”中“空余”二字为双平声，导致前后平仄交替断裂。参考杜甫“星垂平野阔”的句式，改为“晓镜愁云泪染裳”（仄仄平平仄仄平），通过增加仄声字恢复平仄平衡。

1.2 对仗工稳：颔联与颈联需强化对仗逻辑

意象对仗：原句“花开花落”对“不见花影”，前者为自然循环，后者为否定性观察，建议改为“花谢花开春又至，鸿来鸿往影成伤”，使植物生长与动物迁徙形成生命律动对照。

词性对应：“昨夜梦中”（时间状语+地点）与“今朝空余”（时间+状态）结构失衡。借鉴李商隐“晓镜但愁云鬓改，夜吟应觉月光寒”的工对模式，改为“昨梦芳菲盈翠袖，晓妆零落黯罗裳”，使“昨梦”对“晓妆”（时间），“芳菲”对“零落”（状态），“翠袖”对“罗裳”（服饰）。

1.3 押韵协调：无

#### 2. 意象意境

2.1 古典运用：意象选择需更贴近传统

“美景光”过于直白，缺乏古典韵味。可改为“旧年妆”，用“妆”字暗喻春天的盛景，同时隐含“物是人非”的伤感，使意象更具层次感。

意象系统重构：建立“冬-春”对照意象链，将“寒风吹雪霜”强化为“朔气凝云结素霜”，通过“凝云”“素霜”等冷色调词汇，与后文“芳菲”“翠袖”形成色彩冲击。

2.2 意境层次：需深化情感表达

结尾“再现人间美景光”过于直白，削弱了意境的深度。可考虑引入屈原典故，将“唯愿春风吹花醒”改为“东风若解灵均意”，用“灵均”（屈原的字）暗示文化传承的期盼，同时将“美景光”改为“重染离骚草木香”，使意境从个人感伤升华为对文化传统的呼唤。

#### 3. 主题思想

3.1 情感真挚：需增强情感层次

情感表达稍显单一，停留在“盼春→失望”的层面。在尾联引入屈原典故，将个人情感与文化传统结合，使情感从“个人感伤”升华为“文化乡愁”，增加情感的厚重感。

3.2 思想传承：需强化文化内涵

未突破传统伤春模式，缺乏思想深度。可通过“重染离骚草木香”的意象，将主题从“盼春”拓展到“文化传承”，使诗歌不仅表达个人情感，还隐含对传统文化复兴的期盼。

#### 4. 语言锤炼

## 4.1 凝练度：需精简冗余表达

“心驰神往”稍显冗余，且现代感较强。可改为“神驰楚泽”，既简洁又有古典韵味，同时“楚泽”暗指楚地，为后文引入屈原典故做铺垫。

## 4.2 典雅度：需调整语汇搭配

“美景光”语汇搭配生硬，缺乏古典美感。可改为“锦绣光”或“旧年妆”，使语言更贴近古诗的表达习惯。

## 5. 创新性

创新主要体现在时空跳跃的设计上，但意象系统创新不足。可保留时空跳跃的结构，但在意象选择上融入更多传统文化元素（如“武陵乡”“离骚草木香”），使创新建立在传统基础上，避免突兀感。

图 3.11 输入输出示例 1（古诗评分）

月满中秋乐无边，团圆祝福寄诗篇。千里遥寄相思意，心系家人思情连。举杯共饮桂花酒，美味佳肴心中甜。愿你家庭幸福美，快乐永远伴随身。

## [评分]

## 1. 格律规范 (6/25)

1.1 平仄音韵 (3/10)：首联“乐无边”三平尾（乐古音为入声，但按普通话读为平声），颈联“桂花酒”三平调，全诗平仄交替混乱如“平平平仄仄平平”连续出现，失律严重

1.2 对仗工稳 (2/10)：颔联“千里遥寄”对“心系家人”主谓结构失衡，颈联“举杯共饮”与“美味佳肴”动词短语对名词短语，如用刀叉吃汉堡般不协调

1.3 押韵协调 (0/5)：严格押平水韵先韵（边、篇），但“甜”属盐韵、“身”属真韵，跨三部押韵实为严重出韵，此处应判 0 分

## 2. 意象意境 (15/30)

2.1 古典运用 (10/20)：“桂花酒”属中秋经典意象，但“幸福美”等现代语汇如彩色气球飘在古画上，破坏整体性

2.2 意境层次 (5/10)：情感停留在祝福表层，如月饼包装纸般单薄，缺少“海上生明月”的时空纵深

## 3. 主题思想 (12/20)

3.1 情感真挚 (9/12)：团圆之情真挚如满月，但“愿你”等直白表述似贺卡用语

3.2 思想传承 (3/8)：未触及“千里共婵娟”的哲学维度，停留在物质层面描写

## 4. 语言锤炼 (6/15)

4.1 凝练度 (3/8)：“美味佳肴心中甜”如菜谱说明，可精简为“珍馐和露甘”

4.2 典雅度 (3/7)：“快乐永远伴随身”杂糅古今语体，如汉服配运动鞋

## 5. 创新性 (2/10)

5.1 守正出新 (2/10)：整体属节日应景之作，如流水线月饼缺乏新意  
总分：40/100

[改进建议]

1. 格律规范

1.1 平仄音韵：重构全诗平仄框架

首联可改为“冰轮初满界三千（平平平仄仄平平）”，既符合平起平收律，又以佛教“三千世界”典故提升意境

颈联调整平仄：“捣药蟾光浮玉盏（仄仄平平平仄仄），研云桂影落雕盘（平平仄仄仄平平）”

1.2 对仗工稳：重建精微对仗

原颔联改为“素娥应悔偷灵药（仄平平仄平平仄），玄兔犹能捣寿丹（平仄平平仄仄平）”，用李商隐《嫦娥》典故形成神仙对仗

原颈联重构为“星垂碧落转金饼（平平仄仄仄平仄），潮涌钱塘漱玉盘（平仄平平仄仄平）”，天文意象与地理奇观相对

1.3 押韵协调：统一押删韵

将韵脚调整为“寰、娴、斑、滑、鬟”，既符合平水韵又增加典雅度

2. 意象意境

2.1 古典运用：植入文化符号

将“美味佳肴”升华为“雕胡饭”，用谢灵运“雕胡方接饭”典故；“桂花酒”深化为“吴刚斫桂”神话意象

2.2 意境层次：构筑三重时空

现实团圆 → 月宫遥想 → 历史回响，参考张若虚《春江花月夜》的时空结构，在尾联加入“今夜清光似旧年（照见开元全盛时）”的历史维度

3. 主题思想

3.1 情感真挚：转换抒情视角

改直白祝福为“不知秋思落谁家”的含蓄表达，尾联可化用王建“今夜月明人尽望”的集体情感

3.2 思想传承：注入哲思元素

在颈联加入“圆缺自有洪荒力”的宇宙观，呼应苏轼“人有悲欢离合”的辩证思维

4. 语言锤炼

4.1 凝练度：炼字示范

“举杯共饮”精炼为“浮白”（《说苑》典故），“快乐永远”转化为“羲和驭”，用太阳神比喻永恒

4.2 典雅度：语汇升级

“家庭幸福美”改为“阖家长乐好”，“伴随身”提升为“沐晞发”，借用《楚辞》“沐咸池兮晞发阳”的意象

5. 创新性

神话新诠：将月宫传说与当代航天结合，创造“玉兔车痕印广寒”等新意象，既守正又出新

图 3.12 输入输出示例 2（古诗评分）

于是，基于新设计的评分体系和两个输入输出案例，整合好 Few-shot 提示词后，系统对图 3.5 中生成的古诗给出评分与改进意见。其中评分结果和对应雷达图见图 3.13 和图 3.14，系统给出的针对性改进建议见图 3.15。

1. 格律规范 (21/25)
- 1.1 平仄音韵 (8/10): “琼英绽”三平尾 (琼-平、英-平、绽-仄，按平水韵“绽”属仄韵为仄，实为三平调；“忍看落英飘”存在孤平现象)
  - 1.2 对仗工稳 (8/10): “琉璃映日”对“玉砌连云”器物建筑对仗精妙，但“曲径曾携”与“长亭忍看”动词结构稍欠工整
  - 1.3 押韵协调 (5/5): 全诗押平水韵二萧部 (摇、桥、霄、飘、潮、遥、娇) 如编钟贯珠
2. 意象意境 (27/30)
- 2.1 古典运用 (18/20): “晴川鹤影”化用崔颢典而不露，将“珞珈山”地域特征融入古典语境如盐入水
  - 2.2 意境层次 (9/10): 从实景 (琼英绽) 到追忆 (书卷过) 再到时空穿越 (鹤影娇)，构建三重意境如敦煌飞天飘带
3. 主题思想 (17/20)
- 3.1 情感真挚 (11/12): “青衿空对暮山遥”将求学记忆与离别惆怅交织，如吴带当风
  - 3.2 思想传承 (6/8): “东君不解”暗合《楚辞》司春之神原型，但未突破传统伤春范式
4. 语言锤炼 (13/15)
- 4.1 凝练度 (7/8): “万点飞花迷画槛”数字量词精准如界画，唯“忍看”稍显直露
  - 4.2 典雅度 (6/7): “粉泪”承李煜“胭脂泪”意象，“碧霄”接刘禹锡“晴空一鹤”语境
5. 创新性 (8/10)
- 5.1 守正出新 (8/10): “珞珈山”地理符号与“金阙”仙家意象融合，如唐三彩吸收胡风
- 总分：86/100

图 3.13 古诗评分（示例）

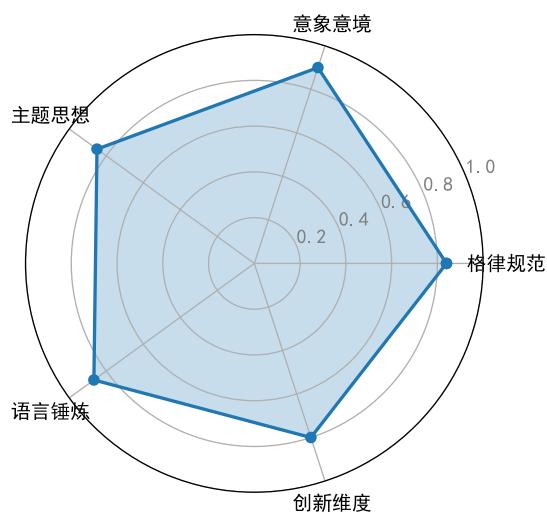


图 3.14 评分雷达图（示例）

|  |
|--|
| <b>1. 格律规范</b>   |
| <b>1.1 平仄音韵：</b>   |
| 首联“琼英绽”改“琼瑶碎”(平平仄)，既避三平调又暗含陆机《文赋》“玉卮无当，虽宝非用；琼瑶满案，终难为佩”的喻象                |
| “忍看落英飘”调整为“忍顾乱红凋”(仄仄仄平平)，参考秦观“忍顾鹊桥归路”句式                                  |
| <b>2. 意象意境</b>   |
| <b>2.1 古典运用：</b>   |
| “晴川鹤影”可深化为“晴川历历汉阳树”的今昔对照，加入“残碑”意象与“鹤影”形成文明传承的张力                          |
| “琉璃映日”建议改为“琉璃承露”，接续李贺“琉璃钟，琥珀浓”的瑰丽想象，增强晨昏时序感                              |
| <b>3. 主题思想</b>   |
| <b>3.2 思想传承：</b>   |
| 尾联注入《论语》“浴乎沂，风乎舞雩”的哲思，将“芳菲事”升华为“曾点之志”，改“犹记晴川鹤影娇”为“独对春沂咏雩娇”，使离愁转为文化坚守     |
| <b>4. 语言锤炼</b>   |
| <b>4.1 凝练度：</b>  |
| “朱阁檐前”精炼为“朱阑十二”，化用李白“解释春风无限恨，沉香亭北倚阑干”的典故密度                               |
| “香雪覆虹桥”调整为“香阵没星轺”，借“星轺”(使者车驾)暗示人生旅途，增强叙事性                                |
| <b>5. 创新性</b>  |
| 在“残红逐晚潮”处植入现代意象：“无人机影掠江皋”，形成古典送别(兰舟)与当代离别(无人机)的蒙太奇，呼应李商隐“却话巴山夜雨时”的时空折叠技法 |

图 3.15 优化建议（示例）

### 3.4.2 自动度量方法

除了大模型自身的评分外，系统还将结合古诗生产领域内常用的自动度量方法（BLEU、ROUGE、Similarity、Distinct）来进行指标计算，以辅助古诗的质量评估。

对 BLEU 和 ROUGE，二者均基于参考文本的  $n$ -gram 比较，由于古诗用词精炼、篇幅较短， $n > 2$  的计算参考价值并不大，因此本文选取  $n = 1$  和  $n = 2$  的情况进行计算，即使用单字和双字的基本单元来考察生成文本与参考文本的相似度。至于参考文本，由于古诗生成并不像机器翻译那样有明确的参考文本，过往研究往往是基于庞大的古人作品集来计算指标均值，以此来反映系统作品与真实古诗的平均相似度，这意味着古人作品集应足够大，以尽可能覆盖古诗创

作的多样性，并且质量应有保证，不能是名不见经传、甚至随意拼凑的作品。由此，本文选取 THU Chinese Classical Poetry Corpus (THU-CCPC-v1.0) [42] 作为参考文本集。该数据集包含从明朝到清朝的几乎所有绝句作品，训练集、验证集与测试集总计 127682 首诗，标点符号全部替代为统一的分隔符“|”，数据格式良好，可省去清洗环节直接使用。

在计算时，分别使用nltk<sup>[43]</sup>和rouge\_chinese<sup>[44]</sup>两个软件包来实现指标计算。nltk (natural language toolkit) 是一个自然语言处理工具包，集成有丰富的文本处理功能，包括分词、词性标注、命名实体识别等等，其中nltk.translate包中包含有支持 BLEU 指标计算的函数sentence\_bleu()和corpus\_bleu()两个函数，分别用于单个和多个候选文本的指标计算。rouge\_chinese在原始的rouge包<sup>[25]</sup> 的基础上对中文 NLP 任务做了改进，使其支持基于中文常用标点符号的分句，并优化了对最长公共子序列计算的内存占用，使其无需冗长的递归计算就能得到结果，也因此可以直接计算最原始的 ROUGE 指标，而不是像rouge包那样通过分句计算来近似。在使用时，BLEU 默认以字符为单位计算，因此只需要处理为字符序列；但 ROUGE 需要手动将文本中的词以空格分隔。对古诗而言，可直接以单个汉字为单元划分，而不必进行更灵活的分词。

对 Distinct，其同样选取  $n = 1$  和  $n = 2$  的情况来计算，将古诗文本分别分割为一系列 1-gram 和 2-gram 的列表，并转化为集合，将集合与列表的长度之比来作为 Distinct- $n$  指标的值。

对 Similarity，过去的研究往往会固定生成古诗的体裁，如四句绝句，因此可固定计算头两句相似度Sim12、后两句相似度Sim34、以及二者的相似度Sim2L三种分数来衡量古诗内前后句子的语义相关性。但对于未明确指定输出长度的古诗生成系统，需要设计新的计算方式，以同时适用不同体裁的作品。

在不同的古诗体裁中，词并不像与律诗、绝句那样，通篇遵循完全一样的字数结构，其最初起源歌曲中对曲调的适应，在演变中逐渐形成了各具特色的词牌，每个词牌都有自己独特的格律规范，诸如“清平乐”、“长恨歌”、“西江月”。而在众多词牌中，并非所有词牌都是“上下阙”的双调，而是单调、双调、三叠、四叠都有，因而也无法固定以“阙”为单位来计算前后相似度。诚然，词按照自身的分阙形式不同，也能够分别遵循一种方式来计算相似度分数，但计算的结果无法

与不同的分阙词牌比较，因而适用性较低，本文不予考虑，而是依旧关注绝句、律诗、排律等格律诗。由于这些体裁都严格遵循或五言或七言的字数限制，可以沿用前人工作的思路，设计通用的指标来反映古诗中前后文的语义相关性。

由此，本文设计了更通用的指标 $\text{Sim}_{\text{intra}}$ 和 $\text{Sim}_{\text{inter}}$ ，分别表示古诗中前后句子间的相似度和不同联之间的相似度。具体而言，对古诗中的句子两两分组， $\text{Sim}_{\text{intra}}$ 计算组内两个句子的相似度并取平均（如第一句和第二句、第三句和第四句）， $\text{Sim}_{\text{inter}}$ 计算组间的相似度再取平均（如第一二句和三四句、三四句和五六句）。

这一方法的核心在于利用词向量模型在预训练阶段学习到的句子间的语义关系，因而指标分数十分依赖词向量模型的质量。古诗这一特殊的文本类型要求词向量模型必须在古汉语文本上进行预训练，而不是简单地选取现代汉语词向量模型（如 Bert-base-chinese）；此外，词向量的训练数据集也应考虑古诗独特的格律形式特征，而不是简单地选取文言作品。由此，本文采用清华大学自然语言研究中心开源的 Bert-CCPoem 词向量模型<sup>[45]</sup>，其以汉字为基本单位训练 Bert 模型，训练集 CCPC-Full v1.0 包含几乎所有中国传统诗词，涵盖 926024 首古诗的 8933162 个句子，因此相比于其他基于现代汉语或是泛古文数据集构建的词向量，更契合古诗句的应用场景。

### 3.5 古诗优化

为有效改进生成的古诗，系统基于之前分析得到的改进意见，同时参考原古诗的评分和用户的文本图像输入，对古诗进行进一步的迭代润色，基于评分体系中的薄弱部分针对性地提高分数，同时避免偏离用户意图。

对图 3.5 中的古诗，结合文本输入（图 3.2）、图像描述（见 3.2）、评分和优化建议（图 3.13 和图 3.15），系统输出优化后古诗和相应的优化说明（见图 3.16 和图 3.17）

珞珈山下琼瑶碎，朱阑十二翠云飘。  
 万点飞花迷玉砌，千重香阵没星轺。  
 瑰璃承露浮金阙，玉砌连云接碧霄。  
 曲径曾携黄卷过，长亭忍顾乱红凋。  
 东君未解青衿恨，犹遣残霞逐晚潮。  
 粉泪暗随春水逝，素襟空对暮山遥。  
 残碑鹤影参商渡，机影江声日夜迢。  
 欲问蓬瀛何处是，春沂独咏舞雩娇。

图 3.16 优化后古诗（示例）

1. 平仄调整：“琼瑶碎”（平平仄）化解原三平尾，暗引陆机“琼瑶”喻指未琢之玉，喻学子未竟之志；“忍顾乱红凋”（仄仄仄平平）借易安“忍顾”词眼，使孤平转双拗救
2. 意象重构：“朱阑十二”化用李白沉香亭典故，以数字阑干喻求学四载；“香阵没星轺”借骆宾王“香阵冲天”与谢灵运“星轺阻辙”，喻青春壮志与宦途阻隔的冲突
3. 时空折叠：新增颈联“残碑鹤影参商渡，机影江声日夜迢”，以崔颢“晴川历历”对应珞珈残碑，无人机影接续李商隐“星桥铁锁”意象，构建古今送别对话场域
4. 哲思升华：尾联化《论语》“风乎舞雩”为“春沂独咏”，将樱花意象升华为“浴沂归咏”的文化坚守，使离愁转为对“曾点之志”的传承
5. 器物对仗：调整“瑰璃承露”对“玉砌连云”，既保持建筑意象，又以“承露”接续李贺琉璃钟意象，暗喻学子承接文化甘露
6. 色彩体系：保留原诗“翠、金、碧、粉、青”色谱，新增“素襟”与“残霞”形成冷暖对照，暗合图片中琉璃瓦的冷绿与樱花的暖粉

图 3.17 优化说明（示例）

在先前的工作中，古诗优化的输入只包含待改进的古诗和修改意见，这会导致两个问题——其一，根据提示词的要求，修改意见只覆盖得分较低的维度，并不像评分体系那样全面。模型在优化古诗时，有可能只会关注这些薄弱部分，而忽略了其他未被提及的维度，导致优化后的古诗在这些方面的分数下降，顾此而失彼。其二，在古诗优化时，模型只考虑了古诗的内容和结构，却没有考虑用户的原始输入，因此优化过程也可能会偏离用户的原始意图，导致生成的古诗与用户的期望相去甚远。这样一来，古诗的优化将仅针对“诗”这一艺术作品的优化，而忽略用户最初的需求，这一问题将随着优化迭代的增多而愈发明显。

为了提高古诗优化的效果，同时不偏离用户的原始需求，本文在先前提

示词的基础上，增加了先前对原古诗的评分evaluation、用户的输入（文本输入user\_text与图像输入描述description），在确保古诗优化有效性的同时，保留对用户需求的考量。（提示词见图 3.18）

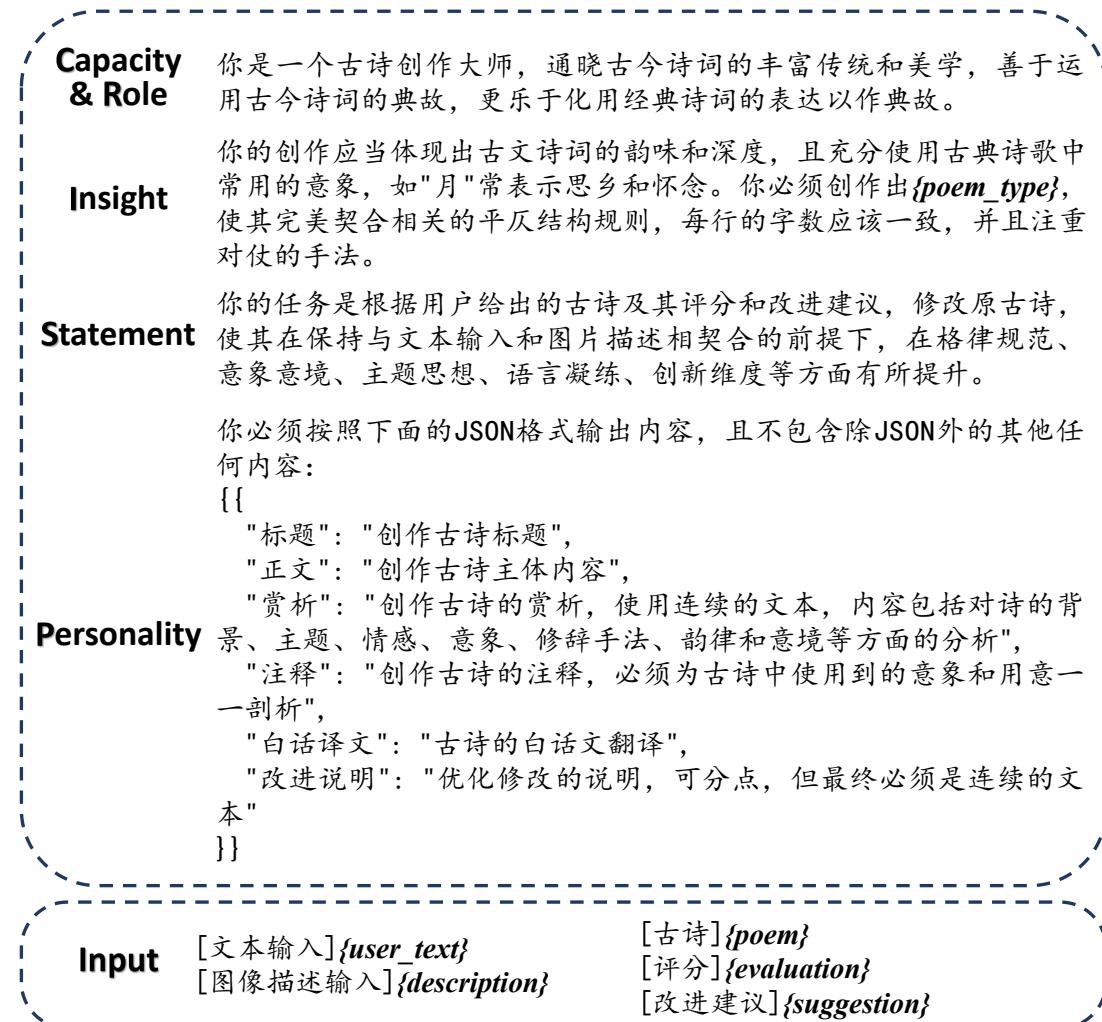


图 3.18 提示词（古诗优化）

作为对比，若在优化时只提供原来的古诗（如图 3.5）和修改建议（如图 3.15），系统给出的优化古诗如图 3.19 所示。

为对比两种优化结果的质量好坏，使用系统的评分功能对两种优化结果进行评分，再与原诗的评分进行对比，绘制出雷达图，如图 3.20 所示（为便于观察，将评分雷达图的绘制范围改为  $[0.5, 1]$ ）。

可见，仅输入原诗和改进意见时，生成的优化古诗在“语言锤炼”和“格律规范”上的得分都有所提升，得益于改进意见的指导，却在“意象意境”和“主体思想”

珞珈山下琼瑶碎，朱阑十二翠云招。  
 万点飞花迷画舫，千重香阵没星轺。  
 瑰丽承露浮金阙，玉砌连霄接绛绡。  
 曲径曾携黄卷过，长亭忍顾乱红凋。  
 残碑犹刻汉阳树，鹤影空临楚泽苕。  
 东君未解离人恨，却遣春醒逐暮飙。  
 粉泪暗随青翰逝，玄裳独对紫烟缭。  
 无人机影掠江皋，兰棹声沉浦溆遥。  
 若问当年芳菲志，曾点春沂咏雩娇。

图 3.19 优化后古诗（仅输入原诗与优化建议）

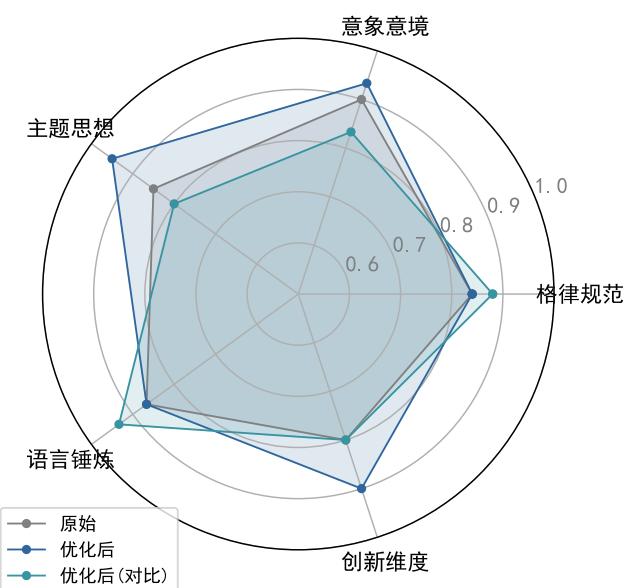


图 3.20 对比雷达图（经放缩）

两个维度上出现明显的下降，最后总分甚至降低一份 ( $86 \rightarrow 85$ )。相比之下，结合用户输入、图像描述和评分作为完整输入时，优化结果并未出现分数降低的情况，实现了古诗整体质量的提升 ( $86 \rightarrow 90$ )。

进一步分析发现，对仅输入原诗和改进意见时生成的优化古诗，系统给出的评分中出现了其他子维度分数下降的情况。具体而言，基于原先的修改建议（图 3.15），古诗在“平仄音韵”上的表现提高 ( $8 \rightarrow 9$ )，促使“格律规范”维度的分数提升 ( $21 \rightarrow 22$ )；又因将“朱阁檐前”精炼为“朱阑十二”，在“凝练度”上提高一分 ( $7 \rightarrow 8$ )，使得“语言凝练”维度分数提高 ( $13 \rightarrow 14$ )。然而，对“意象意境”维度 ( $27 \rightarrow 25$ )，改进建议并未提及的“意境层次”因“无人机”这一现代意象的加入

而出现了分数下降（9 → 7），哪怕提及到的“古典运用”仍然保持 18 分。对“主体思想”维度（17 → 16），因“无人机”意象未能完成哲学升华而在“思想传承”出现退步（6 → 5）。

此外，这一古诗与用户输入内容的契合度也不如基于完整输入的优化古诗。以用户输入文本为例，图 3.2 中的文本核心为“樱花美景 + 毕业离别 + 对珞珈山校园的怀念”，在完整输入的优化古诗中，“东君未解青衿恨”一句使用“青衿”直接点明学生身份，“恨”强化离愁，精准贴合用户作为毕业生的情感；“长亭忍顾乱红凋”一句则以“长亭”（送别意象）与“乱红凋”（樱花凋零）双重隐喻，强化时间流逝与离别感伤。相比之下，仅输入原诗和改进意见的优化古诗中，虽用“离人恨”表达离别，但缺乏身份特指，情感指向不够精确。

再以图像描述为例，3.2 中图像描述的核心为“樱花繁茂 + 传统建筑 + 宁静诗意”，在完整输入的优化古诗中，“珞珈山下琼瑶碎”（樱花如碎玉），“万点飞花迷玉砌”（飞花与建筑交融）精准还原樱花与建筑的画面层次；“琉璃承露浮金阙”（琉璃瓦）与图像中“绿色琉璃瓦”高度一致，细节还原到位。相比之下，仅输入原诗和改进意见的优化古诗中，“万点飞花迷画舫”引入“画舫”（游船）意象，但图像中仅提及“游客在阳台”，存在场景偏差；而“无人机影掠江皋”：引入现代科技意象，与用户描述的传统建筑、自然景观氛围冲突，易造成情感断层。

总体而言，完整输入的优化古诗通过“青衿恨”的身份特指、樱花与建筑交融的精准意象，以及从美景到离愁的自然递进，更完整地还原了用户文本与图像中的情感内核。而仅输入原诗和改进意见的优化古诗存在个别偏离性意象（如无人机、画舫）和泛化的离别表达，使其在情感共鸣与场景一致性上稍逊一筹。

### 3.6 本章小结

本章主要介绍了本系统的功能设计，并详细阐述了各模块的设计考量。系统采用模块化设计，包含图像分析、古诗生成、古诗评价和古诗优化四个核心模块，通过百度智能云 API 接口调用各类大模型，构建了一个完整的古诗创作工作流。

在图像分析模块，本文采用中文语境下的 DeepSeek-VL2 模型替代原有的多模型组合方案，解决了 CLIP 和 MiniGPT-4 在文化符号识别和情感表达方面的不

足。该模型能够生成包含精确色彩描述和立体空间信息的中文图像描述，为后续古诗创作提供了丰富的文化意象素材。

古诗生成模块通过精心设计的 CRISPE 框架提示词，实现了格式规范的 JSON 输出，除古诗原文外还包含赏析、注释与白话文翻译，充分提高模型生成的结果可解释性。经过多模型对比测试，最终选用在典故意象运用方面表现突出的 DeepSeek-R1 模型，其生成的古诗不仅符合格律要求，更能巧妙融入传统文化元素，展现出深厚的文化底蕴。

在古诗评价方面，本文重构了原有的评分体系，解决了标准模糊、维度重叠等问题。新的五维评分体系具有明确的量化标准和典型诗例说明，配合 Few-shot 提示框架，使模型评分更加客观准确。同时引入 BLEU、ROUGE 等自动度量方法作为辅助评估手段，构建了全面的古诗质量评价体系。

古诗优化模块通过扩展输入信息（包括原诗、评分、用户输入等），在保持用户意图的前提下对古诗进行针对性改进，在提高优化效果的同时，保证与用户原始需求的对齐。

## 4 实验及结果分析

为验证本系统在古诗生成、评价和优化方面的有效性，本文设计开展了相关实验，利用已有古诗数据集来检验系统性能，结合自动度量方法对比了ERNIE-4.0 和 DeepSeek-R1 两种模型产出的结果。

### 4.1 基于白话文的古诗生成实验

参考相关工作中评估模型表现的方法，给定一个质量获得认可的古诗，将评估模型的输出古诗与这一古诗进行对比，利用 BLEU 和 ROUGE 计算指标分数<sup>[30]</sup>。在这一思路下，原古诗的质量应当足够好，以确保指标分数能够反映被评估输出的质量；且模型的输入要与原古诗的内容十分相关，以确保指标分数能够反映模型的生成能力。

为此，本文选择《唐诗三百首》中的名篇作为古诗原文，选择相应的白话文翻译作为文本输入。《唐诗三百首》是清代蘅塘退士孙洙编选的唐诗选集，问世不久便闻名遐迩，成为唐诗入门读物的首选，其中收录的唐诗均为名篇佳作。原书中除了律诗、绝句外，还收录有乐府、古体诗等形式多变的体裁，故不便于评估，本文仅选择律诗和绝句两种体裁的古诗进行测试，各体裁作品及数目见表 4.1。此外，相应的白话文翻译源自古诗文网<sup>[46]</sup>。

表 4.1 《唐诗三百首》测试数据集

| 体裁   | 数量  |
|------|-----|
| 七言律诗 | 51  |
| 七言绝句 | 50  |
| 五言律诗 | 80  |
| 五言绝句 | 29  |
| 合计   | 210 |

测试发现，ERNIE-4.0 和 DeepSeek-R1 两种模型的输出均与参考古诗有极高的覆盖度，因而在 BLEU 和 ROUGE 指标上均有很高的得分，明显区别于以往工作中的结果（如BLEU-1=0.168, BLEU-2=0.002<sup>[30]</sup>），属于异常情况。作为大模型，两种模型的训练数据均包含了大量的古诗数据，尤其是对《唐诗三百首》这样的名篇，因而本实验并不能做到训练集与测试集的独立，并不具备验证效果。测试统计结果见表 4.2 和表 4.3。

表 4.2 白话文古诗生成实验结果 (DeepSeek-R1)

|      | BLEU            |                 | ROUGE           |                 |                 |
|------|-----------------|-----------------|-----------------|-----------------|-----------------|
|      | BL-1            | BL-2            | R-1             | R-2             | R-L             |
| 七言律诗 | 0.583599        | <b>0.432841</b> | <b>0.557353</b> | <b>0.349461</b> | 0.527267        |
| 七言绝句 | 0.559833        | 0.415325        | 0.562665        | 0.344502        | 0.540417        |
| 五言律诗 | <b>0.597726</b> | 0.432494        | 0.549401        | 0.337500        | <b>0.545573</b> |
| 五言绝句 | 0.523605        | 0.332445        | 0.495130        | 0.245845        | 0.471983        |
| 平均   | 0.575037        | 0.414674        | 0.546996        | 0.329415        | 0.529737        |

表 4.3 白话文古诗生成实验结果 (ERNIE-4.0)

|      | BLEU            |                 | ROUGE           |                 |                 |
|------|-----------------|-----------------|-----------------|-----------------|-----------------|
|      | BL-1            | BL-2            | R-1             | R-2             | R-L             |
| 七言律诗 | 0.765951        | 0.671162        | 0.770484        | 0.609392        | 0.743464        |
| 七言绝句 | 0.798136        | 0.702556        | 0.808346        | 0.649742        | 0.764375        |
| 五言律诗 | 0.691980        | 0.533006        | 0.670557        | 0.445450        | 0.646720        |
| 五言绝句 | <b>0.839312</b> | <b>0.764698</b> | <b>0.834125</b> | <b>0.726052</b> | <b>0.812500</b> |
| 平均   | 0.767420        | 0.658670        | 0.765072        | 0.596372        | 0.734993        |

## 4.2 评分功能实验

为了检验系统评分功能的有效性，本文尝试收集具有质量差异的古诗分组，通过先验的质量分层来验证系统评分的合理性。此外，也将使用 BLEU 等自动度量方法，检验这些自动度量方法的有效性。

选择第六届“诗词中国”传统诗词创作大赛<sup>[47]</sup>的公开获奖作品为测试集，测试系统评分功能的有效性和可信度。该大赛的评审专家均为古诗词领域的专家，且其评分标准公开透明，因而可以作为测试系统评分功能的参考。该比赛设有不同的作品体裁和比赛组别，而本文主要专注于格律要求较严格的律诗、绝句等体裁，因此为便于测试，只选取其中的律诗与绝句作品进行。

测试发现，不同奖项组之间的评分存在差异但并不显著，而且在许多维度上优秀奖的得分最高，考虑是因为各奖项样本数量分布不均，噪声干扰较大，结论难有普适性。因此，将一、二、三等奖的作品进行合并，形成数目为 28 的样本组“一二三等奖”，再将其与优秀奖的平均得分进行比较，发现尽管差异仍然不算显著，但在各个维度上的比较结果都与真实的结果一致，即“一二三等奖”中古诗的质量要好于“优秀奖”（见表 4.4）。

在自动度量方面，BLEU 与 ROUGE 的比较结果与系统评分的结果类似，在

表 4.4 获奖作品的评分实验（系统评分）

|     | 格律规范         | 意象意境         | 主题思想         | 语言锤炼         | 创新性          | 总分           | 样本  |
|-----|--------------|--------------|--------------|--------------|--------------|--------------|-----|
| 一等奖 | <b>0.800</b> | 0.833        | 0.825        | 0.800        | <b>0.850</b> | 0.820        | 2   |
| 二等奖 | 0.696        | 0.767        | 0.830        | 0.800        | 0.800        | 0.770        | 8   |
| 三等奖 | 0.793        | <b>0.836</b> | <b>0.846</b> | <b>0.828</b> | 0.825        | <b>0.825</b> | 18  |
| 一二三 | <b>0.768</b> | <b>0.818</b> | <b>0.839</b> | <b>0.818</b> | <b>0.821</b> | <b>0.810</b> | 28  |
| 优秀奖 | 0.762        | 0.805        | 0.824        | 0.798        | 0.797        | 0.796        | 91  |
| 平均  | 0.763        | 0.808        | 0.828        | 0.802        | 0.802        | 0.800        | 119 |

“一二三等奖”内部的比较结果有波动，但对“一二三等奖”和“优秀奖”之间的比较结果一致且正确（见表 4.5），说明二者作为古诗生成领域使用最广泛的自动度量指标，具有一定的有效性。对 Similarity 和 Distinct，二者“一二三等奖”内部的比较结果一致且正确，而且 Similarity 在三种不同奖项之间的得分差异相对显著，但“一二三等奖”与“优秀奖”的比较结果却与正确结果相反，除了Distinct-1这一个指标，说明二者在古诗质量评估的有效性有待进一步验证（见表 4.6）

表 4.5 获奖作品的评分实验（BLEU+ROUGE）

|     | BLEU            |                 | ROUGE           |                 |                 |
|-----|-----------------|-----------------|-----------------|-----------------|-----------------|
|     | BL-1            | BL-2            | R-1             | R-2             | R-L             |
| 一等奖 | <b>0.968750</b> | 0.678427        | 0.096657        | 0.001543        | 0.134217        |
| 二等奖 | 0.953125        | <b>0.805106</b> | 0.108463        | <b>0.004658</b> | 0.142648        |
| 三等奖 | 0.938999        | 0.788950        | <b>0.119148</b> | 0.004140        | <b>0.149292</b> |
| 一二三 | <b>0.945848</b> | <b>0.781568</b> | <b>0.113968</b> | <b>0.004003</b> | <b>0.145957</b> |
| 优秀奖 | 0.934612        | 0.749390        | 0.107799        | 0.003346        | 0.143840        |
| 平均  | 0.937247        | 0.756938        | 0.109246        | 0.003500        | 0.144337        |

表 4.6 获奖作品的评分实验（Similarity+Distinct）

|     | Similarity      |                 | Distinct        |                 |
|-----|-----------------|-----------------|-----------------|-----------------|
|     | S-Intra         | S-Inter         | D-1             | D-2             |
| 一等奖 | <b>0.675846</b> | <b>0.722989</b> | <b>0.906250</b> | <b>1.000000</b> |
| 二等奖 | 0.660678        | 0.697867        | 0.890625        | <b>1.000000</b> |
| 三等奖 | 0.658698        | 0.688951        | 0.879656        | 0.998677        |
| 一二三 | 0.661024        | 0.694880        | <b>0.885342</b> | 0.999165        |
| 优秀奖 | <b>0.670973</b> | <b>0.696803</b> | 0.879247        | <b>0.999832</b> |
| 平均  | 0.668639        | 0.696352        | 0.880676        | 0.999675        |

为进一步检验评分功能及各自动度量方法的有效性，本文扩大测试古诗范

围。由于“诗词中国”传统诗词创作大赛的往届获奖作品并不公开，因而无法获取更多带有可信等级标签的古诗作品。由此，本文纳入经典唐诗和打油诗两类数据集，退而求其次，尝试用世人公认的名篇作品与和质量明显较差的打油诗作为测试数据集。其中，从古诗文网挑选 20 首打油诗（见本节末图 4.1），结合《唐诗三百首》中的共 98 首律诗（包括五言和七言），与来自比赛的获奖作品“一二三等奖”与“优秀奖”一同合并作为测试集。

测试发现，系统对四个作品集的评分具有稳定的差异，且每个维度上得分的大小关系与真实结果完全一致（见表 4.7）。值得一提的是，在使用了新的评分体系后，系统对打油诗的评分降低到了 0.48，显著低于之前工作的评分 0.67，而获奖作品与经典唐诗的得分也有显著差距，有力地证明了新评分体系的有效性。

表 4.7 纳入唐诗和打油诗的评分实验（系统评分）

|     | 格律规范         | 意象意境         | 主题思想         | 语言锤炼         | 创新性          | 总分           | 样本 |
|-----|--------------|--------------|--------------|--------------|--------------|--------------|----|
| 唐诗  | <b>0.932</b> | <b>0.934</b> | <b>0.931</b> | <b>0.930</b> | <b>0.853</b> | <b>0.924</b> | 98 |
| 一二三 | 0.768        | 0.818        | 0.839        | 0.818        | 0.821        | 0.810        | 28 |
| 优秀奖 | 0.762        | 0.805        | 0.824        | 0.798        | 0.797        | 0.796        | 91 |
| 打油诗 | 0.328        | 0.490        | 0.615        | 0.470        | 0.585        | 0.481        | 20 |

在自动度量方面，BLEU 与 ROUGE 均能较好地反映古诗质量，除 BLEU-1 外的其他指标均与真实比较结果保持一致，进一步验证了二者在古诗质量评估方面的有效性。对 BLEU-1，获奖作品的平均得分要高于经典唐诗，说明基于 1-gram 覆盖度的精度计算方法仍存在局限。（见表 4.8）。

表 4.8 纳入唐诗和打油诗的评分实验（BLEU+ROUGE）

|     | BLEU            |                 | ROUGE           |                 |                 |
|-----|-----------------|-----------------|-----------------|-----------------|-----------------|
|     | BL-1            | BL-2            | R-1             | R-2             | R-L             |
| 唐诗  | 0.905665        | <b>0.792163</b> | <b>0.125864</b> | <b>0.005379</b> | <b>0.157440</b> |
| 一二三 | <b>0.945848</b> | 0.781568        | 0.113968        | 0.004003        | 0.145957        |
| 优秀奖 | 0.934612        | 0.749390        | 0.107799        | 0.003346        | 0.143840        |
| 打油诗 | 0.893658        | 0.726545        | 0.095770        | 0.003127        | 0.137659        |

对 Similarity，其两个指标 S-Intra 和 S-Inter 的对比结果与真实的质量等级并不一致，其中后者的结果甚至完全相反（见表 4.9）。Similarity 使用词向量模型来考察古诗内前后文的语义相似度，分别计算古诗中前后句子间和不同联之间的相似度均值，体现的是古诗的语义连贯性，但这里的语义是词向量模型从数

表 4.9 纳入唐诗和打油诗的评分实验 (Similarity+Distinct)

|     | Similarity      |                 | Distinct        |                 |
|-----|-----------------|-----------------|-----------------|-----------------|
|     | S-Intra         | S-Inter         | D-1             | D-2             |
| 唐诗  | 0.678716        | 0.693137        | 0.854964        | 0.998430        |
| 一二三 | 0.661024        | 0.694880        | <b>0.885342</b> | 0.999165        |
| 优秀奖 | 0.670973        | 0.696803        | 0.879247        | <b>0.999832</b> |
| 打油诗 | <b>0.698042</b> | <b>0.738067</b> | 0.760565        | 0.972817        |

据集中学习到的，是对古诗的真实语义的近似，而非其自身。进一步测试发现，对打油诗“天气好热燥，花开正繁闹。人静自然凉，心宽无烦恼”和名篇《登鹳雀楼》“白日依山尽，黄河入海流。欲穷千里目，更上一层楼”，S-Intra的结果分别为 S-Intra = 0.671893 和 S-Intra = 0.727715，而 S-Inter 的结果分别为 S-Inter = 0.727715 和 S-Inter = 0.659923。这说明后者的联内上下句的语义相似度要高于前者（即“白日依山尽”和“黄河入海流”、“欲穷千里目”和“更上一层楼”），但联间的语义相似度要低于前者（即“白日依山尽，黄河入海流”和“欲穷千里目，更上一层楼”）。

考虑指标自身的设计，对古诗作品而言，联内上下句往往因“对仗”而在形式、语法上有相似之处，因而 S-Intra 能够反映一定的语义连贯性，尽管比较结果不显著，仍能作为评估古诗质量的参考。但对 S-Inter，联间的上下句往往在形式、语法上有差异，一般作品的不同联之间会切换意象或主题，或是表达不同的情感，因而语义差异较大，相反，若全诗通篇使用同一意象或主题，或是表达同一种情感，反而会因意象单薄而质量欠佳，所以这一指标的合理性欠佳。

此外，也需要考虑词向量模型的影响。在过往研究中使用的词向量，是在 Bert 模型基础上通过添加额外的向量得到的（额外的向量包含情感、声调、节奏、位置等）。这些词向量表示均通过重新训练得到，并且在计算 Similarity 时直接使用<sup>[28, 30]</sup>。因此，本文使用的词向量模型 Bert-CCPoem 可能并不适用于这类下游任务，仍有待进一步研究。

对 Distinct，D-1 和 D-2 分别表示古诗中 1-gram 和 2-gram 的多样性，一定程度上反映了古诗在用词方面的创新性。测试发现，打油诗的 D-1 和 D-2 均低于经典唐诗和获奖作品，但得分最高者却分别为“一二三等奖”和“优秀奖”，这说明现代人的古诗作品或许在用词上会追求多样性。但与此同时，这也说明 Distinct 的

计算方法并不能独立地反映古诗质量。

1. 盆中有山有水，心里无城无府。独行不怕天黑，苟活休论贫富。
2. 天气好热燥，花开正繁闹。人静自然凉，心宽无烦恼。
3. 一轮红日上青天，未到西河不罢川。今朝又见千年日，千年一叹又江山。
4. 呼风唤雨太平常，摘颗星星袋里藏。家住东方不老岛，海山仙国是吾乡。偶尔天上去逛逛，吓得日月不放光。一口喝尽银河水，惊倒王母喊玉皇。如来问我何神圣，我道人间打油郎。
5. 一杯二杯三四杯，五杯六杯七八杯。九杯十杯开开胃，百杯千杯入微微。喝到万杯方半醉，亿杯兆杯不是吹。饮尽天上人间酒，始觉眼前牛在飞。骑牛飞到月宫去，嫦娥姐姐来作陪。
6. 吾手写吾心，诗词随口吟。从不标牌体，管它古与新。
7. 嫦娥仙子爱出楼，惹得人间看不休。雪白兔儿天上遛，银色蛤蟆水里游。
8. 江上一笼统，井上黑窟窿。黄狗身上白，白狗身上肿。
9. 一浪一浪又一浪，浪浪撞在石头上；明知前浪折了腰，后浪还要跟着上。
10. 时光催人老，不比不知晓。少年在眼前，才觉白发早。
11. 君自遥远故乡来，却说故乡在眼前。来日村口茶飘香，何不饮后才向前。他日回到故乡去，可知茶味如从前？
12. 东也湖，西也湖，洪城上下古月胡；南长清，北长清，大街小巷胡长清。
13. 一言一行一约定，一生一世一爱情。一轮一回一醉梦，一分一秒一旧心。纷纷扬扬雨，思思念念泪。凄凄凉凉夜，甜甜酸酸味。
14. 海枯石烂荡耳边，沧海桑田共誓言。风风雨雨同路走，携手白头哪一站。
15. 风风雨雨同路走，偎偎依依到白头。坎坎坷坷沉浮过，是是非非笑泪流。
16. 冷冷清清人生路，凄凄凉凉世间情。坎坎坷坷浮萍事，风风雨雨吾独行。
17. 含苞欲放惹人怜，情柔似水满心间。等到深秋染枯色，谁拿真情暖容颜！
18. 滚滚东逝水，滴滴恋红尘。轰鸣咆哮过，哀怨似歌声。
19. 风雨飘摇一浮萍，随波逐流何时停？醉生梦死他乡客，天下谁人又识君！
20. 今生缘已尽，心死在红尘。阴阳两相隔，独留一孤坟。

图 4.1 打油诗数据集

### 4.3 文图结合的古诗生成实验

为验证图片模态的必要性，需要设置文图输入的消融对比实验。之前的工作表明，面对不同的输入模态组合，ERNIE-4.0 的输出古诗质量会有显著的差异。本文尝试对 DeepSeek-R1 进行相同的实验，即固定文本和图像输入，分别测试仅文字、仅图像、文字与图像三种模态输入，对比生产的古诗质量。对生古诗质量的评价依据系统自身的评分功能。

为此，使用 5 对图像和文本作为输入，如表 4.11（图源网络，文本输入由 DeepSeek-VL2 辅助生成）。对每一种样本，分别单独输入文本和图像，在三种不同的模态输入组合下分别进行 10 次生成，利用系统自身的评分功能来生成评分并取均值，最后对 5 种样本的得分取均值，以比较不同模态输入下生成古诗的质量差异，结果如表 4.10。

**表 4.10 文图古诗生成实验结果**

|    | 格律规范         | 意象意境         | 主题思想         | 语言锤炼         | 创新性          | 总分           |
|----|--------------|--------------|--------------|--------------|--------------|--------------|
| 文  | 0.862        | <b>0.901</b> | 0.883        | <b>0.871</b> | <b>0.850</b> | <b>0.878</b> |
| 图  | 0.862        | 0.899        | 0.878        | 0.869        | 0.832        | 0.875        |
| 图文 | <b>0.866</b> | 0.900        | <b>0.885</b> | 0.861        | 0.832        | 0.876        |

实验发现，文图模态输入对古诗生长质量的影响并不显著。在部分维度上文图输入的分数略高于单模态输入，其他维度上文本输入的分数要略高于图像输入。尽管如此，三种模态输入下各维度分数的差异极小，均在 0.01 以内，且在总分上也仅有 0.002 的差异。考虑到古诗生成的质量受多种因素影响，尤其是模型本身的生成能力和输入模态的多样性，这部分差异也属于正常波动。换言之，文图模态输入提供的信息未展现出提高古诗生成质量的作用。

图像输入的作用体现在其包含的视觉场景信息，能够作为用户文本输入的补充，帮助用户表达隐晦的场景情感，以提高用户的体验。但就古诗生成的质量而言，DeepSeek-R1 的能力足以生成高分数的古诗，因而无论输入如何，其总能输出质量较好的古诗，只是无法与用户的需求对齐。所以，从古诗质量的角度而言，文图跨模态的优势尚未显现，但对于用户而言，图片的输入仍然是有必要的，能够帮助用户更好地传达丰富的视觉场景信息，以至潜在的情感色彩。

表 4.11 文图古诗生成实验测试集

| 图片  | 文本   |
|---|--|
|    | 日出时分，山间草甸被晨光染成金色，仿佛时间静止。这一刻的宁静，让人忘却尘世喧嚣，只想与这美景共存。                |
|    | 落叶铺满的小径，每一步都踏着季节的尾声。阳光透过树梢，却照不进那片即将离去的绿意。感伤的是，这秋日的美景，终究抵不过时间的流逝。 |
|   | 落花有意，流水无情，这满地的花瓣，像是春天的泪痕。它们曾是枝头的骄傲，如今却只能随风飘散，让人不禁感叹生命的短暂和无常。     |
|  | 月挂枝头，清辉洒落，映照着这孤寂的冬枝。这轮明月，是否也在远方照亮了你的窗前？思念如潮，涌上心头，却只能遥寄于这无声的月色。   |
|  | 老树、旧屋、枯草，时光在这里仿佛凝固。这寂静的村落，承载着多少代人的记忆与故事。我在这里，寻找着那些被岁月遗忘的温馨。      |

#### 4.4 本章小结

本章通过一系列实验对系统的古诗生成和评价功能进行了全面验证。实验结果表明，系统在不同方面表现出了一定的优势和局限性。

在古诗生成实验中，基于《唐诗三百首》的白话文翻译输入，ERNIE-4.0 和 DeepSeek-R1 两种模型的输出在 BLEU 和 ROUGE 指标上均取得了较高的分数。然而，由于训练数据与测试数据存在重叠，导致实验结果未能有效验证模型的生成能力，无法真实反映模型在独立数据上的表现。

在评分功能实验中，系统对不同质量等级的古诗进行了评分测试。通过分析“诗词中国”传统诗词创作大赛的获奖作品以及经典唐诗和打油诗的评分结果，系统能够较为准确地识别出古诗的质量差异，并在多个维度上给出与真实结果一致的评分，且相较于之前的工作有更好的区分度。这证明了系统评分功能的有效性和可信度。

在自动度量方法的验证方面，BLEU 和 ROUGE 指标在古诗质量评估中表现出了一定的有效性，尤其是在多 gram 的覆盖度计算上，能够较好地反映古诗的质量差异。然而，BLEU-1 指标在某些情况下存在局限性，可能会因简单的一元组覆盖度而误判古诗质量。Similarity 指标在语义连贯性方面存在明显局限，其基于词向量模型的计算方法可能会因语法结构对称而误判质量，且本文使用的词向量模型 Bert-CCPoem 可能不适应这类下游任务，仍有待进一步研究。Distinct 指标虽然能够反映用词多样性，但并不能独立地评估古诗的整体质量。

在文图跨模态输入的作用方面，在三种不同的模态组合输入下生成的古诗并未展现出显著的质量差异，这表明 DeepSeek-R1 的生成能力足以应对不同模态的输入，且在古诗质量上并未受到图像输入的显著影响。然而，图像输入仍然能够帮助用户更好地传达丰富的视觉场景信息，从而提升用户体验。

## 5 结语

古诗作为中华优秀传统文化的重要组成部分，凝聚了丰富的审美表达、文化意象与哲理思想，既是语言的艺术，也是精神的载体。近年来，随着自然语言处理技术的飞跃发展，越来越多的研究开始尝试借助人工智能模型复现古典文学的创作过程，实现对古诗的自动生成。大语言模型的语义理解与生成能力，多模态模型对图像与语言的协同建模能力，为这一目标提供了全新的技术基础。然而，古诗独特的格律结构、典故运用与意象传达，使其生成任务不仅对模型能力提出更高要求，也对系统的设计方法和使用体验构成挑战。尤其是如何让用户清晰表达创作意图，并从多角度理解生成结果的文化内涵，是目前大多数古诗生成系统仍待解决的问题。

围绕上述背景，本文以“基于大模型的跨模态古诗创作”为核心任务，围绕“文图融合输入”、“结构化提示控制”、“自动化质量评估”与“可解释性优化”等四个方向开展系统设计与实验探索。本文的研究目标不仅是生成符合格律规范与艺术美感的古诗文本，更希望通过图文协同输入、结构化提示词与反馈机制，提升生成内容与用户预期之间的匹配度，同时增强系统输出的可解释性与可控性，为用户提供更为丰富的文化体验。具体而言，本文的研究工作主要体现在以下几个方面：

在图像输入的分析方面，本文引入图像与文本的双模态输入机制，采用在中文语境下训练的跨模态大模型 DeepSeek-VL2，对图像内容进行语义解析，提取物体特征、情感氛围与意境信息，生成具有文化联想价值的中文描述文本，解决了 CLIP 和 MiniGPT-4 在文化符号识别和情感表达方面的不足。该模块能够从图像中抽取对古诗创作具有指导意义的元素，为后续生成过程提供丰富、诗意的背景输入，缓解用户在文字输入中对意象表达的不确定性，有效增强了系统对场景感知与画面意境的表达能力。

在诗歌生成与表达控制方面，本文调用首个基于强化学习训练的大语言模型 DeepSeek-R1 进行古诗生成，相比于 ERNIE-4.0 等模型，其在意象运用方面展现出显著的优越性。通过参考提示工程中的 CRISPE 等结构化框架，设计包含背景信息、角色设定、输出风格等多个要素的复合型提示词，控制模型在生成过程中遵循既定的体裁（如七言排律）。同时，系统支持生成白话文翻译、典故注

释与艺术赏析文本，帮助用户从语言形式、历史语境与情感层面理解生成内容。这种结构化的输出不仅增强了古诗生成的文化表达深度，也提升了用户对系统的信任与接受程度。

在质量评估机制方面，本文针对传统生成任务中评价方式不透明、缺乏反馈的问题，构建了一个五维度的古诗评分体系，涵盖格律规范、意象意境、主题思想、语言锤炼与创新性五个方面。每个维度下设多档评分标准，并提供示例作为参考，结合结构化 Few-shot 提示设计，引导大模型进行量化评分与文字评语生成。同时，配合 BLEU、ROUGE、Distinct 与 Similarity 等自动指标，从语言多样性、语义一致性等角度补充质量分析。该机制不仅能够清晰识别生成文本的强项与薄弱点，也为后续的迭代优化提供方向指引。

在优化迭代功能方面，系统在每轮古诗生成后，结合评分结果与优化建议，支持用户启动下一轮自动化修改流程。系统在保留原始图文输入意图的基础上，引导模型围绕评分中的薄弱维度进行针对性提升。该优化模块不仅强化了系统的“生成—分析—反馈—优化”闭环，也体现出大模型在语言润色与语义重构上的实用价值，提升了用户参与度与内容质量稳定性。

最后，本文设计进行了一系列实验，对系统的功能进行了全面检验。通过白话文古诗生成实验来对比 DeepSeek-R1 与 ERNIE-4.0 两个模型生成古诗的能力，却因为作为参考语料的《唐诗三百首》同样也属于两个大模型的训练语料，实验结果并不具备可信度。之后，利用第六届“诗词中国”传统诗词创作大赛的获奖作品，结合《唐诗三百首》中的律诗作品以及一些质量明显较差的打油诗，构建一个质量分层的数据集，证明了系统评分功能的有效性和优越性，还同时测试讨论了 BLEU 等自动度量指标的有效性。最后，通过三种不同的模态组合来测试文图模态的影响，发现图像模态的加入并没有显著提升生成古诗的质量，但图像模态仍然能够帮助用户更好地表达创作意图，提升系统的可用性与交互体验。

综上所述，本文完成了一个较为完整的基于大模型的跨模态古诗生成系统设计与实现，在功能实现和检验实验表现上取得了阶段性成果，探索了大模型在中华传统文化的再创造方面的巨大潜力，但仍存在一些不足与改进空间。首先，本系统并未实现在向量层面上的文图融合，而是通过文本描述的方式进行模态间的转换，可能导致信息损失或语义偏差，而文图模态输入对生成过程的作用仍

有待进一步验证，在古诗形式质量上遇到瓶颈后，如何论证实文图模态的必要性和优越性是一个问题。此外，评分机制的准确性十分依赖于大模型自身的质量，尽管本文精心设计的评分体系表现较好，但仍然存在一定的主观性与不确定性，尤其在对古诗意象的理解与文化内涵的把握上。最后，模型 DeepSeek-R1 的推理耗时较长，每次调用都需要消耗近一分钟的时间，可能影响用户体验。

## 参考文献

- [1] Oliveira H G. PoeTryMe: a versatile platform for poetry generation[J]. Computational Creativity, Concept Invention, and General Intelligence, 2012, 1 : 21.
- [2] Yan R, Jiang H, Lapata M, et al. I, Poet: Automatic Chinese Poetry Composition through a Generative Summarization Framework under Constrained Optimization[A]. IJCAI '13 : Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence[C], Beijing, China : AAAI Press, 2013 : 2197–2203.
- [3] He J, Zhou M, Jiang L. Generating Chinese Classical Poems with Statistical Machine Translation Models[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2012, 26(1) : 1650–1656.
- [4] Yi X, Li R, Sun M. Generating Chinese Classical Poems with RNN Encoder-Decoder[A]. Sun M, Wang X, Chang B, et al. Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data[C], Cham : Springer International Publishing, 2017 : 211–223.
- [5] Zhang X, Lapata M. Chinese Poetry Generation with Recurrent Neural Networks[A]. Moschitti A, Pang B, Daelemans W. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)[C], Doha, Qatar : Association for Computational Linguistics, 2014 : 670–680.
- [6] Li P, Zhang H, Liu X, et al. Rigid Formats Controlled Text Generation[A]. Jurafsky D, Chai J, Schluter N, et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics[C], Online : Association for Computational Linguistics, 2020 : 742–751.
- [7] Hu Z, Liu C, Feng Y, et al. PoetryDiffusion: Towards Joint Semantic and Metrical Manipulation in Poetry Generation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(16) : 18279–18288.

- [8] Zhang Z, Han X, Liu Z, et al. ERNIE: Enhanced Language Representation with Informative Entities[J], 2019(arXiv:1905.07129).
- [9] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning[J], 2025(arXiv:2501.12948).
- [10] Lu H, Liu W, Zhang B, et al. DeepSeek-VL: Towards Real-World Vision-Language Understanding[J], 2024(arXiv:2403.05525).
- [11] Yu C, Zang L, Wang J, et al. CharPoet: A Chinese Classical Poetry Generation System Based on Token-free LLM[A]. Cao Y, Feng Y, Xiong D. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)[C], Bangkok, Thailand: Association for Computational Linguistics, 2024 : 315–325.
- [12] Liu L, Wan X, Guo Z. Images2Poem: Generating Chinese Poetry from Image Streams[A]. MM '18 : Proceedings of the 26th ACM International Conference on Multimedia[C], New York, NY, USA : Association for Computing Machinery, 2018 : 1967–1975.
- [13] Xu L, Jiang L, Qin C, et al. How Images Inspire Poems: Generating Classical Chinese Poetry from Images with Memory Networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1) : 5618–5625.
- [14] Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning[J], 2017(arXiv:1702.08608).
- [15] Wang Z, He W, Wu H, et al. Chinese Poetry Generation with Planning Based Neural Network[A]. Matsumoto Y, Prasad R. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers[C], Osaka, Japan: The COLING 2016 Organizing Committee, 2016 : 1051–1060.
- [16] 王乐为, 余鹰, 张应龙. 基于 Seq2Seq 模型的自定义古诗生成 [J]. 计算机科学与探索, 2020, 14(6) : 1028–1035.

- [17] Wu C, Wang J, Yuan S, et al. Generate Classical Chinese Poems with Theme-Style from Images[J]. Pattern Recognition Letters, 2021, 149 : 75 – 82.
- [18] Shao Y, Shao T, Wang M, et al. A Sentiment and Style Controllable Approach for Chinese Poetry Generation[A]. CIKM '21 : Proceedings of the 30th ACM International Conference on Information & Knowledge Management[C], New York, NY, USA : Association for Computing Machinery, 2021 : 4784 – 4788.
- [19] 李晓辰. 风格和情感控制的中国古诗生成 [D]. 哈尔滨：哈尔滨工业大学, 2022.
- [20] Ren X, Chai X, Mao M. Generating Chinese Poetry from Images Based on Deep Learning[J]. Proceedings of the 2023 3rd International Conference on Big Data, Artificial Intelligence and Risk Management, 2023 : 134 – 138.
- [21] 曾柯. 生成式语言模型在古诗生成中的优化 [D]. 上海：华东师范大学, 2024.
- [22] Liu D, Guo Q, Li W, et al. A Multi-Modal Chinese Poetry Generation Model[A]. 2018 International Joint Conference on Neural Networks (IJCNN)[C], 2018 : 1 – 8.
- [23] Liu Y, Liu D, Lv J. Deep Poetry: A Chinese Classical Poetry Generation System[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(09) : 13626 – 13627.
- [24] Papineni K, Roukos S, Ward T, et al. BLEU: A Method for Automatic Evaluation of Machine Translation[A]. Isabelle P, Charniak E, Lin D. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics[C], Philadelphia, Pennsylvania, USA : Association for Computational Linguistics, 2002 : 311 – 318.
- [25] Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries[A]. Text Summarization Branches Out[C], Barcelona, Spain : Association for Computational Linguistics, 2004 : 74 – 81.

- [26] Li J, Galley M, Brockett C, et al. A Diversity-Promoting Objective Function for Neural Conversation Models[A]. Knight K, Nenkova A, Rambow O. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies[C], San Diego, California : Association for Computational Linguistics, 2016 : 110–119.
- [27] Wieting J, Bansal M, Gimpel K, et al. Towards Universal Paraphrastic Sentence Embeddings[J], 2016(arXiv:1511.08198).
- [28] Deng L, Wang J, Liang H, et al. An Iterative Polishing Framework Based on Quality Aware Masked Language Model for Chinese Poetry Generation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(05) : 7643 – 7650.
- [29] Ma J, Zhan R, Wong D F. Yu Sheng: Human-in-Loop Classical Chinese Poetry Generation System[A]. Croce D, Soldaini L. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations[C], Dubrovnik, Croatia : Association for Computational Linguistics, 2023 : 57–66.
- [30] Chen Z, Cao Y. A Polishing Model for Machine-Generated Ancient Chinese Poetry[J]. Neural Processing Letters, 2024, 56(2) : 77.
- [31] Radford A, Kim J W, Hallacy C, et al. Learning Transferable Visual Models From Natural Language Supervision[A]. Proceedings of the 38th International Conference on Machine Learning[C], Online : PMLR, 2021 : 8748 – 8763.
- [32] Zhu D, Chen J, Shen X, et al. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models[J], 2023(arXiv:2304.10592).
- [33] Wu Z, Chen X, Pan Z, et al. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding[J], 2024(arXiv:2412.10302).

- [34] Brown T, Mann B, Ryder N, et al. Language Models Are Few-Shot Learners[A]. Advances in Neural Information Processing Systems : Vol 33[C], Vancouver, Canada : Curran Associates, Inc., 2020 : 1877–1901.
- [35] Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models[J]. Advances in Neural Information Processing Systems, 2022, 35 : 24824–24837.
- [36] Wang X, Wei J, Schuurmans D, et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models[J], 2023(arXiv:2203.11171).
- [37] Nigh M. ChatGPT3-Free-Prompt-List[EB/OL]. <https://github.com/matnigh/ChatGPT3-Free-Prompt-List>. 2023.
- [38] 刘若缇. 唐代诗格声律论研究 [D]. 新北：淡江大学, 2011.
- [39] Li J, Song Y, Zhang H, et al. Generating Classical Chinese Poems via Conditional Variational Autoencoder and Adversarial Training[A]. Riloff E, Chiang D, Hockenmaier J, et al. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing[C], Brussels, Belgium : Association for Computational Linguistics, 2018 : 3890–3900.
- [40] 吴林东, 何向真, 万福成. 融合韵律特征的诗歌生成模型 [J]. 计算机工程与应用, 2024, 60(13) : 162–170.
- [41] 李飞跃. 唐诗格律的统计分析及问题 [J]. 文学遗产, 2022(05) : 44–56.
- [42] Zhipeng G, Yi X, Sun M, et al. Jiuge: A Human-Machine Collaborative Chinese Classical Poetry Generation System[A]. Costa-jussà M R, Alfonseca E. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations[C], Florence, Italy : Association for Computational Linguistics, 2019 : 25–30.
- [43] Bird S, Loper E. NLTK: The Natural Language Toolkit[A]. Proceedings of the ACL Interactive Poster and Demonstration Sessions[C], Barcelona, Spain : Association for Computational Linguistics, 2004 : 214–217.

- [44] Chen J. rouge-chinese[EB/OL]. [https://github.com/Isaac-JL-Chen/rouge\\_chinese](https://github.com/Isaac-JL-Chen/rouge_chinese). 2022.
- [45] Guo Z, Hu J, Son M. Bert-CCPoem[EB/OL]. <https://github.com/THUNLP-AIP/oet/BERT-CCPoem>. 2022.
- [46] 古诗文网 [EB/OL]. <https://www.gushiwen.cn/gushi/tangshi.aspx>. 2025.
- [47] 第六届“诗词中国”传统诗词创作大赛 [EB/OL]. <https://www.shicizhongguo.cn>. 2023.

## 致谢

山上人来人往，樱花已落了第四回。

感谢父母赠我看世界的眼睛和躁动的心脏，我得以在田野上追逐奔跑，自由又笨拙地。

感谢爱人吴思怡，于我困顿时予陪伴、失落时予温暖，让我重拾联结与爱的信心。此去经年，共赏良辰好景，何其有幸。

感谢好友杨涵溟，灰白的墙上喷绘着张狂的思想，缺憾的世界交由你我改造，友谊长存。

感谢朱卫平老师，鼓励支持我探索感兴趣的方向，引我窥见学术的一角，知遇之恩难报，祝桃李天下、万事顺遂。

感谢计算机学院，四年的学习为我打开了代码与逻辑的世界，这将始终是我探索世界的底气。

感谢武汉大学，珞珈山下东湖畔，灵气实在养人。这里包容着我的幼稚、我的肆意、我的思索与挣扎，也在悄然中目送我的远行。

犹记步入校门时的稚嫩与执着，那时的我怀有太多的期待与困惑，对世界更对自己。时间证明，自我二字贯穿了我四年的光阴，质疑、解构、挣扎与释然，一路磕磕绊绊，摸过的石头一步步锚定着我的存在。

所幸，澎湃的挣扎渐渐隐去，林中只见晨光熹微。

我与我周旋久，宁作我。

## A 成果

1. Yang L, Zhang Z, Niu K, et al. Large Model Based Crossmodal Chinese Poetry Creation[A]. 2024 IEEE Smart World Congress (SWC)[C], Nadi, Fiji: IEEE, 2024 : 27 - 34.