

# CCAC2025

## 第五届中国情感计算大会评测研讨会

Evaluation Seminar in the 5th Chinese Conference on  
Affective Computing

廖健 虞剑飞 连政 金琴 赵金明

魏忠钰 张伟男 杨亮 樊小超 陈鑫

主编

2025年7月

四川 成都



# 第五届中国情感计算大会评测委员会

## 评测主办机构

中国中文信息学会情感计算专业委员会

## 评测组织机构

中国人民大学

启元实验室

复旦大学

哈尔滨工业大学

大连理工大学

新疆师范大学

山西大学

太原科技大学

厦门大学

国家语言资源监测与研究平面媒体中心

## 评测委员会

CCAC 评测工作组:

廖 健 山西大学

虞剑飞 南京理工大学

连 政 中国科学院自动化研究所

赛道一组织:

金 琴 中国人民大学、启元实验室

赵金明 中国人民大学、启元实验室

赛道二组织:

魏忠钰 复旦大学

张伟男 哈尔滨工业大学

赛道三组织:

杨 亮 大连理工大学

赛道四组织:

樊小超 新疆师范大学

赛道五组织:

廖 健 山西大学

王素格 山西大学

陈 鑫 太原科技大学

郑泽芝 厦门大学

杨尔弘 国家语言资源监测与研究平面媒体中心

评测顾问专家：

王素格 山西大学

黄萱菁 复旦大学

赵妍妍 哈尔滨工业大学

杨 亮 大连理工大学

蒋昌建 复旦大学

评测委员（按姓氏拼音排序）

白泽文 大连理工大学

曾景杰 大连理工大学

程道翔 新疆师范大学

黄兆培 中国人民大学、启元实验室

李鹏帅 山西大学

梁敬聪 复旦大学

林嘉昱 复旦大学

吴国正 中国人民大学、启元实验室

肖瑞宇 哈尔滨工业大学

徐 迪 新疆师范大学

颜勤沛 新疆师范大学

杨 琦 大连理工大学

伊胜利 新疆师范大学

张佳雯 山西大学

张鑫洁 中国人民大学、启元实验室

# 第五届中国情感计算大会技术评测总体报告

廖健<sup>1</sup> 虞剑飞<sup>2</sup> 连政<sup>3</sup> 金琴<sup>4</sup> 赵金明<sup>4</sup> 魏忠钰<sup>5</sup> 张伟男<sup>6</sup>

杨亮<sup>7</sup> 樊小超<sup>8</sup> 陈鑫<sup>9</sup>

<sup>1</sup>山西大学, 太原, 030006

<sup>2</sup>南京理工大学, 南京, 210094

<sup>3</sup>中国科学院自动化研究所, 北京, 100190

<sup>4</sup>中国人民大学/启元实验室, 北京, 100086

<sup>5</sup>复旦大学数据智能与社会计算实验室, 上海, 200433

<sup>6</sup>哈尔滨工业大学, 哈尔滨, 150006

<sup>7</sup>大连理工大学, 苏州, 116081

<sup>8</sup>新疆师范大学, 乌鲁木齐, 830054

<sup>9</sup>太原科技大学, 太原, 030006

**摘要:** 本文详细描述了第五届中国情感计算大会 (CCAC2025) 技术评测的总体情况。本届评测共设定了五个评测赛道: (1) 多模态交互中的情感变化推理; (2) 第五届智慧论辩评测; (3) 四格漫画理解评测; (4) 中文多模态讽刺计算; (5) 细粒度比拟句情绪分析。共有国内外 121 支队伍报名, 分别参加了不同的评测赛道, 45 支队伍提交了最终结果。本文主要针对评测的组织、语料的标注、评测指标以及评测结果的情况进行了详细的说明。  
**关键词:** 情感计算, 评测, 多模态情绪对话, 智慧论辩, 漫画理解, 讽刺计算, 修辞分析

## Overview of the Technical Evaluation Seminar in the 5th Chinese Conference on Affective Computing

Jian Liao<sup>1</sup>, Jianfei Yu<sup>2</sup>, Zheng Lian<sup>4</sup>, Qin Jin<sup>4</sup>, Jinming Zhao<sup>4</sup>, Zhongyu Wei<sup>5</sup>,  
Weinan Zhang<sup>6</sup>, Liang Yang<sup>7</sup>, Xiaochao Fan<sup>8</sup>, Xin Chen<sup>9</sup>

<sup>1</sup>Shanxi University, Taiyuan, 030006

<sup>2</sup>Nanjing University of Science and Technology, Nanjing, 210094

<sup>3</sup>Institute of Automation, Chinese Academy of Sciences, Beijing, 100190

<sup>4</sup>Renmin University of China / Qiyuan Lab, Beijing, 100086

<sup>5</sup>Data Intelligence and Social Computing Lab, Fudan University, Shanghai, 200433

<sup>6</sup>Harbin Institute of Technology, Harbin, 150006

<sup>7</sup>Dalian University of Technology, Dalian, 116081

<sup>8</sup>Xinjiang Normal University, Urumqi, 830054

<sup>9</sup>Taiyuan University of Science and Technology, Taiyuan, 030006

**Abstract:** This paper provides a detailed description of the overall situation of the technical evaluation of the 5<sup>th</sup> China Conference on Affective Computing (CCAC2025). This evaluation consists of five competition tasks: (1) Inference of Emotional Changes in Multimodal Interaction; (2) AI-Debater 2025; (3) Four-panel Comic Understanding Evaluation; (4) Multimodal Chinese Sarcasm Computation; (5) Fine-grained Bini Sentence Emotion Analysis. A total of 121 domestic and international teams registered, participating in different competition tasks, with 45 teams submitting final results. The paper mainly focuses on the organization of the evaluation, annotation of corpora, evaluation criteria, and the results of the evaluation.

**Keywords:** Affective Computing, Evaluation, Multimodal Emotional Dialogue, Intelligent Debate, Comic Understanding, Sarcasm Computation, Rhetorical Analysis

## 1 引言

情感计算作为人工智能领域的一个重要分支,旨在赋予机器理解和处理人类情感的能力,其研究涵盖了自然语言处理、计算机视觉、心理学等多个领域,具有重要的理论意义和应用价值。在情感计算技术深度融入大模型时代的背景下,情感计算正从单一模态特征提取向跨模态协同推理、从静态情感识别向动态认知建模、从表层情绪感知向深层心理机制解析的维度跃迁。

在情感计算的发展历程中,各类评测活动起到了举足轻重的作用。从早期的 COAE、TREC 到国际知名的 SemEval,这些评测不仅推动了情感计算技术的进步,也为学术界和产业界提供了交流的平台。它们通过对不同赛道和数据的评测,不断推动着情感计算技术的创新与发展。2022年,第二届情感计算大会(CCAC)开始组织技术评测,目前情感计算大会评测已经成功举办了四届,逐渐成为我国情感计算领域最专业、最权威的评测活动之一,有力推动了情感计算理论和技术的研究发展与技术应用。本届评测由中文信息学会情感计算专委会主办,由中国人民大学、启元实验室、复旦大学、哈尔滨工业大学、大连理工大学、新疆师范大学、山西大学、太原科技大学、厦门大学以及国家语言资源监测与研究平面媒体中心联合承办。

本年度评测体系以“多模态认知智能的突破性发展”为核心命题,系统构建了面向复杂现实场景的情感计算评估框架,本年度评测创新性设置五大技术攻坚方向,涵盖了情感计算领域的多个热点问题。包括:(1)“多模态交互中的情感变化推理”,在多模态交互维度突破情感动态演化的时空建模瓶颈;(2)“第五届智慧论辩评测”,在认知推理层面深化论辩逻辑的因果链条解析能力;(3)“四格漫画理解评测”,通过视觉叙事系统解构漫画语义的跨模态映射机制;(4)“中文多模态讽刺计算”,依托多模态讽刺识别建立反讽意图的跨通道验证范式;(5)“细粒度比拟句情绪分析”,关注本体-喻体-喻底的多层关联深化比拟修辞格与情绪表达关联的认知。这些评测任务不仅涵盖自然语言理解、计算机视觉、认知心理学等核心领域的交叉创新,更通过“情感-认知-表达”三维评估体系,着力解决大模型时代情感计算面临的模态鸿沟、逻辑可信度、文化语境适应性等关键问题,为构建具有社会感知能力的下一代智能系统提供基准验证与理论支撑。

通过这些评测赛道,我们旨在建立并完善情感计算研究的基础数据集和评测标准,为情感计算领域的技术进步提供有力支持。在本次评测中,我们见证了众多优秀团队和研究成果的涌现,这些成果不仅为学术界提供了宝贵的研究经验,也为产业界带来了新的应用前景。

本文具体安排如下:第二节介绍 CCAC2025 技术评测的赛道设置,包括每个赛道的详细说明。第三节说明参加评测的情况以及各个单位的评测结果,最后给出结论和展望。报告附录整理了各赛道获奖队伍撰写的技术报告。

## 2 赛道描述

CCAC2025 技术评测共设置 5 个赛道，每个赛道下细分为若干个赛题。各赛道之间没有必然的联系，参赛队伍可以任选一个或者多个赛道。下表 1 给出了各赛道的大致情况。对于每个赛道我们全部人工进行答案标注，在发布评测语料的同时也向各个参赛队伍提供了部分训练集和验证集，但是不限定参赛队伍只使用我们所提供的训练样例，可以使用任意外部资源（词典、知识库、工具等），但需注明使用的外部资源。本小节首先介绍 CCAC2025 技术评测的时间安排以及参赛队伍情况，其次对于每个赛道进行详细的说明。

表 1 CCAC2025 技术评测赛道简表

Tab.1 Summary of CCAC2025 Technology Evaluation Tasks

| 编号   | 组织单位  | 赛道名称          | 赛道描述   |
|------|---|---------------|--|
| 赛道 1 | 中国人民大学<br>AI·M <sup>3</sup> 实验室<br>启元实验室    | 多模态交互中的情感变化推理 | 输入双人多模态对话片段，包含对话视频、对话文本、每句话的情感状态，此外还会指定出要关注的交互者及其出现情感变化的位置，要求输出该交互者出现对应情感变化的原因，包括多模态的刺激因素，交互者的认知评价以及交互者最终的情感反应。  |
| 赛道 2 | 复旦大学<br>哈尔滨工业大学                             | 第五届智慧辩论评测     | 本届智慧辩论评测包含自主论辩任务、自主论辩赋能社会调查两个任务，其中任务一包含辩题识别，论据挖掘，辩论构建和自主论辩生成四个子任务。   |
| 赛道 3 | 大连理工大学                                      | 四格漫画理解评测      | 包含两个任务，分别从逻辑理解和上下文推理的方面来评估模型对四格漫画的理解能力。要求模型准确判断该漫画正确的阅读顺序，并依据漫画整体的情节、画面风格以及前后关联等上下文信息，生成一段被掩码画面格的文本描述，还原漫画情节。  |
| 赛道 4 | 新疆师范大学                                      | 中文多模态讽刺计算     | 中文多模态讽刺计算旨在综合分析深度融合文本、音频和视觉等多种模态的信息，以实现讽刺表达方式的全面理解和精准识别，包括两个任务：1) 识别多模态数据中含有反讽意味的视频片段；2) 对于反讽视频片段，识别其细粒度的分类。   |
| 赛道 5 | 山西大学<br>太原科技大学<br>厦门大学<br>国家语言资源监测与研究平面媒体中心 | 细粒度比拟句情绪分析    | 本任务细粒度分析比拟句涉及本体、喻体、喻底（本体和喻体间的相似依据）的多层关联，准确建模情绪，旨在深化比拟修辞格与情绪表达关联的认知。包含 2 个阶段，共计 3 个任务。第一阶段要求（1）从对比拟句所蕴含的情绪类别进行判断，（2）从中抽取细粒度本体及喻体要素。第二阶段根据给定的本体和喻体，结合比拟句子信息判断喻底。 |

## 2.1 CCAC2025 技术评测时间安排

本届评测从评测启动到评测论坛召开前后历时近 7 个月，分为评测赛道征集、评测大纲发布、评测数据准备、评测数据集、评测结果发布、评测论坛报告等不同阶段。表 2 给出了具体的时间安排：

表 2 CCAC2025 技术评测时间表  
Tab.2 Schedule of CCAC2025 Technology Evaluation

| 时间                   | 事项                   |
|----------------------|----------------------|
| 2025 年 1 月 18 日      | CCAC2025 技术评测正式启动    |
| 2025 年 1 月 19 日      | 发布 CCAC2025 评测赛道征集启事 |
| 2025 年 4 月 1 日       | 任务发布与报名启动            |
| 2025 年 5 月上旬         | 训练集语料发布              |
| 2025 年 6 月上旬         | 测试集语料发布              |
| 2025 年 6 月中旬         | 提交截止                 |
| 2025 年 6 月下旬         | 比赛结果公布               |
| 2025 年 7 月 18 日-20 日 | CCAC2025 大会召开及颁奖典礼   |

## 2.2 参赛队伍情况

CCAC2025 技术评测得到了国内外同行的大力支持，共有 121 支队伍报名参加评测，最终有 45 支队伍提交结果。表 3 针对给出了每个赛道评测单位参加的情况。

表 3 CCAC2025 技术评测参赛情况  
Tab.3 Number of Participants in Each Tracks

| 赛道   | 参赛队伍数 | 提交结果单位数 |
|------|-------|---------|
| 赛道 1 | 37    | 9       |
| 赛道 2 | 23    | 14      |
| 赛道 3 | 13    | 4       |
| 赛道 4 | 29    | 9       |
| 赛道 5 | 19    | 9       |
| 总数   | 121   | 45      |

## 2.3 赛道一

### 2.3.1 任务背景

多模态交互中的情感变化推理旨在理解交互过程中目标对象情感变化的多模态影响因素，在自然人机交互、教育、医疗等场景具有广泛的应用。与传统的多模态情感识别任务不同，情感变化推理关注于对交互者情感原因及情感产生过程的深度理解。该任务需要充分挖

掘多模态情境当中的情感刺激因素，建模交互者由刺激引发的认知评价，进而推断出其情感反应。本评测项目鼓励参赛者从不同模态不同角度对多模态对话中的情感变化进行建模和推理，旨在推动多模态交互中情感理解相关研究的发展。

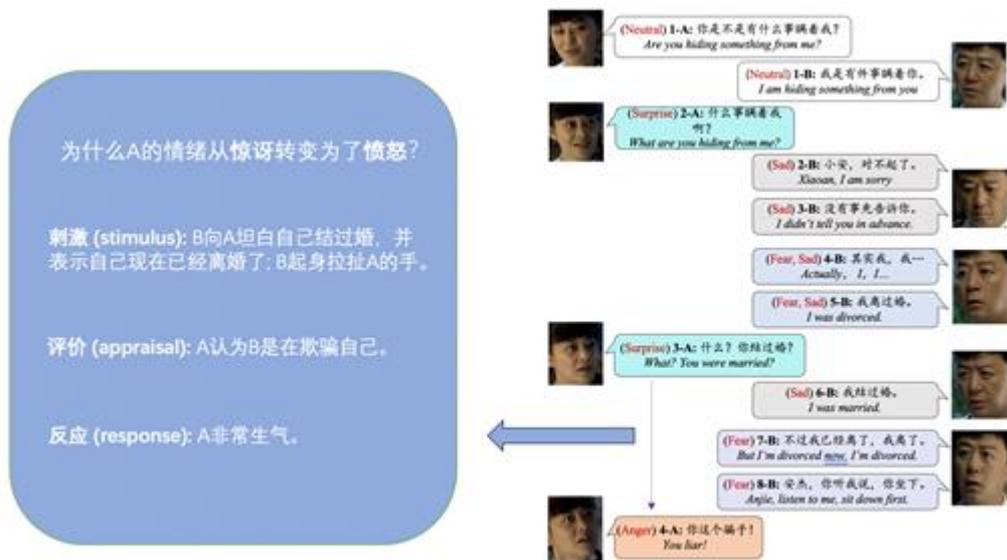
### 2.3.2 任务介绍

本届多模态对话中的情感识别评测任务将采用 INSIDE (emotIon chaNge reaSonIng In multimoDal convErsations) 数据集作为支撑数据集，该任务的具体介绍如下：

多模态交互中的情感变化推理评测任务旨在对于多模态交互场景中交互者出现的情感变化情况进行深度理解。任务输入是双人多模态对话片段，包含对话视频、对话文本、每句话的情感状态，此外还会指定出要关注的交互者及其出现情感变化的位置，要求输出该交互者出现对应情感变化的原因，包括多模态的刺激因素，交互者的认知评价以及交互者最终的情感反应。

### 2.3.3 数据集描述

本次技术评测使用的标注数据集为 INSIDE 数据集，由中国人民大学 AI·M<sup>3</sup>实验室提供，在 M3ED 数据集的基础上进一步标注而成。INSIDE 数据集共计 4147 个样本，包含丰富的情感互动。我们将发布视频片段所对应的视频信息、文本信息以及对应的标注信息，其中文本信息和标注信息将以 JSON 格式发布，数据样例如下：



### 2.3.3 评测指标

本次评测将基于 METEOR 和 BERTScore 两项指标的综合排名进行评定。具体方法如下：首先，分别根据 METEOR 和 BERTScore 两项指标对所有参赛队伍进行独立排名；然后，将每支队伍在两项指标中的排名位次相加，总和最小的队伍位列前茅。若两支队伍的综合排名总和相同，则以 BERTScore 的高低作为最终排序依据，得分较高者优先。

具体任务细节参见：

[https://github.com/AIM3-RUC/MECR\\_CCAC2025](https://github.com/AIM3-RUC/MECR_CCAC2025)

## 2.4 赛道二

### 2.4.1 评测背景

在本届大会中，我们将举办第五届智慧论辩评测。论辩是人类智慧的一项重要技能，在诸多人类活动中承担着不可或缺的作用。计算论辩技术关注机器对人类论辩过程的理解和模仿，广泛应用于决策辅助、写作支持和逻辑审查等场景，也可以在社会仿真中发挥重要作用，于近年来逐渐成为人工智能研究的新兴重要分支。本届智慧论辩评测由中国中文信息学会情感计算专委会（CIPS-CCAC）主办，复旦大学数据智能与社会计算实验室和哈尔滨工业大学社会计算与信息检索研究中心共同承办，欢迎各界人士参与。

### 2.4.2 评测内容

本届智慧论辩评测包含网络媒体内容论辩理解、自主论辩生成、论辩赋能社会调查仿真三个任务。

#### 赛题一：自主论辩任务

该赛题包含辩题识别，论据挖掘，辩论构建和自主论辩生成四个子任务。在最终评测中，取四个子任务的平均分数作为最终分数。

##### 子任务一：辩题识别

给定一个论点和一个辩题对，判断当前论点是否属于当前辩题，属于当前辩题输出 1，不属于当前辩题输出 0。

##### 数据样例

- 输入：{'辩题': '应该教学生竞争还是合作?', '论点': '我们应该更加重视小学教育中的合作'}
- 输出：1

##### 评价指标：Acc

##### 子任务二：论据挖掘

给定一个论点和一个论据，判断当前论据是否可以支持当前论点，可以支持当前论点输出 1，不可以支持当前论点输出 0。

##### 数据样例

- 输入：{'论点': '通过合作，孩子们可以学习人际交往技能，这对所有学生的未来生活都很重要', '论据': '研究表明，在合作的过程中，孩子们可以学习如何倾听他人的意'}

见，如何与他人沟通，如何全面思考，甚至在发生冲突时如何与其他团队成员妥协  
}

- 输出：1

**评价指标：Acc**

### 子任务三：辩论构建

给定一篇立论陈词稿和陈词稿中的一个论点，判断当前论点在陈词稿中的结构类型。结构类型包括主要论点、普通论点和论点前提，直接输出论点结构类型。

#### 数据样例

- 输入：{'陈词稿': '(以上省略) 首先，通过合作，孩子们可以学到人际交往的技巧，这对所有学生未来的生活都意义重大。我们从团队合作中学到的不仅是如何与他人达成相同的目标，更重要的是如何与他人相处。(以下省略)'; '论点': '通过合作，孩子们可以学到人际交往的技巧，这对所有学生未来的生活都意义重大。'}  
• 输出：主要论点

**评价指标：Acc**

### 子任务四：自主论辩生成

给定一个辩题及对应持方，由参赛模型自动生成一篇一辩立论陈词稿。

#### 数据样例

- 输入：{'辩题': '死刑是否应该合法化', '持方': '正方'}  
• 输出：'尊敬的评委、对方辩友、各位观众：今天，我们讨论的辩题是"死刑应当合法化"。我方坚定认为，死刑作为刑罚体系的组成部分，具有不可替代的法治价值与社会功能。我将从以下四个层面阐述我方观点。(以下省略)'

**评价指标：人工评价**

## 赛题二：论辩赋能社会调查仿真

基于人口统计学信息构建大模型仿真个体，通过多智能体辩论机制激发观点交锋，促使个体在认知迭代中形成更精准的社会调查响应。

#### 数据样例

- 输入：{'个体信息': '{"AGE": "Middle-aged (36-65 years old)", "GENDER": "Male", ...}', '问卷题目': 'Some people believe that we should spend less money for defense. Which of the following best describes your view? Options: A.Refused B.Decrease C.Increase'}  
• 输出：C

**评价指标：**

50%多智能体辩论得分: Debatrix (点此了解更多) 与人工评价  
50%最终回答 F1 Score

### 2.4.3 评测数据

网络媒体内容辩论理解数据集: 中文数据集, 包含论点所属的辩题类别、论点论据间的关系等标注信息。在 PE 辩论挖掘数据集的基础上使用 DeepSeek-R1 翻译, 并由专业人员进行校对、修正, 整理为 JSON 文件。

自主辩论生成数据集: 中文数据集, 包含辩题与一辩立论陈词稿的数据对。辩题来源于近年知名华语辩论赛赛题, 由 DeepSeek-R1, GPT4-turbo, Gemini-2.5 等先进闭源大模型自动构建, 并由专业人员进行校对、修正, 整理为 json 文件。

社会调查仿真数据集: 源于某国某年的真实社会调查数据, 清洗标签缺失、拒绝回答等样本后整理为 json 文件。

具体任务细节参见:

<http://www.fudan-disc.com/sharedtask/AIDebater25/tracks.html>

## 2.5 赛道三

### 2.5.1 评测背景

当前, 大模型在单一模态理解方面表现优异, 但在处理融合叙事逻辑与情感线索的复杂多模态场景时仍存在技术瓶颈。四格漫画常作为多模态信息的综合载体, 是以四个画格组成的短篇叙事漫画, 遵循“起承转结”结构, 通过简洁画面和对话呈现完整情节, 常用于幽默、讽刺或生活化主题, 强调节奏紧凑与结局反转。其四联画板通过视觉叙事、文字对白和空间布局的协同作用构建连贯的语义逻辑链, 对现有模型理解能力提出了多维度挑战。因此, 我们基于该漫画数据, 设定了漫画理解评测任务, 探索模型的综合认知能力。

### 2.5.2 评测内容

#### 赛题一: 四格漫画逻辑理解

本任务评估模型对于输入四格漫画阅读顺序的逻辑理解能力。四格漫画第一格设定场景, 第二格发展情节, 第三格转折, 第四格收尾或搞笑, 整体结构紧凑。本任务为模型提供一个经过特殊处理的四格漫画, 该漫画的某两个画面格的顺序被交换, 要求模型准确判断该漫画正确的阅读顺序。

#### 数据说明:

对于中文漫画 (zh), 后缀为 \_geshi 的漫画表示元漫画, \_sp01 \_sp12 \_sp23 \_sprandom 分别表示交换 panel01, panel12, panel23 以及随机交换漫画四个 panel 的漫画变体, 同时我们给出了漫画变体的正确阅读顺序。

对于英文漫画（en），前缀为 true 的漫画表示元漫画，\_sp01 \_sp12 \_sp23 分别表示交换 panel01, panel12, panel23 的漫画变体。同时我们给出了漫画变体的正确阅读顺序。

在验证集和测试集中，中文漫画四个 panel 会被随机打乱，要求模型识别正确阅读顺序，英文漫画只会交换两个相邻 panel（sp01、sp12、sp23），要求模型识别正确阅读顺序。

**输入：**漫画

**输出：**正确阅读顺序，比如元漫画阅读顺序（0 1 2 3），sp01（1 0 2 3），sp12（0 2 1 3），sp23（0 1 3 2）注意使用空格连接

## 赛题二：四格漫画上下文推理

本次任务评估模型对于输入四格漫画上下文信息的理解能力。通过对输入四格漫画的某一画面格进行随机掩码，要求模型依据漫画整体的情节、画面风格以及前后关联等上下文信息，生成一段被掩码画面格的文本描述，还原漫画情节。

**数据说明：**

对应中文和英文漫画，我们均给出了漫画四个 panel 的文本描述。

在验证集和测试集中，我们会给出随机 mask 掉的一个 panel 的元漫画，以及剩下三个 panel 的文本描述，要求模型还原被 mask 的 panel 的描述。

**输入：**被随机 mask 一个 panel 的漫画以及剩下三个未被 mask 的 panel 的文本描述

**输出：**被 mask 的 panel 的文本描述

### 2.5.3 评测指标

**赛题一：**Macro-F1 score

**赛题二：**ROUGE, BLUE

具体任务细节参见：

<https://cut-viper-fec.notion.site/ccac2025-readme-1ebc5d03b30b8007bc90e9fc6bb74905>

## 2.6 赛道四

### 2.6.1 评测背景

讽刺作为一种复杂的语言现象，广泛存在于日常对话、文学作品及娱乐节目中。它通过表面上看似正面或中立的语言来传达负面或批评性的含义，根据其表现形式的不同，可以细分为多个类别。多模态讽刺计算任务旨在综合分析并深度融合文本、音频和视觉等多种模态的信息，以实现讽刺表达方式的全面理解和精准识别。这一领域的研究不仅提升了机器理解复杂语言现象的能力，增强了人机交互的自然性和准确性，还在多媒体内容分析、社交媒体监控以及跨文化交流等多个方面展现了重要的应用价值。

### 2.6.2 评测内容

中文多模态讽刺计算包括两个任务：1) 识别多模态数据中含有反讽意味的视频片段；2)

对于反讽视频片段，识别其细粒度的分类。

### 赛题一：多模态讽刺识别

多模态讽刺识别任务是指，对于给定的多模态讽刺数据（包括文本、音频、视频等），其目标在于结合文本语义、语调、表情、肢体语言等多模态信息，判断其是否具有讽刺意味。

子任务一，多模态讽刺识别：对于给定的多模态（文本数据标和对应的视频片段与音频片段），参赛模型需判断当前文本数据是否为讽刺语句。输出为 2 种标签：“1”表示该句子为讽刺语句，“0”表示该句子为非讽刺语句。

数据样例为文本数据和相应的视频片段与音频片段，文本数据为目标语句。

- 输入：但是刚才我看了一下，以前的这个造型，靓颖海豚界也有杀马特是吗？
- 输出：1
- 输入：张靓颖当之无愧的实力派的女歌手，因为唱歌的时候经常能发出滋儿滋儿的声音，所以被大家叫“海豚公主”。
- 输出：0

评价指标：macro F1。

### 赛题二：多模态讽刺类别识别

对于给定的多媒体数据（文本数据和对应的视频片段与音频片段），参赛模型需判断当前评论数据属于哪个讽刺类别。讽刺类别[2-3]说明及输出标签如下：

- a. Sarcasm（嘲笑），与事态相矛盾并持批评态度的评论，标注为 0；
- b. Irony（反语），与事态相矛盾但没有明显批评态度的评论，标注为 1；
- c. Satire（讽刺文学），看似支持但包含潜在分歧和嘲讽的评论，标注为 2；
- d. Understatement（轻描淡写），降低或破坏事态重要性的评论，标注为 3；
- e. Overstatement（夸大），用明显夸张的措辞描述事态的评论，标注为 4；
- f. Rhetorical question（反问），用反问语句引导读者推理出与实际情况明显相悖的含义的评论，标注为 5。

讽刺类别与 ID 的对应关系

```
{  
    "-1": "非讽刺",  
    "0": "Sarcasm",  
    "1": "Irony",  
    "2": "Satire",  
    "3": "Understatement",  
    "4": "Overstatement",  
    "5": "Rhetorical Question"  
}
```

数据样例

数据样例为文本数据和相应的视频片段与音频片段，文本数据为目标语句。

输入：但是没有见过现场解雇员工的，对吧？

输出：0

输入：还有分享《姐姐妹妹站起来》，这就是按错了。

输出：1

输入：怕江豚吃醋。

输出：2

输入：因为我不了解那些非冠军选手的生活状态。

输出：3

输入：但是我觉得张靓颖老师，近几年发展得一直都非常的好，作为一个季军。

输出：4

输入：但，对就像你刚刚的笑一样，全美人民都蒙了，因为从中国来个 Dolphin 干嘛？不是应该来个 panda 吗？

输出：5

**评价指标：** macro F1

### 2.6.3 评测数据

中文多模态讽刺计算数据集数据来源于国内著名的脱口秀节目《吐槽大会》。该数据集通过时间戳将视频内容切分为多个片段，每个视频片段均进行了详细的标注，包括：编号（subtitleNo）、说话人（speaker）、对应视频片段的起始结束时间（start、end）、文本内容（sentence）、是否幽默（isHumor）、幽默类别（humorType）、是否讽刺（isSarcasm）、讽刺目标（sarcasmTarget）、讽刺类别（sarcasmType）。具体文本标注标签如下：

"讽刺数据"

```
{
  "id": "2_1_3",
  "speaker": "张邵刚",
  "start": "00:07:45,900",
  "end": "00:07:53,027",
  "sentence": "笑什么笑，就你们弄的那个破维秘的氛围，不靠我这撑一下，这场子体面吗？",
  "isHumor": 1,
  "speakerEmotion": 1,
  "humorType": 4,
  "isSarcasm": 1,
  "sarcasmTarget": "李诞",
```

```

    "sarcasmType": 5
  }

  "非讽刺数据"
  {
    "subtitleNo": "2_1_7",
    "speaker": "池子",
    "start": "00:05:46,231",
    "end": "00:05:47,386",
    "sentence": "拐都拿起来了都。",
    "isHumor": 0,
    "speakerEmotion": 1,
    "humorType": -1,
    "isSarcasm": 0,
    "sarcasmTarget": "",
    "sarcasmType": -1
  }

```

其中" id": "2\_1\_3"表示其为第2季第1期的第3条文本数据,非讽刺及幽默时,其相应类别 sarcasmType、humorType 被标注为-1。

具体任务细节参见:

<https://github.com/MuYuU0/CCAC2025-Chinese-multimodal-sarcasm-calculation>

## 2.7 赛道五

### 2.7.1 评测背景

比拟是中文常见的修辞手法,可分为比喻、拟人和拟物三种类型。比拟通过将本体(描述对象)与喻体(比拟对象)建立相似性联系(喻底),以生动形象的方式传递情感和语义。为了增强文本情绪表达效果,比拟句常出现在文学文本、社交媒体等场景中,对机器深层挖掘文本语义及情绪信息提出了挑战。因此,本任务细粒度分析比拟句涉及本体、喻体、喻底(本体和喻体间的相似依据)的多层关联,准确建模情绪,旨在深化比拟修辞格与情绪表达关联的认知,推动计算语言学与认知科学的交叉研究,从而为文学分析、舆情分析及人机交互提供支撑。

### 2.7.2 评测内容

本届中文比拟情绪分析评测包含2个阶段,共计3个任务。第一阶段要求(1)从对比拟句所蕴含的情绪类别进行判断,(2)从中抽取细粒度本体及喻体要素。第二阶段根据给

定的本体和喻体，结合比拟句子信息判断喻底。3个任务采用统一的数据集，由于1条比拟句可能出现多个“本体-喻体-喻底”对，则每1条数据会提供若干个<本体、喻体、喻底、情绪>四元组标签。以下为各任务的具体介绍：

### 第一阶段：

#### 任务 1：比拟句情绪分析

目标：给定一个比拟句，判别该句子的情绪类别。

输出：8类情绪标签，0-7分别代表无情绪、乐、好、怒、哀、惧、恶、惊。

#### 任务 2：细粒度本体、喻体特征抽取

目标：抽取句子中所有本体（描述对象）及喻体（比拟对象）片段对。

输出：假设一句比拟句中包含 n 组<本体-喻体>对，输出<本体 1, 喻体 1>, .....<本体 n, 喻体 n>。

### 第二阶段：

#### 任务 3：<本体-喻体>对喻底判别

目标：给定比拟句信息和一组<本体-喻体>对，以单选题形式进行比拟喻底的判别。

问题：片段对（本体<sub>k</sub>, 喻体<sub>k</sub>）喻底是什么？

选项：A 选项 1 B 选项 2 C 选项 3 D 选项 4

输出：正确选项标签。

## 2.7.3 评测指标

任务 1 评价指标：宏平均-F1

任务 2 评价指标：F1（模糊匹配）

实体抽取采用模糊匹配：对于模型抽取的实体 e，人工标注的答案实体 a，覆盖度定义为： $s(e, a) = |e \cap a| / |e \cup a|$ ，若  $s(e, a) \geq 0.6$ ，则视为实体抽取正确。对于一组<本体-喻体>对，需本体、喻体实体同时满足正确条件时，视为<本体-喻体>对抽取正确。

任务 3 评价指标：ACC。

最终排名依据：三个任务的指标均值。

## 2.7.4 评测数据

在第一阶段任务中，训练集以及验证集均由六列数据组成，训练集包括 3582 个句子，4000 个本体及喻体片段对。验证集包括 897 个句子，1000 个本体及喻体片段对。测试集由两列数据组成，包括 4148 个句子。

在第二阶段任务中，训练集以及验证集均由十列数据组成，训练集包括 4000 条数据，验证集包括 1000 条数据。测试集由九列数据组成，包括 5000 条数据。

具体任务细节参见：

<https://github.com/sxu-nlp/Bini-EmoAnalysis/>

### 3 评测结果

对于每一项赛道，我们都会给出每个赛道部分提交队伍的评测排名结果。

#### 3.1 赛道一

赛道一的最终评测排名结果如下表 4 所示。

表 4 赛题 1 前 6 名评测结果  
Tab 4 The Details of Top 6 Submitted Runs in Track 1

| 序号 | 队伍名             | METEOR | BERTScore |
|----|-----------------|--------|-----------|
| 1  | 山西财经大学李大宇       | 0.5024 | 0.8100    |
| 2  | UESTC#01        | 0.4662 | 0.7944    |
| 3  | 深北莫 NLP         | 0.4701 | 0.7927    |
| 4  | NUSTM_Emotion   | 0.4595 | 0.7850    |
| 5  | 西华大学 525        | 0.3387 | 0.7486    |
| 6  | 智能语音与情感交互 AI4AI | 0.3329 | 0.7223    |

赛道一获奖队伍信息如下：

表 5 赛道一获奖队伍信息  
Tab 5 Information on the winning teams of Track 1

| 奖项  | 队伍名           | 参赛单位           | 队员                               | 指导教师    |
|-----|---------------|----------------|----------------------------------|---------|
| 一等奖 | 山西财经大学李大宇     | 山西财经大学         | 李大宇, 李旻, 陈鑫, 张文跃                 | 李大宇, 李旻 |
| 二等奖 | UESTC#01      | 电子科技大学         | 蒋培元, 张瀚文, 张宗舜, 彭雪枫, 杨嘉业, 刘露, 何义会 | 刘峤, 刘瑶  |
| 二等奖 | 深北莫 NLP       | 深圳北理莫斯科大学      | 张子蕴, 李焱杰, 胡宇宣, 李子宣, 刘惟楚          | 李成明     |
| 三等奖 | NUSTM_Emotion | 南京理工大学文本挖掘实验室  | 曹涵宇, 李阳阳                         | 虞剑飞, 夏睿 |
| 三等奖 | 西华大学 525      | 西华大学智能技术与系统实验室 | 倪思源, 洪陈, 齐庆, 罗 桢荣                | 李显勇     |

|     |                 |        |                        |         |
|-----|-----------------|--------|------------------------|---------|
| 三等奖 | 智能语音与情感交互 AI4AI | 北京邮电大学 | 吴俊阳, 王卓, 田天翊, 李启飞, 练慧俊 | 高迎明, 李雅 |
|-----|-----------------|--------|------------------------|---------|

### 3.2 赛道二

赛道二两个赛题的最终评测排名结果如下表 6 和 7 所示。

表 6 赛道二提交评测结果-赛题 1 自主论辩任务

Tab 6 The Details of Submit Runs in SubTrack 2.1

| 排名 | 队伍         | 成绩 (平均 Acc) |
|----|------------|-------------|
| 1  | 认知仿真       | 0.866       |
| 2  | HLT-Track1 | 0.813       |
| 3  | 正相关        | 0.790       |
| 4  | EX         | 0.785       |
| 5  | wzkkk      | 0.665       |

表 7 赛道二提交评测结果-赛题 2 论辩赋能社会调查仿真

Tab 7 The Details of Submit Runs in SubTrack 1

| 排名 | 队伍         | 成绩 (平均 Acc) |
|----|------------|-------------|
| 1  | 正相关        | 0.701       |
| 2  | EX         | 0.654       |
| 3  | HLT-Track2 | 0.563       |

赛道二获奖队伍信息如下：

表 8 赛道二获奖队伍信息

Tab 8 Information on the winning teams of Track 2

|      | 奖项  | 队伍名        | 参赛单位                      | 队员  | 指导教师 |
|------|-----|------------|---------------------------|---|------|
| 赛题 1 | 一等奖 | 认知仿真       | 中国科学院软件研究所天基综合信息系统全国重点实验室 | 李罗政, 陈昌明                                  | 李凌波  |
|      | 二等奖 | HLT-Track1 | 哈尔滨工业大学(深圳)计算机科学与技术学院     | 靳伯骏, 王华杰, 张宸华, 黄煜聪, 陈冠荣, 何宇航, 孙洋, 鲍建竹, 涂耿 | 徐睿峰  |
|      | 二等奖 | 正相关        | 齐鲁工业大学(山东省科学              | 赵学官, 张恒, 滕                                | 司加胜  |

|      |     |             |                              |                                       |            |
|------|-----|-------------|------------------------------|---------------------------------------|------------|
|      |     |             | 院)                           | 叶青, 高天智, 齐泽楠                          | 鹿文鹏        |
|      | 三等奖 | EX          | 新加坡管理大学 (SMU); 中国科学院沈阳自动化研究所 | 赖逸斌                                   | 傅博         |
|      | 三等奖 | wzkkk       | 广西大学电气工程学院                   | 王子恺                                   | 黄清宝        |
| 赛题 2 | 一等奖 | 正相关         | 齐鲁工业大学(山东省科学院)               | 赵学官, 张恒, 滕叶青, 高天智, 齐泽楠                | 司加胜<br>鹿文鹏 |
|      | 二等奖 | EX          | 新加坡管理大学 (SMU); 中国科学院沈阳自动化研究所 | 赖逸斌                                   | 傅博         |
|      | 二等奖 | HLT-Trac k2 | 哈尔滨工业大学(深圳)计算机科学与技术学院        | 王华杰, 靳伯骏, 孙洋, 张宸华, 黄煜聪, 陈冠荣, 何宇航, 鲍建竹 | 徐睿峰        |

### 3.3 赛道三

赛道三的最终评测结果如下表 9 所示。

表 9 赛道三提交评测结果-赛题 1 四格漫画逻辑理解

Tab 9 The Details of Submitted Runs in Track 3.1

| 排名 | 队伍          | 英文任务分数 | 中文任务分数 | 赛题 1 总分 |
|----|-------------|--------|--------|---------|
| 1  | 二次元情探       | 0.3984 | 0.4127 | 0.4056  |
| 2  | SCNU_JETKUN | 0.3281 | 0.3651 | 0.3466  |
| 3  | 不吃香菜队       | 0.3828 | 0.1905 | 0.2866  |
| 4  | 所以然         | 0.0000 | 0.4762 | 0.2381  |

表 10 赛道三提交评测结果-赛题 2 四格漫画上下文推理

Tab 10 The Details of Submitted Runs in Track 3.2

| 排名 | 队伍  | 英文任务分数 | 中文任务分数 | 赛题 2 总分 |
|----|-----|--------|--------|---------|
| 1  | 所以然 | 0.5162 | 0.6182 | 0.5672  |

赛道三获奖队伍信息如下：

表 11 赛道三获奖队伍信息

Tab 11 Information on the winning teams of Track 3

| 奖项     | 队伍名称  | 参赛单位          | 队员                      | 指导教师       |
|--------|-------|---------------|-------------------------|------------|
| 任务一优胜奖 | 二次元情探 | 河北大学数学与信息科学学院 | 刘志林, 郑安遇                | 王硕         |
| 任务二优胜奖 | 所以然   | 齐鲁师范学院        | 张文雅, 张云杰, 李翰林, 陈文锟, 刘德坤 | 李成飞<br>王春雨 |

### 3.4 赛道四

赛道四的最终评测结果如下表 12 所示。

表 12 赛道四提交队伍评测结果

Tab 12 The Details of Submitted Runs in Track 4

| 排名 | 队伍名称   | 最终得分   |
|----|--------|--------|
| 1  | 言传身教   | 0.5989 |
| 2  | 所以然    | 0.5964 |
| 3  | 我是大明星队 | 0.5487 |
| 4  | 天山棱镜   | 0.5389 |
| 5  | HIT416 | 0.5107 |
| 6  | 梯度消失   | 0.506  |
| 7  | 不用取名   | 0.4685 |
| 8  | 西瓜崽    | 0.4648 |
| 9  | 神仙保佑队  | 0.4573 |

赛道四获奖队伍信息如下：

表 13 赛道四获奖队伍信息

Tab 13 Information on the winning teams of Track 4

| 奖项  | 队伍名称 | 参赛单位   | 队员                           | 指导教师     |
|-----|------|--------|------------------------------|----------|
| 一等奖 | 言传身教 | 北京邮电大学 | 冷基栋、王斯然、谭嘉辰、黄林迪              | 闫强、易兰丽   |
| 二等奖 | 所以然  | 齐鲁师范学院 | 李翰林, 张云杰, 侯晓杰, 张文雅, 陈文锟, 刘德坤 | 李成飞, 王春雨 |

|     |        |              |                    |     |
|-----|--------|--------------|--------------------|-----|
| 二等奖 | 我是大明星队 | 中国科学院计算技术研究所 | 徐辉、杨其彘、沈武韬         | 张瑾  |
| 三等奖 | 天山棱镜   | 新疆大学         | 李炜杰、陈冠熊、张怡、胡少涛、朱佳宁 | 赵新元 |
| 三等奖 | HIT416 | 哈尔滨工业大学计算学部  | 鄂福顺                | 吴锐  |
| 三等奖 | 梯度消失   | 安徽工程大学       | 郭灏、耿治发             | 皇苏斌 |

### 3.4 赛道五

赛道五的最终评测结果如下表 14 所示，按照三个赛题的指标均值排序。

表 14 赛道五评测结果

Tab 14 The Details of Submitted Runs in Track 5

| 序号 | 队伍             | 赛题 1(Macro-F1) | 赛题 2(F1) | 赛题 3(ACC) | 平均     |
|----|----------------|----------------|----------|-----------|--------|
| 1  | IIE06          | 0.5731         | 0.7186   | 0.9935    | 0.7617 |
| 2  | 地铺稀客队          | 0.5516         | 0.6667   | 0.9895    | 0.7359 |
| 3  | 2025HIT 细粒度小分队 | 0.5023         | 0.6071   | 0.9745    | 0.6946 |
| 4  | 所以然            | 0.4468         | 0.5881   | 0.9850    | 0.6733 |
| 5  | ICKF           | 0.4670         | 0.5443   | 0.8840    | 0.6318 |
| 6  | 啊对对队           | 0.3948         | 0.5787   | 0.9165    | 0.6300 |
| 7  | 苏格拉            | 0.4199         | 0.4611   | 0.9085    | 0.5965 |
| 8  | DHfusion       | 0.4137         | 0.2893   | 0.8870    | 0.5300 |
| 9  | 我是大明星队         | 0.4696         | 0.1956   | 0.8680    | 0.5111 |

赛道五获奖队伍信息如下：

表 15 赛道五获奖队伍信息

Tab 15 Information on the winning teams of Track 5

| 奖项  | 队伍名   | 参赛单位         | 队员             | 指导教师 |
|-----|-------|--------------|----------------|------|
| 一等奖 | IIE06 | 中国科学院信息工程研究所 | 赵语林；孙业奇；李雨晴    | 林政   |
| 二等奖 | 地铺稀客队 | 北京理工大学       | 张兴辰；刘艺程；杨成；周祐超 | 史树敏  |

|         |                   |                    |                                     |         |
|---------|-------------------|--------------------|-------------------------------------|---------|
|         | 2025HIT<br>细粒度小分队 | 哈尔滨工业大学            | 范云天; 焦彦华; 刘铎; 张泽晰;<br>王梓名; 王渝凯; 白宸旺 | 苏统华     |
| 三等<br>奖 | 所以然               | 齐鲁师范学院             | 李翰林; 刘德坤; 张思锐; 赵北,<br>邵圣禹; 毕谊诚; 刘静暄 | 李成飞; 刘斌 |
|         | ICKF              | 西华大学               | 袁希鹏; 袁召毅; 徐欢; 兰诺天                   | 刘佳      |
|         | 啊对对队              | 西华大学计算机与<br>软件工程学院 | 仇朝廷; 王鸿洋; 范文豪                       | 杜亚军     |

## 4 结论与展望

本文对 CCAC2025 技术评测进行了详细的介绍和说明, 描述了赛道设置、语料建设、评测指标和评测结果。本次评测共设有 5 个评测赛道: (1)“多模态交互中的情感变化推理”; (2)“第五届智慧论辩评测”; (3)“四格漫画理解评测”; (4)“中文多模态讽刺计算”; (5)“细粒度比拟句情绪分析”。

赛道一聚焦多模态对话情感变化推理任务, 紧密契合人工智能情感计算领域的前沿研究方向, 对提升人机交互系统的情感认知能力具有重要价值。从评测结果可见, 头部队伍在 METEOR 与 BERTScore 双指标上取得突破性进展, 展现出我国在多模态情感因果推理领域的技术实力。参赛方案普遍采用跨模态注意力机制、动态情感轨迹建模等技术路径, 有效实现了对话文本、语音韵律、微表情等多模态刺激因素的联合分析。值得关注的是, 评测队伍的技术方案仍存在可持续改进之处: 情感因果链的完整度存在瓶颈; 复杂对话场景下的情感推理稳定性不足; 多模态特征对齐精度有待提升。建议后续研究重点突破: 1) 构建细粒度多模态情感事件图谱, 强化刺激因素与认知评价的映射关系建模; 2) 开发动态情感状态追踪框架, 增强长对话场景下的推理鲁棒性; 3) 探索跨模态语义蒸馏技术, 提升异构模态信息的融合质量。本次评测为多模态情感对话领域提供了重要的基准测试体系, 未来可通过引入对抗训练、因果推断等先进范式, 推动多模态情感理解技术向更深层次的认知推理迈进。

赛道二第五届智慧论辩评测以计算论辩技术为核心, 构建了覆盖论辩全流程的立体化评估体系, 推动大模型技术在决策支持、社会仿真等场景的深度应用。从赛题一自主论辩任务 (平均 Acc 0.798) 的表现看, 头部队伍在辩题识别、论据支持判断等基础任务中展现出优异的语义理解能力。赛题二论辩赋能社会调查仿真 (平均 Acc 0.619) 的评测结果揭示了更复杂的技术挑战。现有模型在社会属性建模深度, 尤其是人口统计学特征对辩论策略的驱动的深入融合方面还需开展进一步研究。本次评测为计算论辩技术建立了首个涵盖基础能力验证与社会仿真应用的多层级基准, 未来可通过引入辩论伦理评估维度、构建跨模态论辩数据集 (如融合语音语调、表情信号), 推动该技术在智能决策支持系统中的规模化应用。

赛道三以四格漫画为多模态认知载体, 构建了覆盖逻辑推理与上下文生成的双维度评估体系, 揭示了当前大模型在叙事性多模态理解领域的关键技术瓶颈。基于“起承转结”结构特

征设计的赛题一（逻辑理解）与赛题二（上下文推理）形成互补验证，评测结果显示：多模态叙事理解呈现显著技术分化，赛题一头部队伍在英文与中文任务间存在微弱性能差异，但整体正确率不足 40%，反映出模型对漫画空间布局、视觉符号与叙事逻辑的跨模态对齐能力薄弱。生成式推理展现语言模态优势，赛题二生成任务表现显著优于判别任务，但需注意其依赖文本对齐信息的特性。跨语言能力呈现结构性差异暗示中文模型在叙事文化语境建模（如反转梗、生活化隐喻）方面可能存在数据分布优势，但跨语言迁移能力仍需强化。本次评测首次建立了多模态叙事理解的量化评估框架，为后续研究提供了以下关键启示：需突破现有单模态预训练范式，发展漫画专用多模态架构；应加强叙事逻辑链建模与生成式推理的联合训练；建议构建包含跨文化、跨风格漫画的基准数据集，推动技术向真实应用场景迁移。

赛道四聚焦中文多模态讽刺计算，通过“讽刺识别”与“细粒度分类”双重任务，系统评估模型对跨模态讽刺信息的理解能力。在赛题一（多模态讽刺识别）层面，二分类任务的技术成熟度相对较高；赛题二（细粒度讽刺分类）的技术难度呈现指数级提升，6 分类任务中最高 F1 值与最低的差距达 28.7%。评测结果表明未来研究还应重点聚焦：构建跨模态讽刺知识图谱，整合语言学（反讽结构）、心理学（认知失调理论）、传播学（情感动员机制）等多维度特征；开发时序敏感的跨模态融合模型，重点分析并融合视频时空特征，解决现有模型对微表情时序模式的建模缺失问题。本次评测证实，多模态讽刺计算已从单一模态识别进入“模态协同推理”新阶段，后续赛事应重点增设“领域适应性”（如政务讽刺、网络暴力讽刺）和“实时性”专项评测，加速技术转化进程。

赛道五聚焦中文比拟修辞格的多维度解析，围绕情绪分类、细粒度实体抽取和喻底判别构建多层次评估体系，有效检验了模型对修辞认知逻辑的建模能力。从评测结果看，任务 1（情绪分类）的宏观 F1 均值仅 0.478，任务 2（实体抽取）的 F1 均值 0.481，暴露出当前模型在修辞深层语义关联和细粒度结构解析方面的核心瓶颈。任务 1 的显著性能分化揭示了情感类别判别中的两大挑战：其一，比拟修辞通过喻体映射传递的间接情感易产生语义消歧；其二，情绪标签的细粒度区分对语境依赖性强，需突破传统情感词典的表层匹配局限。任务 2 的实体抽取性能反映出模型在模糊匹配机制上的技术差异。任务 3 的高准确率（均值 0.927%）得益于选项设计的约束性，但需注意测试集可能存在“选项提示效应”，未来需构建开放式喻底生成任务以真实检验认知建模能力。未来相关研究应重点聚焦构建修辞认知图谱，建立本体-喻体-喻底-情感的多元关联网络；开发动态语境感知的修辞解析算法，解决“一语多喻”场景下的指派歧义。赛道五的评估体系为中文修辞智能研究提供了重要基准，其揭示的技术挑战将推动计算语言学与认知科学的深度交叉创新。

由于评测时间比较仓促，在评测过程中也出现了一些错误，经过和参赛队伍的不断讨论，也进行了及时的修正；同时，对于评测方法的制定、评测指标的选择、人工标注的质量等都存在值得商榷的地方，需要进一步分析和完善。我们希望通过这些评测赛道，建立并完善情感计算研究的基础数据集和评测标准，推动该领域的研究进展和技术实用化，活跃学术氛围，促进各个研究单位的交流，为未来的评测组织工作提供经验借鉴与参考，为情感计算领域的技术进步提供有力支持。

## 5 致谢

感谢 CCAC2025 程序委员会主席夏睿教授、杨巨峰教授、赵妍妍教授、组织委员会主席李显勇教授、陈晓亮教授、杜亚军教授对本届评测的协调与支持，感谢山西大学王素格教授对评测组织各环节的指导与帮助。感谢各赛道组织团队在语料标注、赛道组织和评测实施过程中的艰苦工作和重要贡献，感谢所有支持和参与本次评测的国内外同仁。

## 附录：获奖队伍技术报告

赛道一：

1. Emotion Change Reasoning in Multimodal Interaction – CCAC2025 Challenge
2. 多模态交互中的情感变化推理技术报告
3. MECR\_CCAC2025：基于 LoRA 微调的多模态情绪变化推理评测方案
4. CCAC2025-MECR 比赛技术报告
5. 面向多模态对话的可解释情绪因果链生成研究
6. 基于多模型提示词工程的多模态交互中的情感变化推理

赛道二：

1. CCAC2025 评测技术报告:赛道 2-task1-自主论辩任务
2. 第五届智慧论辩评测（AI-Debater 2025）辩题识别
3. 第五届智慧论辩评测（AI-Debater 2025）论据挖掘
4. 第五届智慧论辩评测（AI-Debater 2025）辩论构建生成技术报告
5. 第五届智慧论辩评测（AI-Debater 2025）自主辩论生成技术报告
6. CCAC2025 评测技术报告
7. 融合分类与生成技术的中文论辩系统构建
8. 智慧论辩评测 wzkk 队伍技术报告

赛道三：

1. 四格漫画逻辑理解
2. CCAC2025-task3 技术报告：四格漫画理解能力评测

赛道四：

1. 融合幽默-讽刺语义的多模态多任务情感识别方法
2. CCAC2025 赛道四技术报告：基于 Qwen2-Audio 的中文多模态讽刺计算技术研究
3. 基于数据增强与模型集成的中文多模态讽刺计算
4. 中文多模态讽刺识别评测方案
5. 基于预训练模型与 Transformer 融合机制的中文多模态讽刺计算
6. 面向中文多模态讽刺检测的分层多视角融合方法

赛道五：

1. CCAC2025 赛道五 IIE06 队伍技术报告
2. CCAC2025-Task5 技术报告：面向细粒度比拟句情绪分析的上下文微调方法
3. 基于预训练模型的细粒度比拟句情感分析研究
4. CCAC2025-task5 技术报告：细粒度比拟句喻底判别
5. 细粒度情感分析技术简报

# Emotion Change Reasoning in Multimodal Interaction – CCAC2025 Challenge

Li Dayu<sup>1</sup> Li Yang<sup>2</sup> Chen Xin<sup>3</sup> Zhang Wenyue<sup>4</sup>

<sup>1</sup>School of Finance, Shanxi University of Finance and Economics, Taiyuan 030000, China

<sup>2</sup>School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030000, China

<sup>3</sup>School of Information, Shanxi University of Finance and Economics, Taiyuan 030000, China

{corresponding\_emaildomain}

## Abstract

This paper addresses the core challenge of the CCAC 2025 evaluation task "Emotion Change Reasoning in Multimodal Interaction"—deep understanding of the causes and processes of emotional changes. We propose an innovative framework based on Multi-Task Quantized Low-Rank Adaptation (MT-QLoRA). By constructing a unified instruction template to integrate four closely related subtasks (textual stimulus recognition, visual stimulus extraction, cognitive appraisal generation, and emotion response prediction), we effectively model the "stimulus → cognition → emotion" causal reasoning chain. Furthermore, we implement efficient parameter fine-tuning on the Qwen3-32B model using Quantized Low-Rank Adaptation technology. Experiments on the INSIDE dataset from CCAC 2025 (containing 4,147 annotated multi-modal dyadic conversation samples with stimulus-cognition-emotion triples) demonstrate that: 1) The multi-task collaborative mechanism significantly enhances semantic relevance; 2) Low-rank decomposition combined with quantization strategies effectively reduces GPU memory requirements and improves inference speed. Our solution ultimately outperforms baseline models with METEOR 0.5024 and BERTScore 0.81, validating its effectiveness in accurately modeling causal chains of emotional changes and providing an efficient new paradigm for multimodal emotion change reasoning in resource-constrained scenarios.

## 1 Introduction

With the advancement of AI, affective computing has gained significant attention as a critical branch of NLP and social computing. In multimodal interaction scenarios, understanding human emotional dynamics is increasingly vital. Traditional multimodal emotion recognition typically classifies static emotional states but overlooks the dynamic evolution of emotions and their underlying causality. This limitation is particularly evident in applications like human-computer dialogue, online tutoring, and medical consultations, restricting systems' ability to understand and respond to users' emotional fluctuations.

To tackle these challenges, the Fifth China Conference on Affective Computing (CCAC 2025) established a dedicated evaluation track for "Emotion Change Reasoning in Multimodal Interaction." As shown in Figure 1, this task models the complete stimulus → cognition → emotion causal chain through multimodal analysis (text/vision/audio), enhancing HCI empathy while providing practical value for education and healthcare. Current research still faces three key challenges: (1) Difficulty in extracting key stimulus features due to heterogeneous multimodal signals; (2) Implicit expression of cognitive appraisals complicating semantic mapping; (3) Prohibitive resource costs for full fine-tuning of large models, limiting practical deployment.

To address these issues, this paper proposes an emotion change reasoning framework based on Multi-Task Quantized Low-Rank Adaptation (MT-QLoRA), aiming to achieve precise modeling of emotional causal chains in multimodal conversations. By constructing a unified structured instruction template, our method integrates four subtasks—textual stimulus recognition, visual feature extraction, cognitive appraisal generation, and emotion response prediction—into an end-to-end generation pipeline, forming a complete "stimulus → cognition → emotion" reasoning pathway. Simultaneously, through 4-bit quantization and low-rank adaptation techniques, we achieve efficient parameter fine-tuning on the Qwen3-32B large model, reducing GPU memory consumption to 21GB while increasing inference speed by 2.1×.

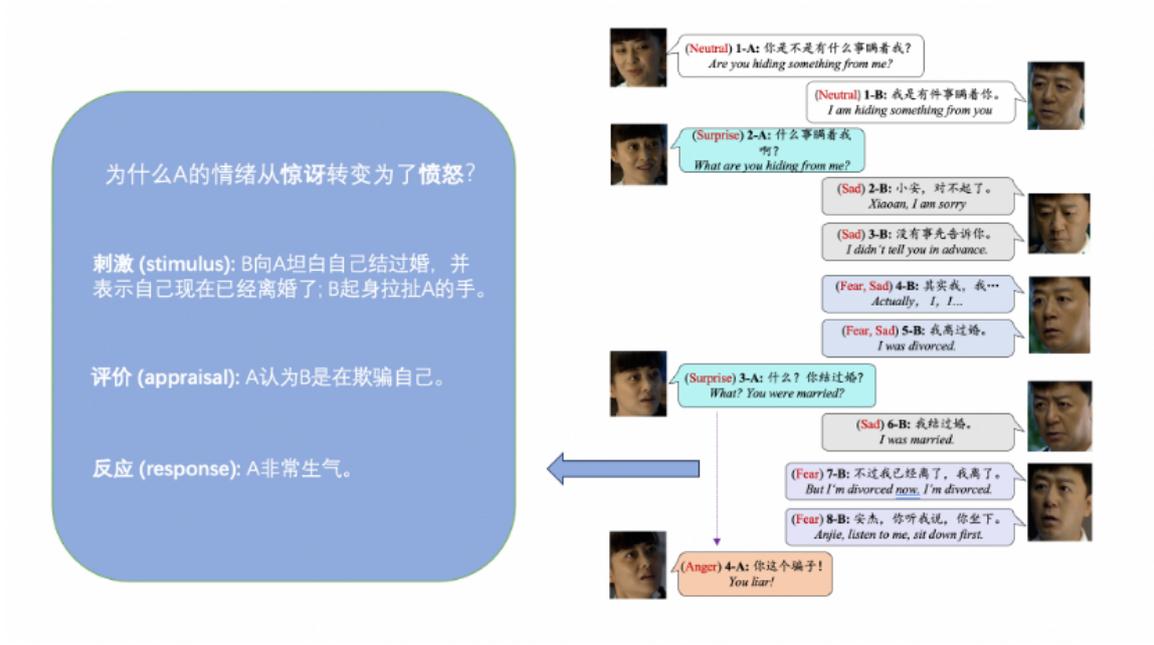


Figure 1: Example of emotion change reasoning task in multimodal interaction. The figure shows a dyadic multi-turn dialogue segment where the target task is to reason how Speaker A’s emotion transitions from surprise to anger. The reasoning process requires identifying: *Stimulus*: B confesses to A about previous marriage and divorce; B stands up and grabs A’s hand. *Cognitive Appraisal*: A believes B is deceiving them. *Emotional Response*: A feels intense anger.

On the INSIDE dataset used in the CCAC 2025 evaluation task, our method achieved significant results: a METEOR score of 0.5024 and BERTScore of 0.81, substantially outperforming baseline models and securing first place in the overall ranking. Experimental results demonstrate that the multi-task collaborative mechanism effectively enhances the model’s understanding of multimodal information, while the quantization and low-rank strategies provide viable pathways for efficient deployment in resource-constrained environments. This research not only validates the effectiveness of the MT-QLoRA framework in emotion change reasoning tasks but also offers new insights and technical support for future studies in multimodal affective computing.

## 2 Methodology

To address the challenges inherent in multimodal emotion shift reasoning, we propose a Multi-Task Quantized Low-Rank Adaptation (MT-QLoRA) architecture. This framework features two fundamental innovations: First, it integrates closely related subtasks into an end-to-end generative pipeline through a unified structured instruction template, thereby simulating the human cognitive process of “Stimulus → Appraisal → Emotion” reasoning. Second, it employs Quantized Low-Rank Adaptation to maintain model performance while substantially reducing computational resource requirements, significantly enhancing deployability in resource-constrained environments. The overall framework comprises three core components: Multimodal Dialogue Encoding, Multi-Task Structured Instruction Construction, and Parameter-Efficient Tuning & Inference.

### 2.1 Multimodal Dialogue Information Encoding

We utilize the Qwen3-32B pre-trained large language model as the foundational backbone for encoding utterance sequences. Given an input sequence  $\mathcal{T} = \{(s_i, e_i, t_i)\}_{i=1}^N$  where  $s_i$  denotes speaker identity,  $e_i$  represents emotion labels, and  $t_i$  corresponds to raw text content, the model generates text representations through the transformation:

$$\mathbf{H}^t = \text{Qwen3-32B}(\mathcal{T}) \in \mathbb{R}^{N \times d} \quad (1)$$

This encoding process captures both linguistic features and emotional cues within the dialogue flow.

## 2.2 Multi-Task Structured Instruction Construction

To facilitate the generation of structured emotion shift quadruples  $\mathcal{R} = (S_{\text{text}}, S_{\text{visual}}, A, R)$ , we design a specialized instruction template with explicit task guidance:

$$\mathcal{I} = \underbrace{[\text{SYS}] \text{Analyze \#speaker's emotion shift from \#prev\_emo to \#curr\_emo in the dialogue:}}_{\text{Task Description}} \oplus \mathcal{T} \oplus \underbrace{[\text{REQ}] \text{Output: <stimulus-text>; <stimulus-visual>. <appraisal>. <response>}}_{\text{Format Constraint}}.$$

This template directs model attention to speaker  $p$ 's emotion transition  $\text{emo}_{k-1} \rightarrow \text{emo}_k$  at position  $k$ , with the [REQ] marker initiating multi-task generation under strict output constraints.

## 2.3 Parameter-Efficient Tuning and Inference

Our parameter-efficient fine-tuning approach implements Quantized Low-Rank Adaptation (QLoRA) on the Qwen3-32B foundation model. This involves dual optimization strategies: 4-bit quantization compresses original floating-point weight matrices  $\mathbf{W} \in \mathbb{R}^{m \times n}$  into 4-bit integers, dramatically reducing memory footprint, while low-rank adaptation injects trainable parameters through the decomposition  $\Delta \mathbf{W} = \mathbf{A} \mathbf{B}^\top$  where  $\mathbf{A} \in \mathbb{R}^{m \times r}$  and  $\mathbf{B} \in \mathbb{R}^{n \times r}$  with  $r \ll \min(m, n)$ . The model subsequently generates outputs via autoregressive decoding:

$$\hat{\mathcal{R}} = \text{Decoder}_{\theta_{\text{Qwen}}}(\mathcal{I}; \mathbf{H}; \Delta \mathbf{W}) \quad (2)$$

producing four essential elements: textual stimuli  $S_{\text{text}}$  (emotion-triggering utterances), visual stimuli  $S_{\text{visual}}$  (expressions/gestures or null), cognitive appraisal  $A$  (psychological interpretation), and emotional response  $R$  (e.g., "p feels angry").

## 2.4 Multi-Task Joint Loss Function

To strengthen synergistic understanding across quadruple components, we formulate a multi-task joint loss function:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{CE}}^{(\text{text})} + \lambda_2 \mathcal{L}_{\text{CE}}^{(\text{visual})} + \lambda_3 \mathcal{L}_{\text{CE}}^{(\text{appraisal})} + \lambda_4 \mathcal{L}_{\text{CE}}^{(\text{response})} \quad (3)$$

Here,  $\mathcal{L}_{\text{CE}}^{(\text{text})}$  and  $\mathcal{L}_{\text{CE}}^{(\text{response})}$  denote classification cross-entropy losses,  $\mathcal{L}_{\text{CE}}^{(\text{visual})}$  handles visual stimulus reconstruction,  $\mathcal{L}_{\text{CE}}^{(\text{appraisal})}$  models appraisal semantic consistency, and  $\lambda_i$  represent task-specific weighting coefficients that balance component contributions.

## 3 Experiments

All experiments were conducted on **4×NVIDIA A100 80GB GPUs** using PyTorch 2.5.1, Unsloth 2025.6.12, and Transformers 4.51.3. Key training parameters include: maximum sequence length of 2,048 tokens, batch size of 8, learning rate of  $2e-4$  with AdamW optimizer. For parameter-efficient fine-tuning, we integrated LoRA with 4-bit NF4 quantization to enhance training efficiency and reduce resource consumption.

To validate the effectiveness of our proposed MT-QLoRA framework for multimodal emotion shift reasoning, we conducted systematic experiments on the INSIDE benchmark dataset. Table 1 compares performance metrics (METEOR and BERTScore) across baseline models and training strategies.

The prompt-based approach using Qwen2.5-7B-Instruct achieved only 0.0631 METEOR and 0.5728 BERTScore under zero-shot conditions, indicating small models struggle to capture complex causal chains without guidance. Performance significantly improved to 0.2921 (one-shot) and 0.3043 (two-shot) METEOR with in-context learning, demonstrating the value of few-shot examples for task understanding.

Table 1: Performance comparison on INSIDE dataset

| Method   | Model                     | Setting           | METEOR        | BERTScore     |
|--|---------------------------|-------------------|---------------|---------------|
| <i>Prompt-based Generation</i>                 |                           |                   |               |               |
| Prompt-based                                   | Qwen2.5-7B-Instruct       | Zero-shot         | 0.0631        | 0.5728        |
| Prompt-based                                   | Qwen2.5-7B-Instruct       | One-shot          | 0.2921        | 0.7138        |
| Prompt-based                                   | Qwen2.5-7B-Instruct       | Two-shot          | 0.3043        | 0.7136        |
| <i>Quantized Supervised Fine-Tuning (QSFT)</i> |                           |                   |               |               |
| QSFT   | DeepSeek-R1-0528-Qwen3-8B | Joint (Epoch 1)   | 0.4781        | 0.7131        |
| QSFT   | DeepSeek-R1-0528-Qwen3-8B | Joint (Epoch 2)   | 0.4799        | 0.7085        |
| QSFT   | Qwen3-8B                  | Joint (Epoch 1)   | 0.5154        | 0.7538        |
| QSFT   | Qwen3-8B                  | Joint (Epoch 2)   | 0.5241        | 0.7550        |
| QSFT   | Qwen3-8B                  | Joint (Epoch 3)   | 0.5217        | 0.7525        |
| QSFT   | Qwen3-32B                 | Joint (Epoch 1)   | 0.5320        | 0.8064        |
| QSFT   | Qwen3-32B                 | Joint (Epoch 2)   | <b>0.5451</b> | <b>0.8092</b> |
| QSFT   | Qwen3-32B                 | Joint (Epoch 3)   | 0.5463        | 0.8081        |
| <i>Subtask Decoupled Training</i>              |                           |                   |               |               |
| QSFT   | Qwen3-8B                  | Textual Stimulus  | 0.2421        | 0.5844        |
| QSFT   | Qwen3-8B                  | Appraisal         | 0.2302        | 0.6062        |
| QSFT   | Qwen3-8B                  | Response          | 0.1661        | 0.5868        |
| QSFT   | Qwen3-8B                  | Combined Subtasks | 0.3023        | 0.6897        |

Instruction fine-tuning with QLoRA substantially enhanced performance. Qwen3-8B reached 0.5241 METEOR and 0.7550 BERTScore after two training epochs. The larger Qwen3-32B model exhibited superior modeling capabilities, achieving 0.532 METEOR/0.8064 BERTScore after just one epoch, and peaking at 0.5451 METEOR/0.8092 BERTScore in the second epoch. This confirms that larger foundation models possess stronger generalization and semantic understanding for emotion shift reasoning.

Subtask-decoupled training (separately optimizing textual stimulus, appraisal, and response components) yielded substantially lower performance (0.3023 METEOR/0.6897 BERTScore) compared to joint training. This validates that multi-task collaborative learning enhances holistic modeling of the "stimulus→appraisal→emotion" causal chain.

## 4 Conclusion

This paper proposes MT-QLoRA, a novel multimodal emotion shift reasoning framework that integrates four subtasks—textual stimulus identification, visual stimulus extraction, cognitive appraisal generation, and emotional response prediction—into a unified causal chain through structured instruction templates. By combining quantization and low-rank adaptation, we enable efficient fine-tuning of large pretrained models (Qwen3-32B). Experimental results demonstrate state-of-the-art performance on the CCAC 2025 INSIDE benchmark, achieving 0.5463 METEOR and 0.8092 BERTScore. Our parameter-efficient approach significantly reduces resource requirements while maintaining performance. Future work will focus on: (1) optimizing multimodal fusion mechanisms, (2) incorporating external knowledge, and (3) developing finer-grained emotion state modeling to enhance understanding of dynamic affective changes in human interactions.

# 多模态交互中的情感变化推理技术报告

蒋培元<sup>1</sup> 张瀚文<sup>1</sup> 张宗舜<sup>1</sup> 彭雪枫<sup>1</sup> 杨嘉业<sup>1</sup> 刘露<sup>1</sup>

何义会<sup>1</sup> 刘峤<sup>1</sup> 刘瑶<sup>2</sup>

(1. 电子科技大学 计算机科学与工程学院, 四川省成都市 邮编 611731;

2. 电子科技大学 信息与软件工程学院, 四川省成都市 邮编 611731)

**摘要:** 本报告围绕多模态交互中的情感变化推理任务展开, 探索了多种方案以提升推理效果。首先, 探究了大语言模型微调方法, 通过设计 prompt 实现了对话情境中情感变化的推断。其次, 我们探究了基于多模态的情感推理方法。具体来说, 基于融合方案的不同, 我们的方案可以划分为两种: Step-by-Step 方案与端到端方案。其中, Step-by-Step 方案通过视觉大模型生成视频内容描述文本, 再与原始对话文本联合建模, 实现情感推理; 端到端模式则直接将视频编码映射至文本空间, 并与文本 token 联合解码生成推理结果。此外, 在端到端方法中, 我们提出了一种关键帧选择策略以提升多模态大模型的长视频理解能力。在验证集上, 我们的最优方法取得了 BERTScore 0.743 和 METEOR 0.475 的性能表现。

**关键词:** 情感推理; 关键帧提取; 多模态融合

中图分类号: TP391

文献标识码: A

## Emotion Shift Reasoning in Multimodal Interaction: Technical Report

Peiyuan Jiang<sup>1</sup>, Hanwen Zhang<sup>1</sup>, Zongshun Zhang<sup>1</sup>, Xuefeng Peng<sup>1</sup>, Jiaye Yang<sup>1</sup>, Lu Liu<sup>1</sup>,

Yihui He<sup>1</sup>, Qiao Liu<sup>1</sup>, Yao Liu<sup>2</sup>

(1. School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China;

2. School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China)

**Abstract :** This report focuses on the task of emotion shift reasoning in multimodal interactions and explores multiple approaches to improve inference performance. Firstly, we investigate fine-tuning strategies for large language models (LLMs), where emotion shifts in dialogue contexts are inferred through carefully designed prompts. Secondly, we explore multimodal emotion reasoning methods. Specifically, based on different fusion strategies, our approaches are categorized into two types: Step-by-Step and end-to-end. In the Step-by-Step approach, a visual-language model is used to generate textual descriptions of video content, which are then combined with the original dialogue text for joint reasoning. In the end-to-end approach, video features are directly mapped into the text space and decoded jointly with text tokens to produce inference results. Moreover, in the end-to-end method, we propose a key frame selection strategy to enhance the ability of multimodal large models to understand long videos. On the validation set, our best-performing method achieves a BERTScore of 0.743 and a METEOR score of 0.475.

**Key words:** Emotion Reasoning; Key Frame Extraction; Multimodal Fusion

## 0 引言

在多模态人机交互场景中, 准确理解和推断视频中说话人的情感变化对于提升人工智能系统的

共情能力至关重要<sup>[1]</sup>。实际对话中, 说话主体的情绪状态是复杂多变的, 即说话主体的情绪状态常随着对话的演进而产生变化。如图 1 所示, 说话人 A 的情绪最初是中立, 但在交谈过程中受到另一名对话者影响而转变为愤怒。因此, 除了正确

识别说话者的情绪状态外，推理出引起情绪状态转变的原因也是实现智能人机交互的关键一步<sup>[2]</sup>。然而，大多数现有对话情感识别方法主要关注情感的变化结果，忽略了情感转变原因，这使得难以理解模型的内在判别逻辑。因此，如何全面建模对话中的情感变化因果关系是情感计算中的一个重要人物。



图 1: 对话中的情感转变

当前，情感转换因果推理面临的主要挑战在于如何有效整合多模态情感线索实现鲁棒的因果推理<sup>[3]</sup>。传统单模态方法在面对真实世界丰富多样的情感表达时经常因视角有限导致其鲁棒性和精度受到局限。已有研究表明，将视觉、语音和文本信号相结合可以提升对话情感识别的性能。

为此，本文围绕“多模态交互中的情感变化推理”任务，设计了面向该问题的模型方案。首先，我们探究了模型在仅使用单模态（仅使用文本数据）数据时的因果推理能力。其次，我们在此单模态基础上进一步探究了引入视觉模态后模型的性能表现。具体来说，我们探索了两种多模态融合推理方案：

(1) **Step-by-Step 方案**：借助预训练的视觉大语言模型对视频内容进行描述生成，将得到的关键画面语义信息转换为文本，再与原始对话文本融合输入情感推理模型。

(2) **端到端方案**：构建统一的多模态模型，直接将视频帧与音频编码映射到与文本相同的表示空间，并与对话文本的 token 一同输入模型解码层。通过端到端训练，使模型直接学习从视频和文本的联合表示中生成情感推理结果，实现视觉-文本信息的深度融合。

在方案(2)中，为了确保大模型能够有效理解长视频信息，我们需要对视频中的关键帧（存在情感语义信息转换的帧）进行预抽取以过滤掉视频中的冗余信息。

## 1 技术方案

### 1.1 数据分析

(1) 考虑到视频是时序数据，转折处的情感

线索必然在情感转折之前发现，因此需要对输入视频进行裁剪，避免转折后的视频段信息干扰；

(2) 数据中情感类别分布存在不均衡，如图 2 所示，需要在 prompt 中引入一些先验信息；

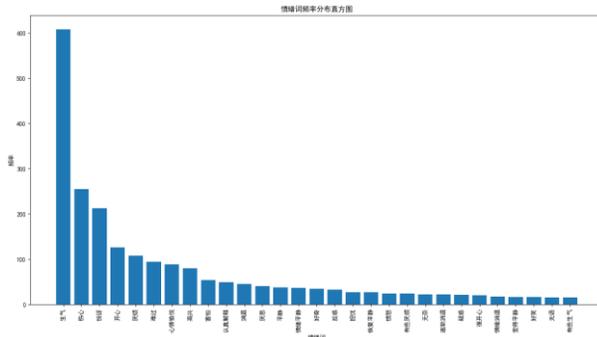


图 2: 情感类别标签分布统计

(3) 数据中的视觉线索稀疏，训练数据中视觉线索标签数据统计分布如图 3 所示，模型如果过于关注视觉线索可能导致整体的效果变差，需要均衡线索贡献。

Distribution of Visual Field Availability in Training Set

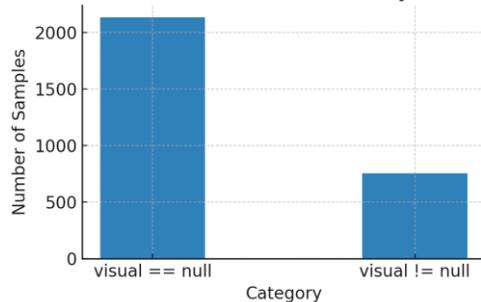


图 3: 训练集视觉线索标签分布

### 1.2 基于文本的大语言模型微调

基础语言模型采用 Qwen2.5-7B 和 Qwen3-14B。旨在探究不同参数量的大语言模型对于处理复杂对话文本情感分析任务的能力。通过针对情感推理任务的微调，优化模型在情感状态识别、文本线索挖掘及因果推理中的表现，确保结果的精准性和鲁棒性。

为实现精准的情感因果推理，我们设计了如下结构化 **Prompt** 模板：

**“instruction”**：“你是多模态情绪因果推理专家。现给你一段对话文本，你需要完成以下任务：

(1) 根据对话中最后一句话的情绪状态如生气、伤心、厌烦、惊讶、开心、难过、平静、高兴、害怕、厌恶、反感、愉悦、无奈、期待、好奇及其组合等 {responses}，从输入中找出导致该情绪的文本刺激线索 (textual stimulus)；

(2) 用一句话概括基于上下文信息所引起的情感转变的原因 (appraisal)。

(3) 用一句话描述当事人可能的应对反应 (response)，如“说话人 (A or B) 感到反感和

气愤\”。

(4) 你必须严格以下面格式输出你的答案：  
{textual stimulus}。{appraisal}。{response}。”

表 1: 基于大语言模型微调的测试结果

| Model      | bertscore | meteor |
|------------|-----------|--------|
| Qwen2.5-7B | 0.559     | 0.487  |
| Qwen3-14 B | 0.589     | 0.509  |

实验结果表明，相同 prompt 情况下，模型参数量越大，情感因果推理能力越强，这与实际的直觉相符。

### 1.3 基于视觉文本的大模型微调方案 (Step-by-Step)

我们借助 Tarsier2 视觉语言基础模型<sup>[4]</sup>与 Qwen2.5-7B, Qwen3-14B 语言模型，先用视觉语言 Prompt 详细描述视频中人物的细微表情与动作，再结合语言 Prompt，让模型融合对话文本与视觉描述，精准识别最后一句话的情感状态和触发线索，完整输出情感转变的因果链条，为多模态情感推理提供丰富且具象的视觉文本双重支撑，同时利用 opus-mt-en-zh 翻译模型实现跨语言推理表达。具体细节如下所示：

视觉语言 Prompt: "Please describe in detail the movements and facial expressions of the characters appearing in this video. It must be detailed"

语言 Prompt: "instruction": "你是一位严谨的多模态情绪分析专家，现已获得：\n完整对话文本（见 input）。\n对应视频的视觉描述（见 video\_description）。\n请在文本线索与视觉线索之间寻找引起最后一句话情感状态的线索，完成情感因果推理，具体步骤如下：

(1) 根据最后一句话的情感状态 {responses}；从上下文中找出引发这个情感状态的 textual stimulus、visual stimulus（无则填“null”）。

(2) 基于线索，输出产生该情感的原因，即 appraisal\n

(3) 最终的输出格式为：\n{textual stimulus}；{visual stimulus}。{appraisal}。{response}。 ”。

表 2: 视觉大模型微调测试结果

| Model               | bertscore | meteor |
|---------------------|-----------|--------|
| Tarsier2+Qwen2.5-7B | 0.63      | 0.473  |
| Tarsier2+Qwen3-14B  | 0.67      | 0.511  |

从表 2 的结果来看，引入额外的场景描述信息可以有效提升模型推理的语义准确率，对于 meteor 得分的帮助不大。这主要是因为 METEOR 主要基于单词级别的精确匹配、词干匹配和有限的同义词映射来计算指标，它对句子整体的逻辑连贯性和语义一致性缺乏敏感度。引入额外的场景描述，有助于模型捕获背景信息、推断隐藏含义，从而提升语义准确率；但这些深层次的推理结果，

往往不会直接反映为与参考答案词汇上的 n-gram 重合度的提升。

### 1.4 基于视觉文本的大模型微调方案 (端到端模式)

在端到端模式中，我们以 CLIP 为视觉编码器、HuBERT 为音频编码器、Qwen2.5-7B Tokenizer 为文本编码器、Qwen2.5-7B 为基础语言模型，利用模态桥接网络将视觉和音频特征映射到文本模态空间，结合对话文本与视觉信息，直接生成符合要求的情感推理结果，高效实现多模态情感变化的精准判断与分析。具体细节如下所示：

模态桥接网络：视觉模态和音频模态各自使用一个 nn.Linear 将特征投影到文本模态的维度空间。

Prompt: 你是一位“多模态情绪因果推理专家”，擅长结合对话文本、视觉帧和音频片段，进行情感刺激识别与因果分析。

给定三部分输入：

- (1) 对话文本；
- (2) 对应的视觉特征；
- (3) 对应的音频特征。

请完成以下四项任务：

- 1) Textual Stimulus: 从最后一句对话中，找出引发当前情绪的主要文本刺激；
- 2) Visual Stimulus: 根据 visual\_features, 推理引发该情绪的视觉刺激；若无，则填“null”；
- 3) Appraisal: 用一句话概括对话上下文和多模态信息所导致的情感转变原因；
- 4) Response: 描述当事人的情绪反应，如“说话人 A 感到惊讶与失望”

输出格式（严禁多余说明与换行）：

(textual stimulus); (visual stimulus)。 (appraisal)。 (response)。

### 1.5 优化策略-关键帧选择策略

通过观察我们发现，视频中的对话过程中往往同时出现多位说话者，且发言频繁切换。针对这种情况，我们提出了以下关键帧提取方案的四个阶段：

(1) 转场检测&粗分段：利用整帧颜色直方图变化或(有音频时)音频能量突变检测转场点，得到一系列分段边界；每两个相邻边界之间，视作为一个“相对连贯”的短段。

(2) 片段内 AU 峰值 Candidate: 对每段中每帧提取面部动作单元 (AU) 激活强度，或仅用嘴部开合度等关键 AU；用 Savitzky - Golay 滤波平滑 AU 曲线，再用 find\_peaks 找 1 - 2 个局部峰值帧，作为该段的候选代表帧。

(3) 多样性&全局配额筛选：汇总所有候选帧，若总量>目标帧数 N，则按“段长”或“峰值高度”为每段分配配额，并在本段候选中用 K-Medoids 聚类选出枢轴帧；若候选不足，则补充段中点或色彩变化最大帧；去重、按时间排序，截取最终 N（预先定义，本项目中默认值为 8）帧。

(4) 人脸框绘制&输出：对每个最终选中的关键帧，使用高置信度的人脸检测器（如 RetinaFace/feat.Detector）重新检测面部位置；在帧图像上用红框（或其他高对比色）标出脸部区域，帮助下游模型聚焦。

表 3：多模态大模型（端到端）微调测试结果

| Model                         | bertscore | meteor |
|-------------------------------|-----------|--------|
| Emogpt-w/o key frame extract  | 0.733     | 0.466  |
| Emogpt-with key frame extract | 0.743     | 0.475  |

其中，Emogpt 是我们所提出的端到端模型的名称，w/o key frame extract 表示没有使用关键帧抽取策略，with key frame extract 表示使用了关键帧抽取策略。由表 3 可知，基于端到端的策略因引入了更多的模态线索，可以获得更高的 bertscore。其中，相较于随机采样策略，我们的关键帧提取策略可以降低信息冗余，补充有效的视觉线索，进一步提升模型生成描述的语义准确率。最终，基于端到端的方案可以获得更高的性能。

## 参考文献

- [1] Shen G, Wang X, Duan X, et al. Memor: A dataset for multimodal emotion reasoning in videos[C]//Proceedings of the 28th ACM international conference on multimedia. 2020: 493-502.
- [2] Yang J, Wei P, Du X, et al. Graph attention based on contextual reasoning and emotion-shift awareness for emotion recognition in conversations[J]. Complex & Intelligent Systems, 2025, 11(7): 1-20.
- [3] Cheng Z, Cheng Z Q, He J Y, et al. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning[J]. Advances in Neural Information Processing Systems, 2024, 37: 110805-110853.
- [4] Yuan L, Wang J, Sun H, et al. Tarsier2: Advancing Large Vision-Language Models from Detailed Video Description to Comprehensive Video Understanding[J]. arXiv preprint arXiv:2501.07888, 2025.

# MECR\_CCAC2025: 基于LoRA微调的多模态情绪变化推理评测方案

张子蕴

深圳北理莫斯科大学

zhangziyun@bit.edu.cn

## 摘要

本文针对CCAC2025多模态情感变化推理评测 (MECR) 任务, 提出结合结构化Prompt与LoRA微调的因果推理方案。该任务需建模情绪变化的刺激 (Stimulus)、认知评价 (Appraisal) 及反应 (Response) 三要素。研究通过数据清洗与结构化处理构建训练集, 设计包含完整字幕信息与情绪变化标注的Prompt模板, 引导模型生成结构化JSON输出。基于Qwen2.5-VL-7B-Instruct模型, 采用四阶段LoRA微调策略 (共18轮), 显著提升模型对情绪因果链的推理能力。最终方案在测试集上取得METEOR 0.4701和BERTScore 0.7927, 排名赛事第三, 验证了技术路线的有效性。

**关键词:** 多模态; 情感推理; 大模型微调

## MECR\_CCAC2025: A Multimodal Emotional Change Reasoning Evaluation Scheme Based on LoRA Fine-Tuning

Ziyun Zhang

Shenzhen MSU-BIT University

zhangziyun@bit.edu.cn

## Abstract

This paper focuses on the Multimodal Emotional Change Reasoning (MECR) task of CCAC2025 and proposes a causal reasoning scheme combining structured Prompt and LoRA fine-tuning. This task requires modeling three elements of emotional changes: Stimulus, cognitive Appraisal, and Response. The study constructs the training set through data cleaning and structured processing, designs a Prompt template containing complete subtitle information and emotional change annotations, and guides the model to generate structured JSON outputs. Based on the Qwen2.5-VL-7B-Instruct model, a four-stage LoRA fine-tuning strategy (with a total of 18 epochs) is adopted, which significantly improves the model's ability to reason about emotional causal chains. The final scheme achieves METEOR 0.4701 and BERTScore 0.7927 on the test set, ranking third in the competition, verifying the effectiveness of the technical route.

**Keywords:** multimodal, emotional reasoning, large model fine-tuning

## 1 引言

近年来，多模态情感推理已成为人工智能领域的一个研究热点。它旨在通过融合文本、视觉、音频等多种模态信息，对人类在真实交互场景中的情绪变化进行深度感知和精准推理。这一技术在虚拟心理咨询、智能助手、人机交互和社交机器人等应用场景中具有重要意义(Amirizani et al., 2024)。

CCAC2025 的多模态情感变化推理评测 (MECR, Multimodal Emotion Change Reasoning) 正是在这一研究背景下提出。该赛道强调模型需要理解视频交互中的情境、捕捉情绪主体的情绪变化，并结合多模态信息输出符合人类推理逻辑的分析结果。与传统情感识别任务不同，MECR 不仅要求识别当前的情绪状态，还需要推理导致情绪变化的“刺激”(stimulus)、主体的“认知评价”(appraisal)及其“情绪反应”(response)。这一任务本质上是情绪变化的因果推理问题，挑战模型在复杂语境下的多模态对齐、情绪理解及因果链建模能力(Cheng et al., 2024)。

在数据层面，MECR 数据集覆盖了多种日常生活和社会场景，视频中包含多名发言人及其情绪标注。每段视频配有精细化标注，包括起止时间、说话人身份、情绪类别和情绪变化信息。这使得模型在处理过程中需要同时考虑时间序列信息、跨模态语义一致性及复杂的情绪演变过程。

本项目基于LLaMA-Factory框架(Zheng et al., 2024)，采用LoRA (Low-Rank Adaptation) 微调技术对Qwen2.5-VL-7B-Instruct大模型进行适配。我们设计了完整的数据预处理、异常样本清理、Prompt工程及多阶段微调策略，以提升模型对视频字幕和视觉信息的情绪推理能力。最终在测试集上取得 METEOR 0.4701 和 BERTScore 0.7927 的成绩，在参赛队伍中排名第三。

本文将系统介绍本项目在CCAC2025 MECR赛道中的技术方案，包括数据处理方法、模型微调策略、实验设计及结果分析，并对未来的优化方向进行探讨。

## 2 技术路线

在本研究中，我们围绕 CCAC2025 MECR 赛道的特点，设计了一条符合技术路线报告要求的完整技术路径。该路径涵盖数据处理、Prompt 模板设计、模型微调以及评测步骤。整体设计不仅满足任务需求，更强化了在多模态交互场景下情绪因果推理的能力。

### 2.1 数据处理与 Prompt 模板设计

MECR数据集由多模态视频、字幕及情绪变化标注组成，数据复杂度高，且存在部分异常样本，如“jiayouernv”目录下的视频元数据损坏无法被PyAV解析。针对这一问题，我们开发了“del\_jiayouernv.py”脚本，自动检测并剔除这些异常样本，确保后续模型训练数据的完整性和一致性。

数据格式转换是预处理中的关键环节。通过“sharegpt\_transfer.py”脚本，我们将原始JSON格式数据转化为适配Qwen2.5-VL-7B-Instruct的ShareGPT格式。在此过程中，我们特别加入了情绪主体、变化前后语句、时间戳等关键信息，为模型提供了完整、结构化的上下文输入。

Prompt模板设计是本技术路线的核心。初期实验表明，传统对话格式难以引导模型生成符合任务要求的情绪因果链推理结果。原始数据结构中，视频字幕和情绪标注分别以“utterance”、“emotion”、“timestamps”呈现，变化信息通过“target\_change\_utt\_ids”和“target\_change\_emos”记录。为了让模型充分理解情绪主体的变化过程及其原因，我们重新设计了 Prompt，使其包含以下元素：

- 完整的字幕信息，明确标注说话人、起止时间及情绪类别。
- 情绪变化标注，以“主体（说话人X）从[情绪1]变为[情绪2]”的方式呈现，列出变化前后语句序号。

3. 明确的任务说明，要求模型完成主体识别、围绕主体进行因果推理，并输出结构化 JSON。
4. 严格格式要求，约束模型输出 `stimulus`、`appraisal` 和 `response` 三个字段。

设计后的 Prompt 样例如下：

```

请根据以下视频字幕内容推理出情境 rationale，并以 JSON 格式输出：
### 视频字幕内容

[1] (说话人: A) 起止时间: [00:00:00:01-00:00:02:19] 情绪: ['Neutral'] 台词: 你先把衣服换了，然后去这个地址。

[2] (说话人: A) 起止时间: [00:00:02:19-00:00:07:20] 情绪: ['Neutral'] 台词: 一会儿会有装修公司的人过去，你去监工，看看有什么可以帮忙的。

[3] (说话人: B) 起止时间: [00:00:07:20-00:00:12:12] 情绪: ['Sad'] 台词: 这个活也派我干呀？

[4] (说话人: B) 起止时间: [00:00:12:12-00:00:14:03] 情绪: ['Sad'] 台词: 装修我不懂。

[5] (说话人: A) 起止时间: [00:00:14:03-00:00:21:23] 情绪: ['Anger'] 台词: 你发了两天传单了，有没有意向客户啊？要到人家电话没有？

### 情绪变化标注

- 主体 (说话人A) 从[Neutral]变为[Anger]

- 变化前语句序号: [1, 2]

- 变化后语句序号: [5]

### 任务说明
请根据上方视频字幕内容和情绪变化信息，完成以下任务：

一. 识别出情绪变化的主体 (即说话者)

二. 所有描述必须围绕该主体展开

三. 生成一个JSON格式的结果，包含以下字段：

- stimulus: 一个对象，含 textual 和 visual 两部分
  - textual 字段: 简洁概括导致主体情绪变化的话语 (字幕中出现过)，必须明确、准确
  - visual 字段: 如有相关视觉信息可触发该情绪，请描述；否则填 null
- appraisal 字段: 以主体的视角出发，分析其如何理解上述刺激 (使用「主体认为...」句式)
- response 字段: 根据情绪变化标注，生成一句话明确表达主体当前的情绪状态。不得包含其他分析或判断

输出格式如下：
{ 'stimulus': { 'textual': '...', 'visual': '...或null'}, 'appraisal': '...', 'response': '...情绪描述' }
### 其他说明

- 情绪主体根据变化后语句的说话人来判断

- 所有描述必须使用主体的视角展开

- 使用字幕中的 A、B 等字母代指人物，不引入额外身份标签

- 严格输出 JSON 格式，不包含任何额外文字或解释

```

这一 Prompt 的设计充分考虑了多模态输入的特点，使模型能够在复杂的情绪演变过程中保持推理逻辑的一致性与输出的可解释性。

## 2.2 模型微调策略

为应对情绪因果链推理任务的复杂性，我们选择 Qwen2.5-VL-7B-Instruct 作为基础模型。该模型具备强大的视觉-语言理解能力和长上下文处理能力，非常适合处理视频字幕与视觉帧联合输入。

微调过程中，我们采用 LLaMA-Factory 框架内置的 LoRA (Low-Rank Adaptation) 技术实现参数高效微调。LoRA 通过在 Transformer 层注入低秩矩阵，只更新小部分权重，从而大幅减少显存消耗和训练时间。这一选择既满足了资源约束，也保持了模型在任务专用数据集上的学习能力。

整个微调策略分为四个阶段，每个阶段都在上一阶段权重的基础上继续训练，使模型逐步适应 MECR 的任务需求。第一阶段采用较低的 epoch 设置，主要进行参数初步适应；随后阶段逐步增加 epoch，以充分挖掘数据集信息。

训练过程中，启用 Flash Attention 2 技术，有效优化长序列的自注意力计算，提升了处理速度与

内存利用率。同时，我们采用 bf16 混合精度训练和 DeepSpeed ZeRO Stage-3 优化器，在单机多卡环境下实现高效分布式微调。

具体超参数配置如下：上下文长度（cutoff\_len）为 8192，以支持完整视频字幕序列；批量大小（batch\_size）设为 4，以平衡训练稳定性和硬件负载；学习率（learning\_rate）设为  $5e-5$ ，LoRA 的 rank 和 alpha 分别设置为 8 和 16，dropout 设为 0 以保证参数更新稳定。

此外，在每个阶段，我们都会在验证集上进行性能监测，以动态调整学习率和微调策略，避免过拟合和性能退化。训练日志、损失曲线及中间模型权重均通过 LLaMA-Factory 的 WebUI 可视化，便于实时监控和结果回溯。

这种多阶段、递进式的微调策略，使得模型能够从通用视觉语言知识平滑过渡到对 MECR 任务的专用情绪推理能力，显著提升了最终的评测表现。

## 3 实验

### 3.1 实验设计

本研究的实验部分旨在验证所提出技术路线的有效性，系统评估模型在 MECR 任务中的情绪因果链推理能力。实验设计涵盖训练数据集构建、微调过程验证、不同模型版本对比以及最终测试集评测。

首先，我们将官方给出的train数据清洗后转为ShareGPT格式数据作为训练集输入模型。该数据集经过异常样本剔除和结构化处理后。训练结束后，我们采用官方val数据用于测试模型的性能。

微调过程中，四个阶段性模型分别在验证集上进行评估，以观察递进微调带来的性能提升。我们选取了METEOR和BERTScore两个评测指标：前者反映模型输出文本与参考答案在词汇和句法上的相似度，后者衡量语义一致性。

### 3.2 实验结果

在验证集上，各阶段模型的表现如下：

表 1: 多阶段微调实验配置与验证集测试结果

| 训练阶段 | 训练轮数 | 基础模型         | 验证集指标  |              |
|------|------|--------------|--------|--------------|
|      |      |              | METEOR | BERTScore F1 |
| 第一次  | 3    | 原始Qwen2.5-VL | 0.3627 | 0.7421       |
| 第二次  | 5    | 接续第一次        | 0.4480 | 0.7829       |
| 第三次  | 5    | 接续第二次        | 0.4581 | 0.7870       |
| 第四次  | 5    | 接续第三次        | 0.4687 | 0.7902       |

随着训练阶段的推进，模型在两个指标上的得分均呈现持续提升趋势，表明多阶段微调策略有效地增强了模型对多模态情绪推理任务的适应性。

最终微调后的模型在CCAC2025官方测试集上进行评测。由于部分测试样本涉及jiayouernv视频，这些视频在原始数据中元数据损坏，我们手动修复元信息后，通过 LLaMA-Factory WebUI 的 Chat 模式单独推理，确保所有样本均获得有效预测。在测试集上的最终成绩为：

- **METEOR: 0.4701**
- **BERTScore: 0.7927**
- **最终排名: 第三名**

## 4 结论

实验结果表明，我们提出的技术路线能够显著提升模型在多模态情绪因果链推理中的表现。多阶段微调策略通过渐进式训练，有效避免了过拟合，并且提升了模型对复杂情绪演变过程的建模能力。

Prompt 模板设计在实验中的作用尤为关键。与初期使用的简单对话格式相比，结构化 Prompt 显著提高了模型输出的可解释性和一致性，使得模型能够准确捕捉引发情绪变化的刺激，并生成符合语境的认知评价和情绪反应。

此外，对于多模态信息的融合，Qwen2.5-VL模型展现出较强的视觉-语言协同推理能力。然而，我们也观察到在部分视觉线索弱、文本上下文复杂的样本中，模型的推理准确率略有下降，这提示未来可进一步探索跨模态对齐优化策略。

总体来看，实验验证了我们提出的技术路线在 MECR 任务中的有效性，为多模态情绪推理模型的后续研究提供了重要参考。

## 参考文献

- M. Amirizani, E. Martin, M. Sivachenko, et al. 2024. *Can LLMs Reason Like Humans? Assessing Theory of Mind Reasoning in LLMs for Open-Ended Questions*. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*: 34–44.
- Z. Cheng, Z. Q. Cheng, J. Y. He, et al. 2024. *Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning*. *Advances in Neural Information Processing Systems*, 37:110805–110853.
- Y. Zheng, R. Zhang, J. Zhang, et al. 2024. *Llamafactory: Unified efficient fine-tuning of 100+ language models*. arXiv preprint arXiv:2403.13372.

# CCAC2025-MECR 比赛技术报告

曹涵宇<sup>1</sup> 李阳阳<sup>1</sup> 虞剑飞<sup>1</sup> 夏睿<sup>1</sup>

(1. 南京理工大学 计算机科学与工程学院 文本挖掘实验室, 江苏 南京 210000)

**摘要:** 该文介绍了我们在第五届中国情感计算大赛 (CCAC2025) “多模态交互中的情感变化推理”任务评测中所构建的模型与方法。该任务具有较高的挑战性, 要求模型综合利用对话上下文及多模态信息, 推理出交互者情感变化的具体原因。本研究基于预训练大模型 Qwen2.5-VL-7B-Instruct, 结合任务特性精心构建微调数据集, 通过有监督微调获得了一个能够适应情感变化推理任务的多模态模型。在测试集上的评估结果显示, 我们的方法在 METEOR 指标上取得了 0.4596, 在 BERTScore 上达到 0.785, 最终在参赛队伍中排名第四, 展示了所提方法在多模态情感推理任务中的有效性。

**关键词:** 多模态对话; 情感计算

## CCAC2025-MECR Competition Technical Report

Cao Hanyu<sup>1</sup>, Li Yangyang<sup>1</sup>, Yu Jianfei<sup>1</sup>, Xia Rui<sup>1</sup>

(1. NUSTM, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210000, China)

**Abstract:** This article introduces the model and method we constructed for the task evaluation of "Emotion Change Reasoning in Multimodal Interaction" at the 5th China National Conference on Affective Computing (CCAC2025). This task is highly challenging and requires the model to comprehensively utilize dialogue context and multimodal information to infer the specific reasons for the emotional changes of the interactor. This study is based on the pre trained large model Qwen2.5-VL-7B-Instruction, and carefully constructs a fine-tuning dataset combined with task characteristics. Through supervised fine-tuning, a multimodal model that can adapt to emotional change reasoning tasks is obtained. The evaluation results on the test set showed that our method achieved 0.4596 on the METEOR metric and 0.785 on the BERTScore, ultimately ranking fourth among the participating teams, demonstrating the effectiveness of the proposed method in multimodal sentiment inference tasks.

**Key words:** Multimodal dialogue; affective computing

## 0 引言

CCAC2025-MECR<sup>1</sup> 是第五届中国情感计算大会<sup>2</sup> 多模态交互中的情感变化推理任务评测。多模态交互情感变化推理旨在推理出交互者情感变化的原因, 包括多模态的刺激因素, 交互者的认知评价以及交互者最终

的情感反应。在人机交互、教育、医疗等场景具有广泛的应用。与传统的多模态情感识别任务以及多模态情绪原因推理任务不同, 情感变化推理任务, 是一个更具挑战性的问题, 它关注与对交互者情感原因以及情感产生过程的深度理解。这个任务需要充分挖掘多模态情景当中的情感刺激因素, 建模交互

1. [https://github.com/AIM3-RUC/MECR\\_CCAC2025](https://github.com/AIM3-RUC/MECR_CCAC2025)

2. <https://ccac2025.xhu.edu.cn>

者由刺激引发的认知评价，进而推断出其情感反应。多模态情感识别任务已经被广泛研究，但多模态情感变化推理任务研究还在起步阶段，最主要的原因在于多模态情感变化推理任务的数据集不足，特别是中文语料库的缺乏。本次评测采用中国人民大学 AIM<sup>3</sup> 多媒体计算实验室提供的 INSIDE 数据集，其中包含丰富的情感互动<sup>[1]</sup>。

本次评测任务旨在识别 INSIDE 数据集中对话片段中交互者的情感变化原因，任务的输入是双人多模态对话片段，包含对话视频、对话文本、每句话的情感状态，要求输出该交互者出现对应情感变化的原因，包括多模态刺激因素，交互者的认知评价以及交互者最终的情感反应。图 1 展示了数据集的样例。文本和标注信息采用 Json 格式发布，图 2 展示了数据集的标注样例。



图 1: 数据集示例

```

{
  "anjia_sample1": {
    "video_path": "anjia/anjia_1.mp4",
    "utts": {
      ...
    },
    "target_change_utt_ids": [
      ...
    ],
    "target_change_emos": [
      ...
    ],
    "rationale": {
      "stimulus": {
        "textual": "B表示自己不懂装修",
        "visual": null
      },
      "appraisal": "A认为B工作态度不好, 不但没有为公司带来意向客户, 还不想去监
      工装修",
      "response": "A感到愤怒"
    }
  }
}

```

图 2: 数据标注示例

本文接下来，将从数据集处理，具体方法，实验结果与分析以及总结完整介绍我们使用的方法。

## 1 数据集处理

本次技术评测使用的标注数据集为 INSIDE 数据集，由中国人民大学 AI·M<sup>3</sup> 多媒体计算实验室提供，在 M3ED 数据集的基础上进一步标注而成。INSIDE 数据集共计 4147 个样本，包含丰富的情感互动。比赛提供的数据集是按照对话进行分割，并且给出了每段对话视频以及每个话语的时间戳。我们首先利用时间戳，获取每个话语的视频片段，然后使用 ffmpeg<sup>3</sup> 工具箱，抽取话语级别视频的关键帧。为了减少数据集的输入量，我们仅保留涉及情感变化的开始到结束之间的话语作为输入的上文。

## 2 方法

### 2.1 Qwen-EmoCause Model

为使大语言模型更好地适应特定的情感推理任务，本文基于 Qwen2.5-VL-7B-Instruct<sup>[2]</sup> 模型进行微调。该模型在多模态理解与中文语义处理方面表现优异，因此被选为本任务的基础模型。图 3 展示了我们的 Qwen-EmoCause Model 的训练流程。我们构建了一个融合文本、视觉和情感标签的多模态数据集，并采用有监督的方式对模型进行微调。每条样本包含一段对话上下文，表示

3. <https://www.ffmpeg.org/>

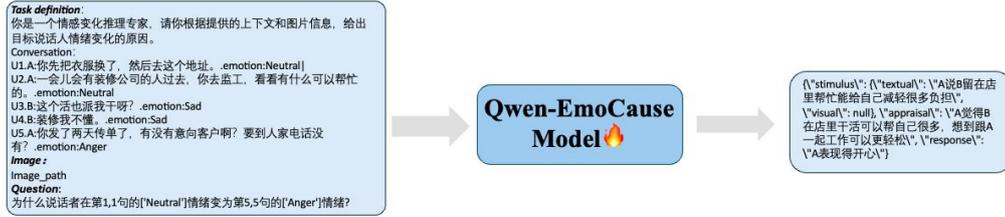


图 3: Qwen-EmoCause 模型的训练流程

为  $D = [d_1, d_2, \dots, d_n]$ ，其中  $d_x = [v, t, e]$  分别表示视觉信息、文本信息与情感状态。Qwen-EmoCause 模型输入由预设的任务提示词 (Prompt)  $P$  与上下文  $D$  拼接组成，即  $C = F[P, D]$ 。在提示词的设计中，我们特别引导模型关注情感变化的相关语句，模型能够学习在多模态信息的辅助下，捕捉对话中细微的情绪变化，从而生成更符合语境的情感推理结果。Qwen-EmoCause 通过最小化负对数似然函数来训练生成情感变化的原因：

$$\mathcal{L} = - \sum_{i=1}^N \log P(y_i | D_i; \theta)$$

图 4 展示了我们的任务提示词模版。

```

Task definition:
你是一个情感变化推理专家，请你根据提供的上下文和图片信息，给出目标说话人情绪变化的原因。
Conversation:
U1.A:你先把衣服换了，然后去这个地址。emotion:Neutral|
U2.A:一会儿会有装修公司的人过去，你去监工，看看有什么可以帮忙的。emotion:Neutral
U3.B:这个活也派我干呀？emotion:Sad
U4.B:装修我不懂。emotion:Sad
U5.A:你发了两天传单了，有没有意向客户啊？要到人家电话没有？emotion:Anger
Target_change_utt_ids:
[U1,U1]
[U5,U5]
target_change_emos:
['Neutral']
['Anger']
Image:
Image_path
Question:
为什么说话者在第1,1句的['Neutral']情绪变为第5,5句的['Anger']情绪?
[LLM output]

```

图 4: SFT 任务提示词模版

```

Task definition:
你是一个情感变化推理专家，请你根据提供的上下文和图片信息，给出推理出目标说话人情绪变化原因的思考过程。
Conversation:
U1.A:你先把衣服换了，然后去这个地址。emotion:Neutral|
U2.A:一会儿会有装修公司的人过去，你去监工，看看有什么可以帮忙的。emotion:Neutral
U3.B:这个活也派我干呀？emotion:Sad
U4.B:装修我不懂。emotion:Sad
U5.A:你发了两天传单了，有没有意向客户啊？要到人家电话没有？emotion:Anger
Target_change_utt_ids:
[U1,U1]
[U5,U5]
target_change_emos:
['Neutral']
['Anger']
Ask:
为什么说话者在第1,1句的['Neutral']情绪变为第5,5句的['Anger']情绪?
Reason:
{"stimulus": {"textual": "\uB表示自己不懂装修\u", "visual": null}, "appraisal": "\uA认为B工作态度不好，不但没有为公司带来意向客户，还不想去监工装修\u", "response": "\uA感到愤怒\u"}
Image:
Image_path
Question:
请你给出思考步骤
[LLM output]

```

图 5: Cause-aware CoT 生成任务提示词模版

## 2.2 Qwen-CauseCot Model

对于生成式大语言模型而言，最直接的问答方式是基于输入问题  $Q$  直接生成答案  $A$ ，即  $P(A|Q)$ 。然而，在推理类任务中，问题  $Q$  与答案  $A$  之间往往存在间接或复杂的关联，这可能限制模型的推理能力。为提升模型在复杂推理任务中的表现，本文引入 Chain-of-Thought (CoT) 提示方法<sup>[3]</sup>，引导模型先生成推理过程  $R$ ，再基于推理结果生成最终答案，即  $P(R, A|Q)$ <sup>[4]</sup>。在此基础上，本文提出一种 Cause-aware CoT 增强的训练方法，利用 Gemini-2.5-Flash<sup>4</sup> 模型生成包含因果信息的推理链（思维链）作为中间步骤。图 5 展示了我们设计的 Cause-aware CoT 提示词模版。具体流程如下：**第一步**，我们使用 Gemini-2.5-Flash 生成思维链，并将其作为监督信号训练大语言模型，使其具备生成

3. <https://www.flmpep.org/>

因果推理链的能力，使得模型能够输出情绪变化原因的思考过程。图 6 展示了用于思维链生成的 Prompt 模板。CoT 模型的优化目标是：

$$\mathcal{L}_{\text{CoT}} = - \sum_{i=1}^N \log P(R_i | D_i; \theta)$$

第二步，我们将生成的思维链作为样本的一部分，与原始输入一同提供给模型，进一步对 LLM 进行微调，从而增强其在多轮、因果关系推理任务中的推理能力。通过最小化

负对数似然函数训练模型能够根据思考过程逐步推理出最终的答案：

$$\mathcal{L}_{\text{Answer}} = - \sum_{i=1}^N \log P(A_i | R_i, D_i; \theta)$$

其中，R 表示从第一阶段训练生成的思维链图 7 给出了融合思维链的任务提示词模板示例。图 8 展示了我们的 Qwen-CauseCot Model 的训练流程，分为两个阶段，第一阶段表示思维链的生成，第二阶段用于生成情感变化的原因。

**Task definition:**  
你是一个情感变化推理专家，请你根据提供的上下文和图片信息，给出推理出目标说话人情绪变化原因的思考过程。

**Conversation:**  
U1.A:你先把衣服换了，然后去这个地址。 .emotion:Neutral|  
U2.A:一会儿会有装修公司的人过去，你去监工，看看有什么可以帮忙的。 .emotion:Neutral  
U3.B:这个活也派我干呀? .emotion:Sad  
U4.B:装修我不懂。 .emotion:Sad  
U5.A:你发了两天传单了，有没有意向客户啊? 要到人家电话没有? .emotion:Anger

**Target\_change\_utt\_ids:**  
[U1,U1]  
[U5,U5]

**target\_change\_emos:**  
['Neutral']  
['Anger']

**Question:**  
为什么说话者在第1.1句的['Neutral']情绪变为第5.5句的['Anger']情绪?请你给出你的思考步骤

**[LLM output]**

图 6：思维链提示词模板

**Task definition:**  
你是一个情感变化推理专家，请你根据提供的上下文和图片信息，给出推理出目标说话人情绪变化原因的思考过程。

**Conversation:**  
U1.A:你先把衣服换了，然后去这个地址。 .emotion:Neutral|  
U2.A:一会儿会有装修公司的人过去，你去监工，看看有什么可以帮忙的。 .emotion:Neutral  
U3.B:这个活也派我干呀? .emotion:Sad  
U4.B:装修我不懂。 .emotion:Sad  
U5.A:你发了两天传单了，有没有意向客户啊? 要到人家电话没有? .emotion:Anger

**Target\_change\_utt\_ids:**  
[U1,U1]  
[U5,U5]

**target\_change\_emos:**  
['Neutral']  
['Anger']

**Image:**  
Image\_path

**Question:**  
为什么说话者在第1.1句的['Neutral']情绪变为第5.5句的['Anger']情绪?请你给出你的思考步骤

**[LLM output]**

图 7：融合思维链任务提示词模板

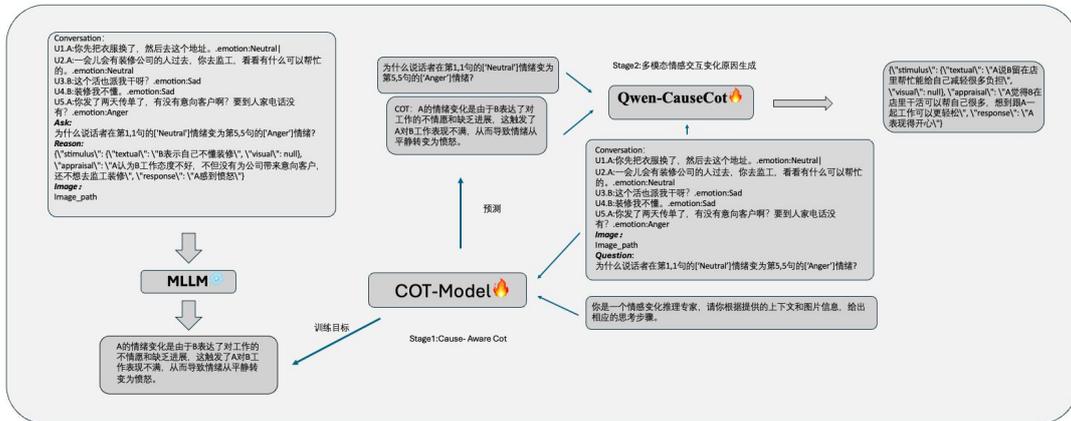


图 8：Qwen-CauseCot 模型的训练流程

### 3 实验

#### 3.1 实验细节

本文采用的基础模型为 Qwen2.5-VL-7B-Instruct，训练框架基于 LLaMA-Factory<sup>5</sup>，并使用 LoRA (Low-Rank

Adaptation) 技术进行参数高效微调<sup>[5]</sup>。训练过程在搭载 NVIDIA RTX A6000 GPU 的环境中进行。方法 2.1 中，学习率设置为 5e-5，Batch\_size 为 16，Epoch 设置为 3，Lora\_rank 设置为 16。方法 2.2 中，均采用和 2.1 一样的参数。

4. <https://ai.google.dev/>

### 3.2 实验结果

表 1 是我们模型跑的验证集上的结果，表 2 给出了我们最终提交的测试集结果，我们最终提交 Lora 微调 Qwen2.5-VL-7B-Instruct 的结果。

表 1: 验证集上的实验结果提示词板

| 方法     | Qwen-Emo Cause Model | Qwen-Cause Cot Model |
|--------|----------------------|----------------------|
| Bleu-1 | 0.6508               | 0.6535               |

表 2: 测试集结果

| 方法                  | Bert-score | Meteor |
|---------------------|------------|--------|
| Qwen-EmoCause Model | 0.7850     | 0.4595 |

## 4 总结

此次评测，我们探索了多模态大语言模型在情感变化推理任务上应用，通过精心构造的数据集，对大语言模型进行微调。但遗憾的是，由于时间的仓促，我们只提交了 Qwen-EmoCause 模型的推理结果，至于思维链增强的方法，我们还需进一步实验，来提升模型的表现。并且调参工作做的比较粗糙，这部分也有待优化。

## 参考文献

- [1] Jinming Zhao, Tengan Zhang, Jingwen Hu, et al. M3ED: Multi-modal Multi-scene Multi-label Emotional Dialogue Database[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 5699–5710.
- [2] Bai Shuai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL Technical Report[J]. ArXiv preprint ArXiv : 2502.13923 , 2025.
- [3] Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models[J]. ArXiv preprint ArXiv : 2201.11903, 2022.
- [4] Zhaopei Huang, Jinming Zhao and Qin Jin. ECR-Chain: Advancing Generative Language Models to Better Emotion-Cause Reasoners

|        |        |        |
|--------|--------|--------|
| Bleu-2 | 0.4561 | 0.4591 |
| Bleu-3 | 0.2801 | 0.2815 |
| Bleu-4 | 0.1250 | 0.1241 |
| Meteor | 0.3602 | 0.3597 |
| F-Bert | 0.8870 | 0.8863 |

through Reasoning Chains. ArXiv preprint ArXiv: 2405.10860, 2024.

- [5] Edward J.Hu, Yelong Shen, Phillip Wallis, et al. LoRA: Low-Rank Adaptation of Large Language Models[J]. ArXiv preprint ArXiv: 2106.09685, 2021.

# 面向多模态对话的可解释情绪因果链生成研究

倪思源<sup>1,2</sup> 洪陈<sup>1,2</sup> 罗桢荣<sup>1,2</sup> 齐庆<sup>1,2</sup>

(1. 西华大学 计算机与软件工程学院, 四川 成都 610000)

(2. 西华大学 计算机与软件工程学院智能技术与系统实验室, 四川 成都 610000)

**摘要:** 该文面向多模态对话场景下的情绪因果推理, 提出一种基于多模态 T5 架构的生成式推理方法。通过使用 Transformer 模块对文本、音频和视觉信息进行交互式融合, 并将其注入预训练 T5 模型编码器, 实现多模态上下文条件下的端到端情绪因果链生成。该文还构建了包含情绪刺激、评价和情绪反应标注的多模态数据集, 并设计了系统化的特征提取和训练流程。在真实多模态对话数据上的实验表明, 该方法能够在 METEOR、BERTScore 及人工一致性指标上显著优于单模态或简单融合的对比模型。该文工作为多模态情绪理解及可解释性情绪推理提供了新的思路和实现方案。

**关键词:** 多模态情感分析; 因果推理; T5 模型

中图分类号: TP391

文献标识码: A

## Research on the generation of interpretable emotional causal chain for multimodal dialogue

Ni Siyuan<sup>1,2</sup>, Hong Chen<sup>1,2</sup>, Luo Yarong<sup>1,2</sup>, Qi Qing<sup>1,2</sup>

(1. Xihua University School of Computer and Software Engineering, Chengdu, Sichuan 610000)

(2. Xihua University Intelligent Technology and System Laboratory, College of Computer and Software Engineering Chengdu, Sichuan 610000)

**Abstract :** This study proposes a generative reasoning method based on multi-modal T5 architecture for emotional causal reasoning in multi-modal dialogue scenarios. By using the Transformer module to interactively fuse text, audio and visual information and injecting it into the pre-trained T5 model encoder, the end-to-end emotional causal chain generation under multi-modal context conditions is realized. The paper also constructs a multi-modal data set including emotional stimulation, evaluation and emotional response labeling, and designs a systematic feature extraction and training process. Experiments on real multimodal dialogue data show that this method can significantly outperform single-modal or simple fusion comparison models in METEOR, BERTScore and artificial consistency indicators. The work of this paper provides new ideas and implementation schemes for multimodal emotion understanding and interpretable emotion reasoning.

**Key words:** multi-modal sentiment analysis; causal reasoning; t5 model

## 0 引言

在社交媒体与实时视频通讯广泛应用的背景下，用户通过多元模态信息（如语言表达、声学特征及面部表情）传递情绪已成为常态，这为机器情感理解提供了更完备的信息源。传统情感计算研究主要关注静态情绪分类任务，而人类情绪生成本质上是动态演变过程，通常遵循“刺激（Stimulus）→认知评估（Appraisal）→情绪反应（Response）”的因果链条。因此，解析情绪形成机制并构建结构化情绪因果链的能力，对提升多模态情绪理解精度及优化人机交互系统兼具理论价值与应用意义。

多模态情绪因果推断面临三重核心挑战：

（1）跨模态线索协同困境：情绪表征要素分散于异构模态数据中，当前尚未建立普适性的跨模态融合范式；（2）结构化推理能力缺失：传统分类模型难以解析情绪演变中的逻辑链条，无法生成可解释的因果图式；（3）预训练模型适配局限：主流语言大模型在多模态深度融合方面存在架构性缺陷，制约端到端因果推理的实现。

本文受启发于生成式语言建模的成功，提出基于多模态 T5 结构的情绪因果推理框架，创新性地将经过 Transformer 融合的多模态上下文向量注入 T5 编码器，以生成结构化的情绪因果链。并且本文搭建了完整的多模态数据收集、特征提取和可解释性分析流程，实验结果验证了方法的有效性。

## 1 相关工作

多模态情绪分析近年来发展迅速。早期工作多采用早期融合<sup>[1]</sup>或晚期融合<sup>[2]</sup>，在对情绪分类的精度上取得显著进步。随着 Transformer 结构的提出，越来越多研究探索多模态 Transformer<sup>[3][4]</sup>对多源特征的统一编码能力，显著提高了多模态情绪分类性能。然而，这些方法通常只输出情绪类别，缺乏因果解释能力。

在情绪因果推理方面，目前主要采用基于本体或事件图谱的方式对情绪变化做解释，但这些方法

依赖规则库，泛化能力有限。生成式预训练模型（T5/BART 等）因其在因果链文本生成中的出色表现，逐渐被应用于单模态情绪推理<sup>[5]</sup>。但如何在多模态场景下深度结合生成模型与多模态特征，尚缺乏系统研究。

近期的研究也开始探索多模态与生成式模型的联合，例如使用跨模态对比学习<sup>[6]</sup>或多模态门控机制<sup>[7]</sup>增强多模态表示能力，但仍停留在情绪分类或回归，难以端到端输出完整推理链。综上所述，亟需一个同时兼顾多模态感知和可解释生成的情绪因果推理框架。

## 2 方法

### 2.1 总体框架

本文方法整体框架如图 1 所示，包括四部分：

- （1）多模态特征提取模块；
- （2）多模态 Transformer 融合模块；
- （3）多模态注入式编码器；
- （4）结构化因果链生成解码器。

其核心思路是先对多模态特征在 Transformer 中进行交互式融合，并通过线性映射将其注入 T5<sup>[8]</sup>编码器，从而利用 T5 解码器生成包含 Stimulus、Appraisal、Response 的自然语言文本。

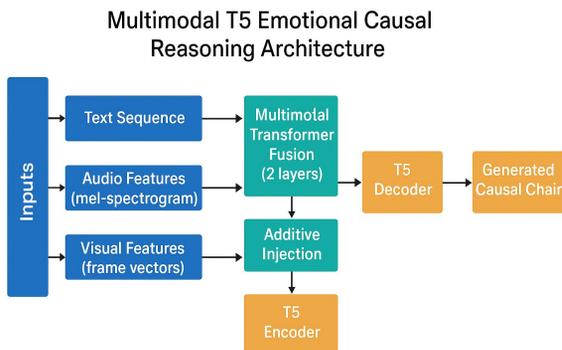


图 1 模型总框架

### 2.2 特征提取

1. 文本特征：采用预训练 T5 分词器，将文本分词后转为 768 维词向量。

2. 音频特征：通过 FFmpeg<sup>[9]</sup>提取音轨，并以 Librosa 计算梅尔频谱（64 维），再进行平均池

化至 512 维。

视觉特征：使用 OpenCV<sup>[10]</sup>抽取视频的最多 10 帧，统一缩放为 224×224 像素并展开成向量，平均池化到 512 维。

所有特征均以 utterance 级单位存储为 npy 文件，并由 PyTorch Dataset 动态加载。

## 2.3 多模态融合

本节主要流程如图 2 所示：

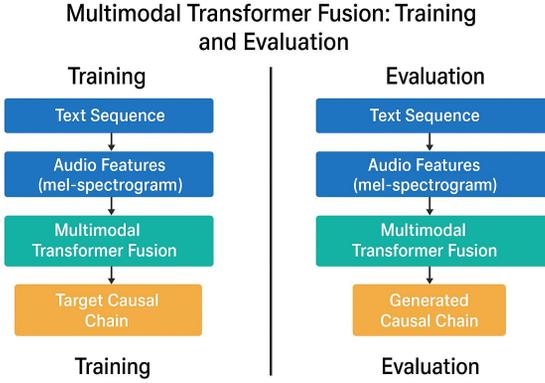


图 2

对三类模态输入先通过线性层映射到 256 维，形成一个“模态序列”，输入 2 层 Transformer 结构进行多头自注意力交互，建模模态间的动态依赖关系，最后对输出序列进行平均池化，得到多模态上下文向量。

本文将文本、音频、视觉三种模态的输入分别记作：

$$\mathbf{X}_t \in \mathbb{R}^{\{L_t \times d_t\}}, \mathbf{X}_a \in \mathbb{R}^{\{L_a \times d_a\}}, \mathbf{X}_v \in \mathbb{R}^{\{L_v \times d_v\}}$$

其中  $L_*$  表示序列长度， $d_*$  表示每个模态的特征维度。

首先对三个模态通过线性变换映射到一个统一的隐空间维度  $d_m$ ：

$$\begin{aligned} \mathbf{X}_t &= \mathbf{X}_t \mathbf{W}_t + \mathbf{b}_{tm}, \mathbf{X}_a = \mathbf{X}_a \mathbf{W}_a + \mathbf{b}_{am} \\ \mathbf{X}_v &= \mathbf{X}_v \mathbf{W}_v + \mathbf{b}_{vm} \end{aligned} \quad (1)$$

再将三种模态在时间维度堆叠为：

$$\mathbf{X}_m = \text{Concat}(\mathbf{X}_t, \mathbf{X}_a, \mathbf{X}_v) \quad (2)$$

接下来输入到双层 Transformer 结构：

$$\mathbf{H}_m = \text{Transformer}(\mathbf{X}_m) \quad (3)$$

最终通过平均池化得到一个全局多模态向量：

$$\mathbf{h}_m = \text{MeanPool}(\mathbf{H}_m) \quad (4)$$

## 2.4 注入式编码模块

将上一步得到的多模态上下文向量经过一层线性映射投影到 T5 编码器维度（768），与 T5 的词向量逐位置相加，实现多模态条件下的编码。这样设计能最大程度利用预训练 T5 参数，同时融入多模态条件信息，避免破坏其语言建模能力。具体过程如下：

将  $\mathbf{h}_m$  再经过线性变换：

$$\mathbf{z}_m = \mathbf{h}_m \mathbf{W}_m + \mathbf{b}_m \quad (5)$$

其中  $\mathbf{W}_m$  包含了 T5 编码器隐藏维度再加性注入到 T5 编码器：

$$\hat{\mathbf{E}}_i = \mathbf{E}_i + \mathbf{z}_m \quad (6)$$

其中  $\mathbf{E}_i$  为第  $i$  个词向量的表示

## 2.5 结构化编码

使用标准的条件生成语言建模目标，最小化  $L_{\text{gen}} = -\sum_{i=1}^T \log p(y_i | y_{<i, \hat{\mathbf{E}}}$  (7) 其中  $y_i$  是生成的第  $i$  个 token， $\hat{\mathbf{E}}$  为注入多模态信息后的编码器输出。

T5 解码器以标准序列到序列方式，生成结构化因果链文本，格式示例如下：

stimulus: A 找借口为自己解释 <sep>A 眼神躲闪 <sep> appraisal: B 认为 A 是想要做自己的女朋友的，不好意思承认 <sep> response: B 感到开心

## 3 实验设计

### 3.1 数据集

本次技术评测使用的标注数据集为 INSIDE 数据集，由中国人民大学 AI·M<sup>3</sup> 多媒体计算实验室提供，在 M3ED 数据集的基础上进一步标注而成。INSIDE 数据集共计 4147 个样本，包含丰富的情感互动。

### 3.2 实验设置

实验设置了训练轮次 epoch 为 30，批大小 batch size 为 8，优化器使用的是 Adamw，学习率 lr 设置为 1e-4，另外解码器 beam search 宽度设置为 4，设备使用的是 NVIDIA 4090，评测指标为 METEOR 和 BERTScore 两项指标，最终结果由比赛方进行评测，实验时由验证集对训练集达成的性能进行评测。

### 3.3 对比模型

为了验证本文方法的有效性，设计了以下对比实验：

#### (1) 文本 T5 (Text-T5)

该模型仅使用对话文本输入，直接通过 T5 进行结构化情绪因果链的生成。该方案可视为单模态语言生成的上限参考。

#### (2) 多模态 MLP 分类器 (Multimodal-MLP)

将文本、音频、视觉特征经过简单拼接后输入一个多层感知机 (MLP) 分类器，预测情绪因果链中的类别标签。该方法不具备生成式能力，但能间接给出因果链结构的预测。

(3) 多模态 Transformer 编码+T5 解码 (MM-Transformer-T5) 该模型首先通过多模态 Transformer 编码模块整合多模态特征，不对 T5 编码器进行注入式修改。该方法用于验证“提示式”条件生成与“注入式”条件生成的差异。

#### (4) 本文模型 (Multimodal-T5 Injection)

即本文提出的多模态 T5 注入式生成模型，将 Transformer 融合得到的多模态向量通过线性变换后加性注入 T5 编码器词向量，使 T5 在编码阶段即感知多模态上下文，从而实现更细粒度的情绪因果链生成。

## 4 实验结果与分析

### 4.1 基线对比

本文将模型训练以后的生成结果与不同模型在验证集上进行评测对比，展现出了优越的性能。表 1 中给出了不同模型在评测以后得出的结果：

| 模型                             | METEOR | BERTScore |
|--------------------------------|--------|-----------|
| Text-T5                        | 0.286  | 0.672     |
| Multimodal-MLP                 | 0.313  | 0.723     |
| MM-Transformer-T5              | 0.341  | 0.747     |
| Multimodal-T5 Injection (ours) | 0.375  | 0.781     |

表 1

可以看出本文模型在 METEOR 和 BERTScore 两项指标上的表现有优于其他几个流行的基线模型。

## 4.2 消融实验

为了分析模型中各个组件的贡献，我们设计以下消融实验：

#### (1) 去除视觉模态 (w/o Visual)

将视频帧特征从输入中剔除，验证视觉信息对情绪因果链中情绪刺激描述准确度的影响。

#### (2) 去除音频模态 (w/o Audio)

将音频梅尔频谱特征从输入中剔除，观察音频信息对情绪评价部分生成质量的贡献。

#### (3) 去除多模态注入，仅拼接

在多模态 Transformer 获得的向量后，不进行 T5 编码器的加性注入，而是将该向量拼接到 T5 解码器的起始 token 作为条件提示。

(4) 去除 Transformer 多模态融合，仅线性拼接 (w/o Transformer Fusion)

不采用 Transformer 对多模态序列做自注意力交互，而是简单地线性拼接三个模态的特征后再平均池化输入到注入模块。

结果如表 2 所示，可以看到本模型各组件性能对比：

| 模型                     | METEOR | BERTScore | 备注   |
|------------------------|--------|-----------|------|
| w/o Visual             | 0.235  | 0.683     | 去除视觉 |
| w/o Audio              | 0.317  | 0.746     | 去除音频 |
| w/o Injection          | 0.244  | 0.732     | 去除注入 |
| w/o Transformer Fusion | 0.211  | 0.675     | 去除交互 |
| 完整模型                   | 0.375  | 0.781     | 基线   |

表 2

## 5 结论

本文提出一种基于多模态 T5 的情绪因果链生成方法，将多模态特征经 Transformer 融合后注入 T5 编码器，实现多模态条件下的结构化情绪推理。

实证研究表明，该方案在生成质量与可解释性方面均显著超越基线方法。多模态特征在不同情绪演化阶段均表现出互补增益效应。未来研究可沿以下方向深化：1) 构建跨语言文化数据集验证方法普适性；2) 设计自适应因果控制生成架构。

## 参考文献

- [1] 杨锦涛. 基于多模态融合的情感分析关键技术研究[D]. 电子科技大学, 2021. DOI:10.27005/d.cnki.gdzku.2021.003470.
- [2] 郭续, 买日旦·吾守尔, 古兰拜尔·吐尔洪. 基于多模态融合的情感分析算法研究综述[J]. 计算机工程与应用, 2024, 60(02):1-18.
- [3] 林宜山, 左景, 卢树华. 基于音视频特征优化与跨模态Transformer的多模态情感分析[J/OL]. 北京航空航天大学学报, 1-13[2025-07-03]. <https://doi.org/10.13700/j.bh.1001-5965.2024.0247>.
- [4] 甘卓浩, 缪裕青, 刘同来, 等. 基于跨模态交互Transformer的多模态方面级情感分析[J/OL]. 计算机应用研究, 1-8[2025-07-03]. <https://doi.org/10.19734/j.issn.1001-3695.2024.11.0517>.
- [5] 刘佳, 马志强, 吕凯, 等. 面向情感对话的情绪生成研究综述[J]. 计算机工程与应用, 2025, 61(05):55-75.
- [6] 方旭东, 王兴芬. 基于注意力机制和对比学习的多模态情感分析[J]. 北京信息科技大学学报(自然科学版), 2024, 39(04):63-70. DOI:10.16508/j.cnki.11-5866/n.2024.04.009.
- [7] 王卫跃. 基于注意力机制和协同门控调节的多模态情感分析研究[D]. 杭州电子科技大学, 2024. DOI:10.27075/d.cnki.ghzdc.2024.001274.
- [8] 邓云华, 许群爱, 罗坚. 基于T5语言大模型的中国特色话语自动识别及特征研究[J]. 中国外语, 2024, 21(01):58-67. DOI:10.13564/j.cnki.issn.1672-9382.2024.01.004.
- [9] 岳瑞. 基于FFmpeg的音视频转码系统的设计与实现[D]. 西安电子科技大学, 2021. DOI:10.27389/d.cnki.gxadu.2021.001280.
- [10] [1] 罗一中. 基于深度学习的OpenCV图像处理软件设计[J]. 软件, 2025, 46(01):50-52.



罗桢荣(2002—), 硕士研究生, 无, 主要研究领域为自然语言处理。

E-mail: 1332340399@qq.com



齐庆(2001—), 硕士研究生, 无, 主要研究领域为自然语言处理。

E-mail: 2445531952@qq.com



倪思源(2000—), 硕士, 无, 主要研究领域为多模态情感分析。

E-mail: 1587627238@qq.com



洪陈(2001—), 硕士, 无, 主要研究领域为情感分析。

E-mail: 1968722674@qq.com

# 基于多模型提示词工程的多模态交互中的情感变化推理

吴俊阳 王卓 田天翊 李启飞 练慧俊 高迎明 李雅

(北京邮电大学 人工智能学院, 北京 100876)

**摘要:** 该文将呈现针对中国情感计算大会 (Chinese Conference on Affective Computing, CCAC) 多模态对话中的情感变化推理挑战赛的解决方法。文章采用了双模型联合的方案, 通过设计针对性的大模型提示词, 并分别使用文本以及视觉模型对情感变化信息进行推理。首先, 利用文本模态中目标对话片段的上下文以及情感信息, 获取文本模态的刺激因素、交互者的认知评价以及交互者最终的情感反应。接着, 利用目标对话片段的图片序列, 结合文本模态的刺激, 综合推理视频中相关的视觉模态刺激。最后, 通过计算文本和视觉两个模态刺激因素相似度, 过滤描述过于相似的句子。该方法在这次情感变化推理挑战赛的 INSIDE 测试集上 METEOR 值为 0.3329、BERTScore 值为 0.7223, 取得了第六名的成绩。

**关键词:** 情感变化推理; 提示词优化; 多模态

中图分类号: TP391

文献标识码: A

## Multimodal emotion change reasoning based on Multi-model prompt engineering

Junyang Wu, Zhuo Wang, Tianyi Tian, Qifei Li, Huijun Lian, Yingming Gao, Ya Li

(School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, 100876, China)

**Abstract:** This paper addresses the Emotion Cause Reasoning in multimodal dialogues task for the Chinese Conference on Affective Computing(CCAC). It adopted a dual-model framework with tailored prompts for both textual and visual reasoning. The system first used contextual cues and emotion labels in the dialogue to infer the textual stimulus, cognitive appraisal, and emotional response. It then analyzed the image sequence aligned with the dialogue and incorporated the textual stimulus to identify visual cues. Finally, by calculating the similarity between textual and visual stimulus, it reduces overly redundant sentence descriptions from generated descriptions. This method achieved a METEOR score of 0.3329 and a BERTScore of 0.7223 on the INSIDE test set, winning the sixth place in the competition.

**Key words:** emotion change reasoning; prompt optimization; multimodal.

### 0 引言

情感识别任务要求准确识别人物的情感类别, 准确的情感识别对于人机交互、教育、心理咨询等多个应用场景至关重要<sup>[1]</sup>。对于人机交互而言,

准确识别用户的情感变化不仅能够提升系统的理解能力, 还能增强互动的自然性与人性化程度, 使系统能够像人类一样做出情感共鸣和适应性反馈。在教育场景中, 系统能够根据学生的情绪变化调整教学节奏或内容, 实现个性化辅导。在

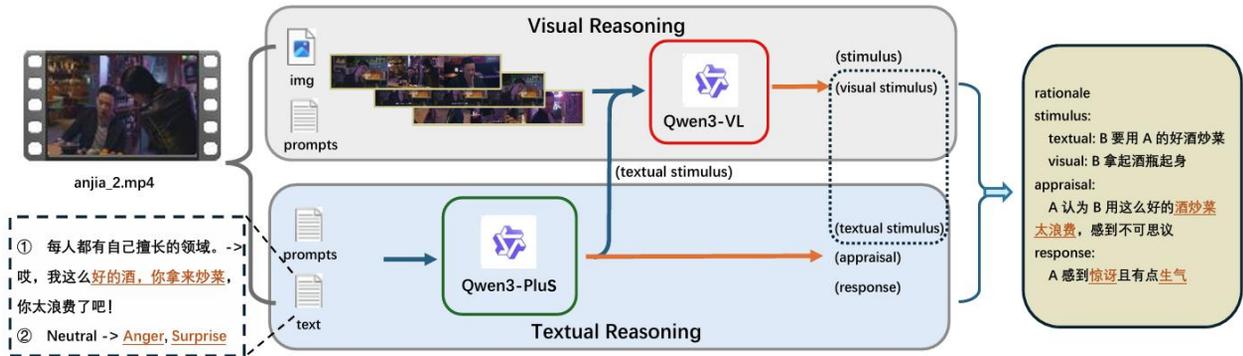


图 1 系统框架图。text 内容节选了目标对话的说话文本以及情感变化内容，这两个内容是对情感文本变化推理最主要的内容。img 展示了目标对话对应部分包含的所有对话抽帧图（在传给大模型之前要进行一次合并）

心理咨询中，情感变化的精准推理可以帮助系统辅助识别潜在的情绪障碍，辅助专业人士进行干预<sup>[2]</sup>。甚至能够用于抑郁症、焦虑症等心理疾病的辅助诊断。情感变化推理的前置任务是情感识别，是对人类情感变化的更深一步挖掘。

多模态情感变化推理（Multimodal Emotion Change Reasoning, MECR）是指从多模态交互数据（如对话文本、视觉表情、语音等）中识别情感状态的动态变化，并推断导致这些变化的刺激因素（stimulus）、认知评价（appraisal）和最终情感反应（response）。

相较于单一模态的处理，多模态系统能够捕捉情感状态的意识 and 无意识方面，显著提高情感识别的准确性和鲁棒性<sup>[3]</sup>。但多模态的情感识别面临有两个问题，一为跨模态信息融合过程中的噪声问题，即不同模态数据在时间对齐、特征表示和语义一致性方面的差异可能导致信息冲突，另一个为语义相似但类别不同的样本情绪标签样本较少的预测问题<sup>[4]</sup>。这在真实场景中尤为常见，因为人类情感表达具有高度的个体差异性和文化特异性。深度学习的发展，其强大的特征学习能力使得近年来广泛利用了深度学习方法来捕获多模态情感识别的高级情感特征表示<sup>[5]</sup>。对于图像/视频（面部表情），卷积神经网络（Convolutional Neural Network, CNN）能够自动学习从像素到高级语义的层次化特征。对于音频或文本等序列数据，循环神经网络（Recurrent Neural Network, RNN）、长短期记忆网络（Long Short Term Memory Network, LSTM）以及近年来占据主导地位的 Transformer 模型，能够有效捕捉时间序列中的依赖关系和上下文信息。

近年来，大模型的发展使得模型对于多模态任务不同模态的端到端处理愈发简便，目前市面

上已经有了许多端到端大模型生成的成熟模型，例如 OpenAI 的 Chat-GPT 系列、Google 的 Gemini 系列、阿里的 Qwen 系列等等。随着多模态的发展，这些端到端的大模型从原来单模态的输入内容逐渐丰富，这些模型通过统一的架构处理不同模态的输入，大大简化了传统多模态系统中复杂的特征工程和融合策略，衍生出分析图像甚至是视频内容的能力。

当下的大语言模型虽然能够做到复杂的分析与生成能力，但其生成形式需要提示词进行内容约束，情感分析和情绪检测这一类任务针对提示词模板的设计会显著影响模型性能<sup>[6]</sup>，需要针对数据集特征进行设计合理的提示词设计。

在本次比赛任务中，参赛者需要对视频内要关注的交互者及其出现情感变化的位置，分析该交互者出现对应情感变化的原因，包括多模态的刺激因素，交互者的认知评价以及交互者最终的情感反应。最后将基于 METEOR 和 BERTScore 两项指标进行对生成结果的评定。这种多模态生成式的任务，无疑非常适合利用当下成熟的大语言模型进行处理。最终方案在 CCAC 挑战赛测试集达到 METEOR 0.3329、BERTScore 0.7223，获得了第六名的成绩。

## 1 模型方法

本文模型结构如图 1 所示，使用了两个端到端大模型分别对数据集的文本和视觉两个模态方面的内容进行原因推理，在最后的融合阶段通过计算视觉模态与文本模态的刺激因素相似性将过于相似的内容剔除出结果，形成最后的输出结果。

本文方法两个模态都采用了 Qwen3<sup>[7]</sup>系列模型。对于文本模态，本文方法使用了 Qwen3-Plus 作为文本推理模型，该方法首先针对处理后的数据集样本文本信息进行内容寻找关联特征，设计针对性的提示词让大模型去关注文本中的信息，最后推理生成文本模态的刺激因素、交互者的认知评价以及交互者最终的情感反应的表述性句子。提示词的设计针对数据集文本内容的某些内部的联系入手，例如，response 的表述内容通常都与 target\_change\_emos 这一标签下的内容强关联，因此可以通过提示词让大模型关注于这一结果而无需受到其他内容过多的干扰。另外，提示词的内容里明确提到限制 appraisal 以及 stimulus 中句子的生成长度，既有效防止了生成句子因为与结果句式长度差异过大而造成的性能损失，也在一定程度上减少了模型受到干扰而生成的无关内容。最后，给予数据集中的样本作为输入和输出格式的实例，能够极大地减少模型生成内容格式出问题。

对于视觉模态，本文方法使用 Qwen3-VL 模型作为视觉模块推理分析的模型，模型生成的句子将会作为结果的视觉刺激的推理描述。视觉输入端的内容是针对视频片段对话进行抽帧合并的大图，针对视觉端的提示词设计，除了给予目标对话文本、目标说话人、目标对话情感变化及其上下文内容这些辅助信息外，还让大模型去关注之前文本推理生成的文本段刺激的内容。数据集当中文本段刺激的内容与视觉段刺激的内容往往是存在相互关联，是文本刺激内容的视觉体现，且内容基本与目标视频片段内的动作内容息息相关，这样做有利于让模型去更好地关注目标时间段内的人物的神情动作而非不相干的人物动作。

在最后的两个模态内容的融合，由于视觉模态推理过程中关注了文本模态的刺激内容，对于大多视觉模态刺激实际上为不存在的数据样本来说，大模型会倾向于生成与文本刺激相同内容的描述作为视觉刺激的内容，因此通过计算两个模态刺激内容的相似值 (METEOR 指标) 来去除那些过于相似的描述内容。

## 2 数据处理

本次测评任务使用的数据集为 INSIDE (emotion change reasoning in multimodal conversations) 数据集，其内容为 M3ED 数据集<sup>[8]</sup>的基础上进一步标注而成。

表 1 数据分布

|      | 训练集   | 验证集  | 测试集  |
|------|-------|------|------|
| 样本数量 | 2892  | 838  | 417  |
| 对话轮数 | 17997 | 5177 | 2577 |
| 语句数量 | 26049 | 7563 | 3739 |

数据集样本均为视频片段，需要分别对其文本标注内容以及原版视频作处理，分别应对文本推理和视觉推理两个部分。在文本标注内容方面，通过对 utts 部分中相同说话人的句子的合并，并在每一个句子前添加情感标签来使得整个对话片段上下文内容更加易读连贯，而且不会丢失相对应的情感信息，具体的保存格式为：“说话人：(情感标签) 说话内容”。同时，对于要关注的情感变化片段，将标签化的信息内容转化为具体的“对话内容 -> 对话内容”的形式，而非原来的 utts\_id，以便让大模型更好地分析具体的情感变化原因。

针对视频视觉信息的处理本文方法采取以说话人句子为界限进行采帧的方式，在一个视频片段中，对说话人的每一个语句，在这段语句中均匀抽取 15 帧图像并在程序中按时间顺序从左往右合成为一个连续的大图作为每个句子的图像表示。在处理数据集文本模态的时候，将包含有要分析的目标句子的时间区间内所有的句子图像依旧是按照从左到右的顺序合并为大图，最终这张大图就是作为图像分析的输入数据。

## 3 实验

### 3.1 实验设置和结果

在实验的过程中，针对不同的模态分别设置了单文本推理生成 (T)、文本视觉独立推理生成 (VT 独立) 以及文本视觉关联推理生成 (VT 关联) 三种情况。第一种情况是推理过程中直接忽略的视觉模态的刺激这一个只会存在于部分样本的特征，只考虑所有样本都一定会出现的与文本相关联的文本刺激因素 (textual stimulus)、认知评价 (appraisal) 和最终情感反应 (response) 这三个句子。第二种 VT 独立和第三种 VT 关联都是指向是否将文本刺激作为视觉推理部分输入内容这一个核心，前者在视觉推理的过程中不考虑文本刺激的内容，后者则反之，在每次进行视觉推理前将文本推理得到的文本刺激内容填入输入的关键词当中。

最终程序运行三种情况生成推理结果指标如下所示：

表 2 不同模态情况 METEOR 值结果

| 模态    | 验证集    | 测试集    |
|-------|--------|--------|
| T     | 0.3390 | -      |
| VT 独立 | 0.3337 | -      |
| VT 关联 | 0.3433 | 0.3329 |

表 3 不同模态情况 BERTScore 值结果

| 模态    | 验证集    | 测试集    |
|-------|--------|--------|
| T     | 0.7221 | -      |
| VT 独立 | 0.7249 | -      |
| VT 关联 | 0.7242 | 0.7223 |

上述得到的结果均为所有生成样本取平均值。

### 3.2 实验结果分析

通过对比三种情况可以发现，三种情况最终得到的 METEOR 和 BERTScore 值差距不大，单文本模态推理已经基本上能够达到该方法的上限水平，对于文本推理的内容质量决定了整个多模态内容推理质量的基本情况，视觉内容更多地是一种辅助的效果，这也符合数据集样本中普遍出现的视觉模态刺激的缺失的情况。

另外，对比文本视觉是否需要相关联的问题，验证集两者的 METEOR 结果差异说明了文本刺激的内容某种程度上与视觉刺激内容相关联这一个猜想的合理性。在没有文本刺激内容的约束下，视觉推理生成的结果会更加倾向于描述所有细节性的视觉动作，由此带来的无中心的描述句子不好去判断句子是否有效，造成数据样本中原本是少数派的视觉刺激在生成结果中大比例地出现，这反而影响了最终结果的生成质量。

## 4 总结

在本文中提出了文本与视觉双模型推理分析的方法，通过设置相对应的提示词内容，对视频标注文本以及抽帧图片进行推理分析情绪变化的原因。通过对单文本推理生成（T）、文本视觉独立推理生成（VT 独立）以及文本视觉关联推理生成（VT 关联）三种情况进行对比，证实了文本生成内容质量在推理结果质量中的性能核心地位，视觉模态作为补充，其提示词中加入文本刺激可提升推理相关性。最终方案在 CCAC 挑战赛测试集达到 METEOR 0.3329、BERTScore 0.7223，获得了第六名的成绩。

## 参考文献

- [1] Cheng Z, Cheng Z Q, He J Y, et al. Emotion-llama: Mul-timodal emotion recognition and reasoning with instruc-tion tuning[J]. Advances in Neural Information Pro-cessing Systems, 2024, 37: 110805-110853.
- [2] Xu Y, Zhao Z, Sun X, et al. Prompt learning with mul-tiperspective cues for emotional support conversation systems[J]. IEEE Transactions on Computational Social Systems, 2025.
- [3] Geetha A V, Mala T, Priyanka D, et al. Multimodal emo-tion recognition with deep learning: advancements, challenges, and future directions[J]. Information Fusion, 2024, 105: 102218.
- [4] Zou S, Huang X, Shen X. Multimodal prompt trans-former with hybrid contrastive learning for emotion recognition in conversation[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 5994-6003.
- [5] Zhang S, Yang Y, Chen C, et al. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advance-ments and future prospects[J]. Expert Systems with Ap-plications, 2024, 237: 121692.
- [6] Mao R, Liu Q, He K, et al. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection[J]. IEEE transactions on affective computing, 2022, 14(3): 1743-1753.
- [7] Yang A, Li A, Yang B, et al. Qwen3 technical report[J]. arXiv preprint arXiv:2505.09388, 2025.
- [8] Zhao J, Zhang T, Hu J, et al. M3ED: Multi-modal mul-ti-scene multi-label emotional dialogue database[J]. arXiv preprint arXiv:2205.10237, 2022.

## 作者介绍

吴俊阳（2002-），硕士在读，主要研究情感识别。

E-mail: wujunyang666@126.com

王卓（2004-），本科在读，主要研究领域自然语言处理。

E-mail: tgcfbwjj@bupt.edu.cn

田天翊（2004-），本科在读，主要研究领域为自然语言处理。

E-mail: tian\_tianyi@bupt.edu.cn

李启飞（1994-），博士在读，主要研究领域为情感识别、抑郁症检测。

E-mail: liqifei@bupt.edu.cn

练慧俊，（1996-），博士在读，主要研究领域为自然语言处理，人机交互。

E-mail: lianhuijunlybl@bupt.edu.cn

高迎明（1989-），博士，讲师，主要研究领域为语音信息处理、人际交互、情感计算等。

E-mail: yingming.gao@bupt.edu.cn

李雅 (1984-), 博士, 副教授, 主要研究领域为语音交互、多模态情感计算。

E-mail: yli01@bupt.edu.cn

# CCAC2025 评测技术报告： 赛道 2-task1-自主论辩任务

李罗政<sup>1</sup> 陈昌明<sup>1</sup> 李凌波<sup>1</sup>

(1. 中国科学院软件研究所 天基综合信息系统全国重点实验室, 北京市, 100190)

**摘要：** 自主论辩任务涵盖辩题识别、论据挖掘、辩论构建和自主论辩生成四个子任务。针对子任务 1-3, 该方案利用大语言模型进行训练数据增强, 通过 LoRA 微调方法并结合模型集成推理, 以提升模型在各项子任务中的表现。子任务 4 为文本生成任务, 该技术采用多智能体协同框架, 搭建了“文本生成-事实校验-评估打分”的辩论稿生成和优化流程, 以克服大模型幻觉, 并综合集成了在线和离线大模型的最优结果, 最终实现了高质量的辩论陈词稿生成。

**关键词：** 大语言模型; 多智能体; 计算论辩;

**中图分类号：** TP391

**文献标识码：** A

## Technical Report of CCAC2025: Track 2-Task 1- Autonomous Debate Task

Luozheng Li<sup>1</sup>, Changming Chen<sup>1</sup>, and Lingbo Li<sup>1</sup>

(1. Institute of Software, Chinese Academy of Sciences, National Key Laboratory of Space Integrated Information System, Beijing 100190)

**Abstract :** This task includes four sub tasks: topic recognition, evidence mining, debate construction, and independent debate generation. For sub tasks 1-3, this approach utilizes large language models for training data augmentation, and combines LoRA fine-tuning method with model ensemble inference to improve the performance of the model in various sub tasks. Subtask 4 is a text generation task, which adopts a multi-agent collaborative framework to build a debate script generation and optimization process of "text generation fact verification evaluation scoring"

to overcome the illusion of large models and ultimately achieve high-quality debate script generation.

**Key words:** Large Language Model, Multi-Agent System; Computational argumentation;

## 0 引言

正文用“本文”, 务必避免过多第一人称的表述。篇幅在 8000 字左右为宜, 参考文献按文中出现顺序引用。<sup>[1]</sup>

随着大语言模型的快速发展, 自然语言处理 (Neural Language Processing, NLP) 任务正从传统的分类任务向更复杂的理解、推理与生成任务演进。计算论辩作为 NLP 的一个重要分支, 涉及论点识别、结构分析、质量评估及自动生成等多个方面内容。本次评测任务聚焦于“自主论辩”, 基本涵盖了完整的辩论建模流程, 对于建立自动化辩论体系具有重要的启发与指导。

由于本次测评使用的语言均为中文, 本方案选取了在中文任务数据集上展现出了优异成绩的 Qwen2.5<sup>[1]</sup>和 Qwen3<sup>[2]</sup>作为本地大模型, 同时采用 Deepseek-v3<sup>[3]</sup>作为在线大模型。基于本地大模型, 本方案设计并完成了自主辩论中文本分类相关的任务。以往研究通过设计具有不同人格的 agent 进行辩论, 最终实现了文本质量的有效评估<sup>[4]</sup>。在此基础上, 在自主辩论生成 (陈词稿生成) 的任务中, 本方案采用类似的多智能体模型, 通过分工协作完成一辩陈词稿的生成。具体而言, 本方案通过设计“生成-评估-事实核查”的多智能体协同生成框架, 有效纠正了大模型的“事实幻觉”并集成了在线大模型和离线大模型的最优生成结果, 从而实现优质、流畅具有逻辑性的陈词稿生成。

本技术方案的主要贡献如下:

- 1) 本方案采用针对性的数据增强方案和模型集成方案, 有效提升了有限监督数据场景下大模型微调的效果。
- 2) 本方案采用“多智能体协同生成框架”进行陈词稿的自动化撰写和修改, 在保证陈词稿逻辑性的基础上, 有效纠正了大模型的“事实幻觉”问题。

# 1 技术方案介绍

## 1.1 子任务 1-3

### 1.1.1 任务介绍

子任务 1-3 包含有辩题识别、论据挖掘和辩论构建。从任务形式上来看，子任务 1,2 属于二分类任务，子任务 3 属于多分类任务。

其中，辩题识别任务旨在判断给定论点是否属于指定辩题范畴。输入为包含“辩题”和“论点”的 JSON 对象，输出为布尔值 (True/False)。该任务要求模型具备较强的语义理解与匹配能力。论据挖掘任务即判断某条论据是否可以支持给定论点。输入为包含“论点”和“论据”的 JSON 对象，输出为布尔值 (True/False)。任务要求模型具备较强的因果关系判断与语义支撑识别能力。辩论构建任务目标为识别论点在立论陈词稿中的结构类型，包括“主要论点”、“普通论点”和“论点前提”。该任务要求模型具备较强的文本结构分析与上下文理解能力。

### 1.1.2 流程设计

基于以上任务特点，本方案使用本地大模型进行数据增强和微调，具体流程如下图 1 所示。

整体上看，每个子任务都采用“数据增强+模型集成”的策略。具体每个子任务的具体操作方式略有不同。

**数据增强。**子任务 1-3 均为文本分类任务，模型的表现一定程度上取决于训练数据的规模和多样性，因此，我们设计了数据增强使用 QWQ-32B-AWQ 模型对原始训练样本进行同义改写，生成多样化的正例；同时构造负例样本，模拟偏离主题的表达方式，提升模型泛化能力。

正例增强策略为同义句改写，包括近义词替换、主被动转换、句法结构重组等。

负例增强策略则根据子任务进行单独设计。在辩题识别任务和辩论构建任务中，负例增强方法与正例增强相同，即对相应标签的训练样例进行同义句改写。在论据挖掘任务中，为了防止模型过度依赖关键词特征，负例创建策略包含同论点负样本生成、跨主题负样本生成和随机采样生成。

**模型微调与集成。**对于以上任务，本方案使用 LLaMA-Factory 进行了多个本地模型 (Qwen2.5-7B-Instruct、Qwen3-8B 等) 的 LoRA 微调，采用多数投票机制生成最终预测结果，提高整体鲁棒性。具体见表 1 所示。

表 1: 集成模型详情

| 模型编号    | 基础模型                | 增强数据量   |
|---------|---------------------|---------|
| Model A | Qwen2.5-7B-Instruct | 262,393 |
| Model B | Qwen2.5-7B-Instruct | 19,714  |
| Model C | Qwen3-8B            | 19,709  |

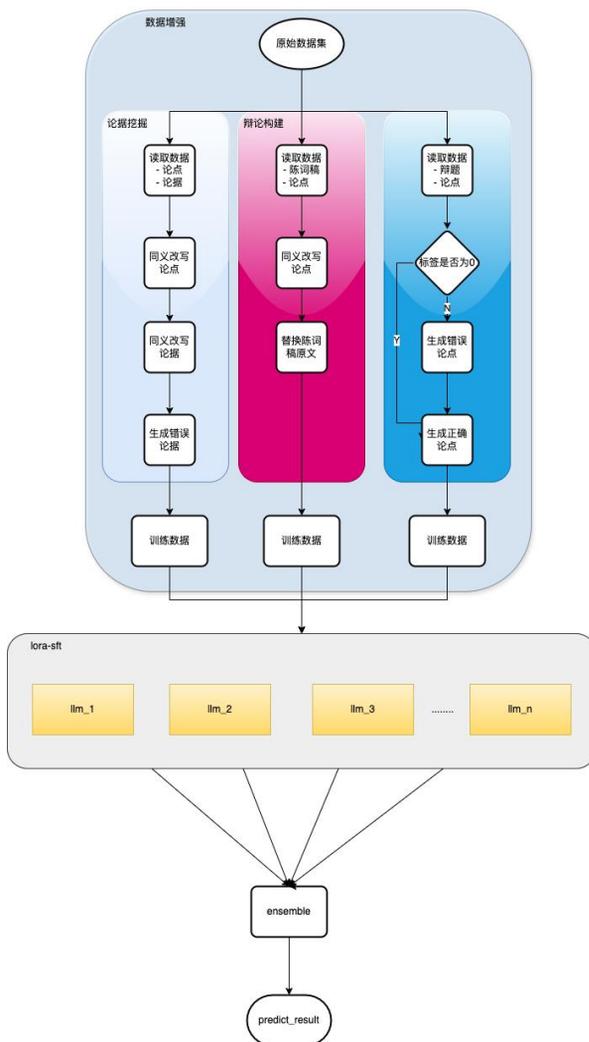


图 1 子任务 1-3 流程设计图

## 1.2 子任务 4

### 1.2.1 任务介绍

本子任务给定辩题及对应持方，由模型自动生成立论陈词稿。训练集数据仅有 338 条，考虑到该

训练数据规模过小，即使使用大模型 Lora 微调技术，也不足以训练得到能够完成文本生成任务的模型。因此，本方案并未使用训练集数据，而是结合在线和本地大语言模型，搭建了多智能体协同写作框架，用以完成自主论辩的生成。

### 1.2.2 模型设计

多智能体协同生成框架设计如图 2 所示，整体可以分为 Generator，Fact Checker 和 Evaluator 三个智能体。其中，Generator（生成器）由开源（Qwen2.5-14B-Int8）和闭源(Deepseek-v3)两个大模型为后端的智能体组成，基于给定辩题与持方，结合 Prompt-1 生成陈词稿初稿，为后续流程提供基础素材。Fact Checker（事实核查器）的功能包括事实抽取、事实核查和内容修订。即从生成器产出内容里提取包含的事实信息，明确需核查的事实点。然后借助搜索引擎 API，搜索出与待核查事实相关的内容作为候选内容。最后，结合搜索网页得到的相似内容（可认为更具有真实性）和原始陈词稿进行内容改写，对内容中事实性问题进行修正，输出更可靠的辩论素材。Evaluator（评估器）即结合“Prompt-2（评分提示词）”，对经事实核查后的两模型生成内容（来自不同生成器路径），从辩论逻辑、说服力等维度评估，并形成总分。最终汇总，选取得分更高的陈词稿内容作为“辩论稿（最终结果）”。

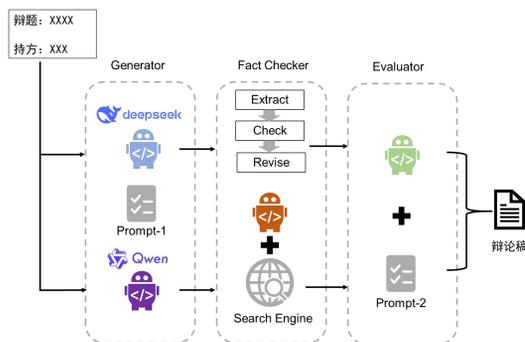


图 2 子任务 4 模型设计图

## 2 实验结果

### 2.1 子任务 1-3

按照任务要求，本技术方案选用准确率（acc）

作为评价指标。随机选取训练集的 80% 数据进行微调，剩余 20% 数据作为验证集。本技术方案的表现如表 2 所示

表 2：子任务 1-3 准确率（Acc）

| 任务名称 | 准确率 1  | 准确率 2  |
|------|--------|--------|
| 辩题识别 | 93.88% | 98.87% |
| 论据挖掘 | 67.59% | 73.27% |
| 辩论构建 | 84.17% | 94.37% |

其中，准确率 1 表示仅使用 Lora 微调本地大模型的准确率，并未使用数据增强和模型集成。准确率 2 表示使用本技术方案后集成的准确率。需要指出，最终提交的版本使用了全部的训练数据，因此最终的准确率与表 2 中的数据可能略有出入。

### 2.1 子任务 4

按照测评规定，子任务 4 由人工进行打分。为了模拟人工打分，本技术方案从完整性、逻辑性、创新性等维度归纳总结，形成 Prompt-2（评分提示词）。然后基于提示词和在线大模型 Deepseek-v3，形成评智能体，如图 2 所示，模拟专家进行陈词稿的打分评估。

使用该评估智能体，本方案可进行陈词稿生成的评估比较。本方案集成了在线大模型 Deepseek-v3 和本地大模型 Qwen2.5-14B，并选取其中评估更高的答案最为最终结果。

本方案统计了 20 个测试集的平均评分，结果如表 3 所示。

表 3 子任务 4 方案结果

| 技术方案        | 评分（满分 100） |
|-------------|------------|
| 本方案         | 91.85      |
| Deepseek-v3 | 89.03      |
| Qwen2.5-14B | 87.45      |

结果显示，本技术方案通过多智能体的设计，可进行有效的事实性知识改写，最终的评分高于 Deepseek-v3。

## 3 总结

本团队围绕 CCAC2025 评测赛道 2-task1 “自主论辩任务”，针对辩题识别、论据挖掘、辩论构建三个子任务设计了统一的数据增强-推理集成的解决方案，为自主论辩生成任务设计了多智能体协同生成框架。本技术方案综合使用 Prompt 引导、数据增强、模型微调与集成投票等手段，使大模型更适应相应的子任务，进而提升整体的表现。特别地，本方案通过多智能体协同的事实核查，最大限度降低了大模型的幻觉问题，进而增强论据的说服力，以期得到更优的自主生成结果。

## 参考文献

- [1] Yang A , Yang B , et al. Qwen2. 5 Technical Report[J]. arXiv:2412. 15115, 2024.
- [2] Yang A, Li A, Yang B, et al. Qwen3 technical report[J]. arxiv preprint arxiv:2505.09388, 2025.
- [3] Liu, Aixin, et al. Deepseek-v3 technical report[J]. arxiv preprint arxiv:2412.19437 2024.
- [4] Chan, Chi-Min, et al. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate[C]// The Twelfth International Conference on Learning Representations, 2024.

# 第五届智慧论辩评测（AI-Debater 2025）辩题识别

张宸华<sup>1</sup> 靳伯骏<sup>1</sup> 王华杰<sup>1</sup> 黄煜聪<sup>1</sup> 陈冠荣<sup>1</sup> 何宇航<sup>1</sup> 孙洋<sup>1</sup>  
鲍建竹<sup>1</sup> 徐睿峰<sup>1</sup>

(1. 哈尔滨工业大学（深圳），广东省 深圳市 518055)

**摘要：**辩题识别是实现智慧辩论不可或缺的基础任务，其目的在于确保论点始终围绕辩题。近年来，大语言模型大量的知识储备和强大的推理能力使其有一定的逻辑判断能力，但是，现有的主流开源模型为通用模型，对于论辩领域论点和辩题契合度的判断针对性并不强。因此，本工作以 Qwen3-14B 为基座模型，引入官方论辩数据与合成数据，采用 LoRA 方法进行高效微调。实验结果表明，本方案在官方验证集上优于基础模型，取得了当前任务中的最优表现。

**关键词：**计算论辩；大语言模型；合成数据

中图分类号：TP391

文献标识码：A

## Argument topic identification in AI-Debater 2025

Chenhua Zhang<sup>1</sup>, Bojun Jin<sup>1</sup>, Huajie Wang<sup>1</sup>, Yucong Huang<sup>1</sup>, Guanrong Chen<sup>1</sup>, Yuhang He<sup>1</sup>, Yang Sun<sup>1</sup>,  
Jianzhu Bao<sup>1</sup> and Ruifeng Xu<sup>1</sup>

(1. Harbin Institute of Technology, Shenzhen, Shenzhen, Guangdong 518005, China)

**Abstract :** Argument topic identification is a fundamental task essential for enabling intelligent debating, aiming to ensure that arguments consistently align with the given topic. In recent years, large language models have demonstrated a certain degree of logical reasoning capability due to their extensive knowledge base and powerful inference abilities. However, mainstream open-source models are generally designed for broad applications and lack specificity in evaluating the alignment between arguments and topics within the debate domain. Therefore, this work employs Qwen3-14B as the base model, incorporating both official debate datasets and synthetic data, and utilizes the LoRA technique for efficient fine-tuning. Experimental results show that the proposed approach outperforms the base model on the official validation set, achieving state-of-the-art performance for this task.

**Key words:** Computational Argumentation; Large Language Models; Synthetic Data

## 0 引言

在人类社会的诸多交流形式中，论辩作为表达立场、传递观点、影响他人决策的重要方式，长久以来被视作智慧与理性的象征。近年来，计

收稿日期：20XX-XX-XX；定稿日期：20XX-XX-XX

基金项目：基金名（基金号）；基金名(基金号)

算论辩 (Computational Argumentation) 逐渐成为自然语言处理领域的研究焦点，其核心目标是让机器具备理解、生成和组织论辩内容的能力，助力其在智能决策、教育辅导等场景中落地应用。

在此背景下，第五届智慧论辩评测围绕自主论辩任务、自主论辩赋能社会调查两大任务展开，其中自主论辩任务包含多个关键子任务，如辩题识别、论据挖掘、辩论构建与自主论辩生成。在本次比赛中，官方在 PE 论辩挖掘数据集的基础上通过 DeepSeek-R1 等大模型翻译和人工校验，构建了网络媒体内容辩论理解、自主辩论生成、社会调查仿真等一系列论辩数据集。

本报告聚焦于辩题识别子任务：此任务旨在判断辩题和论点之间的相关性，是构建辩论的最基础环节。随着以 GPT-4 为代表的大型语言模型 (Large Language Models, LLM) 的广泛应用，模型在自然语言生成与理解和逻辑推理能力取得显著进步。但当前主流开源大模型 (如 LLaMA、Qwen、Deepseek 等) 多面向通用任务设计，虽具备较强的多任务泛化能力，在处理需复杂逻辑推理的论辩任务时却存在一定局限。

为此，本工作在通用模型基础上，利用官方提供的论辩数据集及基于大模型的合成数据进行 LoRA 微调，以提升模型对辩题识别任务的适配性。实验结果表明，该方案相较于官方基准模型有显著提升，模型参数已经开源至 huggingface<sup>1</sup>。

## 1 相关工作

近年来，针对自主辩论任务中训练数据不足或结构复杂性高的问题，部分研究开始尝试引入合成数据 (Synthetic Data) 以提升模型的论辩能力和结构感知能力。在 CCAC 2023 的 AI-Debater2023 任务中，HITSZ-HLT 在使用了合成数据来补充数据集，从而提升了自主辩论任务的表现效果，这验证了合成数据在自主辩论任务中能够增强模型语言理解能力、缓解数据稀缺和提升结构建模性能方面的有效性<sup>[1]</sup>。

## 2 技术方案介绍

本方法的框架示意图如图 1 所示，其主要包含两部分：基于大模型生成的数据合成 (Synthetic Data Generation, SDG)、数据构建 (Data Construction, DC)、基于合成数据的有监督微调 (Supervised Fine-Tuning, SFT)。

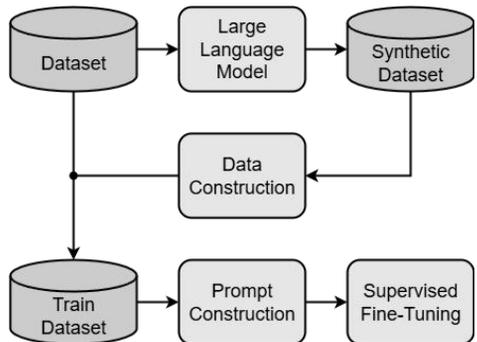


图 1: 本工作框架图

### 2.1 基于大模型生成的数据合成

本任务基于 Task 3 的合成数据，针对 Task 1 构建数据。

### 2.2 数据构建

首先从 Task 3 的合成数据中排除验证集的辩题，保证合成数据中不存在验证集的辩题，确保验证时的公平性。下面介绍 Task 1 的数据构造方法：

**正样本构造：**对于合成数据中的辩题下的主要论点，普通论点，论点前提构造为正样本。

**负样本构造：**随机替换正样本的辩题为另一辩题或论点替换为另一辩题的论点。

在构造过程中保证正负样本的数量均衡性。辩题识别构造样本示例如表 1 所示。

表 1: 辩题识别构造样本示例

| 辩题              | 论点                                    | 标签 |
|-----------------|---------------------------------------|----|
| 人类是否应该进行外太空探索。  | 由于宇航员要在无氧的太空中探险，科学家们被驱动去创新氧气罐。        | 1  |
| 学生是否应该被禁止玩电子游戏。 | 大多数学生屈服于电脑游戏的吸引力，牺牲了本可以用于学习、睡眠和体育活动的时 | 1  |

<sup>1</sup> [https://huggingface.co/dagshfqb/qwen14B\\_ccac2025/tree/main](https://huggingface.co/dagshfqb/qwen14B_ccac2025/tree/main)

|               |                           |   |
|---------------|---------------------------|---|
| 间。            |                           |   |
| 为自己工作还是为雇主工作。 | 许多人可能会来自其他社区，寻求在大学的教育或就业。 | 0 |
| 科技进程的负面影响。    | 全球化在推动各领域进步方面发挥了重要作用。     | 0 |

### 2.3 基于合成数据的有监督微调

为了进一步提升模型在辩论构建任务中的表现，本工作结合官方提供的辩论数据集以及基于大型语言模型生成的合成数据（Synthetic Data），对模型进行了有针对性的微调。

在辩题识别子任务中，每条训练数据由三个部分组成：辩题、论点以及对应标签。模型的目标是根据给定的辩题和论点，判断论点是否属于辩题。

在微调过程中，我们采用 LoRA（Low-Rank Adaptation）技术进行高效的有监督训练，并设计了如下形式的提示词（prompt）作为模型输入，使用的 prompt 如下表 2 所示：

|  |
|--|
| <p>你是一位出色的辩论专家。给定一个辩题和一个论点，你需要判断给定的论点是否属于这个辩题。</p> <p>## 辩题：<br/>{topic}</p> <p>## 论点：<br/>{argument}</p> <p>请你回答此论点是否属于此辩题(属于/不属于)：</p> |
|--|

表 2: LoRA 微调过程中使用的 prompt

## 3 实验设置

本工作初始尝试过的基座模型有 Qwen3-8B 和 Qwen3-14B。经过我们的测试，发现在 Qwen3-14B 上辩论构建任务的性能最佳，故选择 Qwen3-14B 进行实验。

在 LoRA 的配置方面，设置 `lora_r=16`，`lora_alpha=32`，`dropout=0.1`，

`target_modules=["q_proj", "k_proj", "v_proj", "o_proj"]`。

在训练配置方面，设置 `epoch=2`，`batch_size=16`，`gradient_accumulation_steps=2`，`learning_rate=2e-4`，`warmup_ratio=0.0`，`weight_decay=0.01`，最大输入长度 200。

## 4 实验结果分析

本工作所有的推理操作都按照图 1 的形式构造 prompt。实验部分将从性能指标和生成样例两个角度来分析。

在实验过程中，我们将官方提供的辩论构建数据集按照 9:1 的比例划分为训练集和验证集，并将合成数据合并至训练集中。本实验在训练集中严格排除验证集和测试集的样本数据，确保训练集中不包含用于评估的论点，以避免数据泄漏。随后使用训练集进行模型的微调训练，并在验证集上进行性能评估。评估指标采用准确率。

为进一步评估合成数据的有效性，我们在对比实验中引入了两组设置：一组仅使用官方数据进行训练作为 baseline，另一组在原始训练集的基础上加入由大模型生成的合成数据进行扩充。

实验结果如表 3 所示，可以看到，Qwen3-14B + LoRA + 合成数据表现最优。值得注意的是，合成数据对于任务一最终性能有一定提升，但提升不明显。我们认为通过合成更多样化的数据，能使模型熟悉丰富多样的语言表达和论辩风格，进而提升模型的性能；但是，合成数据依然面临引入低质量样本的问题，限制了模型性能的进一步提升。

表 3: 不同模型/方法在辩题识别任务中的表现

| 使用的模型/方法                      | 准确率 (%) |
|-------------------------------|---------|
| bert-base-chinese Fine-Tuning | 87.2    |
| Qwen3-8B LoRA                 | 89.5    |
| Qwen3-14B LoRA                | 90.1    |
| Qwen3-14B LoRA+合成数据           | 90.6    |

## 5 总结

针对现有主流开源通用模型不能很好地满足特定领域的特定需求的问题，本工作通过在合成数据上的微调 lora 方法，使得模型能够更好地适配辩题识别任务。实验结果表明了合成数据的高质量和有效性。

## 参考文献

- [1] Jiayu Lin, Guanrong Chen, Bojun Jin, et al. Overview of AI-Debater 2023: The Challenges of Argument Generation Tasks, arXiv, 2023, arXiv:2407.14829

# 第五届智慧论辩评测（AI-Debater 2025）论据挖掘

张宸华<sup>1</sup> 靳伯骏<sup>1</sup> 王华杰<sup>1</sup> 黄煜聪<sup>1</sup> 陈冠荣<sup>1</sup> 何宇航<sup>1</sup> 孙洋<sup>1</sup>  
鲍建竹<sup>1</sup> 徐睿峰<sup>1</sup>

(1. 哈尔滨工业大学（深圳），广东省 深圳市 518055)

**摘要：**论据挖掘是实现智慧辩论不可或缺的基础任务，其目的在于确保论据始终支持论点。近年来，大语言模型大量的知识储备和强大的推理能力使其有一定的逻辑判断能力，但是，现有的主流开源模型为通用模型，对于论辩领域论据和论点是否支持的判断针对性并不强。因此，本工作以 Qwen3-14B 为基座模型，引入官方论辩数据与合成数据，采用 LoRA 方法进行高效微调。实验结果表明，本方案在官方验证集上显著优于未微调模型与无合成数据的微调方案，取得了当前任务中的最优表现。

**关键词：**计算论辩；大语言模型；合成数据

中图分类号：TP391

文献标识码：A

## Evidence mining in AI-Debater 2025

Chenhua Zhang<sup>1</sup>, Bojun Jin<sup>1</sup>, Huajie Wang<sup>1</sup>, Yucong Huang<sup>1</sup>, Guanrong Chen<sup>1</sup>, Yuhang He<sup>1</sup>, Yang Sun<sup>1</sup>,  
Jianzhu Bao<sup>1</sup> and Ruifeng Xu<sup>1</sup>

(1. Harbin Institute of Technology, Shenzhen, Shenzhen, Guangdong 518005, China)

**Abstract :** Evidence mining is a fundamental task essential for enabling intelligent debating, aiming to ensure that supporting arguments consistently align with their corresponding claims. In recent years, large language models have demonstrated a certain degree of logical reasoning capability due to their extensive knowledge base and powerful inference abilities. However, mainstream open-source models are typically designed for general-purpose applications and lack specificity in assessing the support relationship between arguments and claims within the debate domain. To address this limitation, this work employs Qwen3-14B as the base model, incorporating both official debate datasets and synthetic data, and adopts the LoRA technique for efficient fine-tuning. Experimental results show that the proposed approach significantly outperforms both the non-fine-tuned model and fine-tuning without synthetic data on the official validation set.

**Key words:** Computational Argumentation; Large Language Models; Synthetic Data

## 0 引言

在人类社会的诸多交流形式中，论辩作为表达立场、传递观点、影响他人决策的重要方式，长久以来被视作智慧与理性的象征。近年来，计

收稿日期：20XX-XX-XX；定稿日期：20XX-XX-XX

基金项目：基金名（基金号）；基金名(基金号)

算论辩 (Computational Argumentation) 逐渐成为自然语言处理领域的研究焦点, 其核心目标是让机器具备理解、生成和组织论辩内容的的能力, 助力其在智能决策、教育辅导等场景中落地应用。

在此背景下, 第五届智慧论辩评测围绕自主论辩任务、自主论辩赋能社会调查两大任务展开, 其中自主论辩任务包含多个关键子任务, 如辩题识别、论据挖掘、辩论构建与自主论辩生成。在本次比赛中, 官方在 PE 论辩挖掘数据集的基础上通过 DeepSeek-R1 等大模型翻译和人工校验, 构建了网络媒体内容辩论理解、自主辩论生成、社会调查仿真等一系列论辩数据集。

本报告聚焦于论据挖掘子任务: 此任务旨在判断论据是否支持论点, 是构建辩论的重要环节之一。随着以 GPT-4 为代表的大型语言模型 (Large Language Models, LLM) 的广泛应用, 模型在自然语言生成与理解和逻辑推理能力取得显著进步。但当前主流开源大模型 (如 LLaMA、Qwen、Deepseek 等) 多面向通用任务设计, 虽具备较强的多任务泛化能力, 在处理需复杂逻辑推理的论辩任务时却存在一定局限。

为此, 本工作在通用模型基础上, 利用官方提供的论辩数据集及基于大模型的合成数据进行 LoRA 微调, 以提升模型对论据挖掘任务的适配性。实验结果表明, 该方案相较于官方基准模型有显著提升, 模型参数已经开源至 huggingface<sup>1</sup>。

## 1 相关工作

近年来, 针对自主辩论任务中训练数据不足或结构复杂性高的问题, 部分研究开始尝试引入合成数据 (Synthetic Data) 以提升模型的论辩能力和结构感知能力。在 CCAC 2023 的 AI-Debater2023 任务中, HITSZ-HLT 在使用了合成数据来补充数据集, 从而提升了自主辩论任务的表现效果, 这验证了合成数据在自主辩论任务当中能够增强模型语言理解能力、缓解数据稀缺和提升构建模性能方面的有效性<sup>[1]</sup>。

## 2 技术方案介绍

本方法的框架示意图如图 1 所示, 其主要包含两部分: 基于大模型生成的数据合成 (Synthetic Data Generation, SDG)、数据构建 (Data Construction, DC)、基于合成数据的有监督微调 (Supervised Fine-Tuning, SFT)。

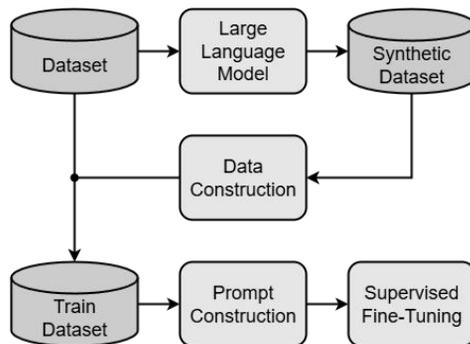


图 1: 本工作框架图

### 2.1 基于大模型生成的数据合成

本任务基于 Task 3 的合成数据, 针对 Task 2 构建数据。

### 2.2 数据构建

首先从 Task 3 的合成数据中排除验证集的论点, 保证合成数据中不存在验证集的论点, 确保验证时的公平性。下面介绍 Task 2 的数据构建方法:

合成数据每一个样本可以看成图结构, 其中结点为论元组件 (主要论点, 普通论点, 论点前提), 边为结点之间是否支持 (source, target, Support/Attack)。

正样本构造: 从边集找到 support 的边, 以此构造正样本

负样本构造:

1) 从边集找到 Attack 的边, 以此构造负样本

2) 一个普通论点的论点前提和另一个普通论点构造负样本

3) 不同普通论点的论点前提构造负样本

在构造过程中保证正负样本的数量均衡性。

论据挖掘构造样本示例如表 1 所示。

1

表 1: 论据挖掘样本示例

| 论点                                     | 论据                                  | 标签 |
|--|-------------------------------------|----|
| 面对面工作可以间接提高项目的质量和效率。                   | 误解减少了，冲突也随之减少，促进了团队成员之间更好的沟通。       | 1  |
| 科技使得个人能够与彼此保持更密切的联系。                   | 人们利用互联网或智能手机在不同地点之间分享信息。            | 1  |
| 城市的医疗技术远远优于乡村，我相信这一好处大于科技对人类可能产生的负面影响。 | 因此，建议退休人员去公司或医院与现有员工进行定期交流，填补知识的差距。 | 0  |
| 另一种观点认为，隐瞒信息可以帮助防止不必要的恐慌和社会动荡。         | 政府应确保在灾难或疫情爆发期间及时向媒体披露关键信息。         | 0  |

### 2.3 基于合成数据的有监督微调

为了进一步提升模型在辩论构建任务中的表现，本工作结合官方提供的辩论数据集以及基于大型语言模型生成的合成数据（Synthetic Data），对模型进行了有针对性的微调。

在论据挖掘子任务中，每条训练数据由三个部分组成：论点、论据以及对应标签。模型的目标是根据给定的论点和论据，判断论据是否支持论点。

在微调过程中，我们采用 LoRA（Low-Rank Adaptation）技术进行高效的有监督训练，并设计了如下形式的提示词（prompt）作为模型输入，使用的 prompt 如下表 2 所示：

|   |
|---|
| <p>你是一位出色的辩论专家。给定一个论点和一个论据，你需要判断给定的论据是否支持这个论点。</p> <p>## 论点：<br/>{argument}</p> <p>## 论据：<br/>{evidence}</p> <p>请你回答此论据是否属于此论点(支持/不支持)：</p> |
|---|

表 2: LoRA 微调过程中使用的 prompt

## 3 实验设置

本工作初始尝试过的基座模型选择 Qwen3-14B 进行实验。

在 LoRA 的配置方面，设置  $lora\_r=16$ ， $lora\_alpha=32$ ， $dropout=0.1$ ， $target\_modules=["q\_proj", "k\_proj", "v\_proj", "o\_proj"]$ 。

在训练配置方面，设置  $epoch=2$ ， $batch\_size=16$ ， $gradient\_accumulation\_steps=2$ ， $learning\ rate=2e-4$ ， $warmup\_ratio=0.0$ ， $weight\_decay=0.01$ ，最大输入长度 200。

## 4 实验结果分析

本工作所有的推理操作都按照表 2 的形式构造 prompt。实验部分将从性能指标和生成样例两个角度来分析。

在实验过程中，我们将官方提供的辩论构建数据集按照 9:1 的比例划分为训练集和验证集，并将合成数据合并至训练集中。本实验在训练集中严格排除验证集和测试集的样本数据，确保训练集中不包含用于评估的论点，以避免数据泄漏。随后使用训练集进行模型的微调训练，并在验证集上进行性能评估。评估指标采用准确率。

为进一步评估合成数据的有效性，我们在对比实验中引入了两组设置：一组仅使用官方数据进行训练作为 baseline，另一组在原始训练集的基础上加入由大模型生成的合成数据进行扩充。

实验结果如表 3 所示，可以看到，Qwen3-14B + LoRA + 合成数据表现最优。这一结果验证了合成数据在提升模型在论据对论点支持逻辑判断能力方面的有效性，表明了任务二由于更加复杂，更多的数据让模型学到的判断模式更加有效，性能有所提升，也表明通过合成额外的数据是提升自主辩论系统性能的可行路径。

表 3: 不同模型/方法在论据挖掘任务中的表现

| 使用的模型/方法                         | 准确率 (%) |
|----------------------------------|---------|
| bert-base-chinese<br>Fine-Tuning | 68.2    |
| Qwen3-14B LoRA                   | 71.1    |

## 5 总结

针对现有主流开源通用模型不能很好地满足特定领域的特定需求的问题，本工作通过在合成数据上的 lora 微调方法，使得模型能够更好地适配论据挖掘任务。实验结果表明了合成数据的高质量和有效性。

## 参考文献

- [1] Jiayu Lin, Guanrong Chen, Bojun Jin, et al. Overview of AI-Debater 2023: The Challenges of Argument Generation Tasks, arXiv, 2023, arXiv:2407.14829

# 第五届智慧论辩评测（AI-Debater 2025）辩论构建生成技术报告

黄煜聪<sup>1</sup> 靳伯骏<sup>1</sup> 王华杰<sup>1</sup> 张宸华<sup>1</sup> 陈冠荣<sup>1</sup> 何宇航<sup>1</sup> 孙洋<sup>1</sup>  
鲍建竹<sup>1</sup> 徐睿峰<sup>1</sup>

(1. 哈尔滨工业大学（深圳），广东省 深圳市 518055)

**摘要：**辩论构建是计算论辩任务中的关键一环，其目标是识别立论陈词稿中各个论点在整体结构中的角色类型。随着大型语言模型的发展，模型在语言理解与结构建模方面具备了更强的能力，但主流开源大模型（如 Qwen、LLaMA 等）面向通用任务，针对辩论构建任务的适应性仍有限。因此，本工作以 Qwen3-14B 为基座模型，引入官方论辩数据与合成数据，采用 LoRA 方法进行高效微调。实验结果表明，本方案在官方验证集上显著优于未微调模型与无合成数据的微调方案，取得了当前任务中的优异表现。

**关键词：**计算论辩；大语言模型；合成数据

中图分类号：TP391

文献标识码：A

## The technical report of debate construction generation in AI-Debater 2025

Yucong Huang<sup>1</sup>, Bojun Jin<sup>1</sup>, Huajie Wang<sup>1</sup>, Chenhua Zhang<sup>1</sup>, Guanrong Chen<sup>1</sup>, Yuhang He<sup>1</sup>, Yang Sun<sup>1</sup>, Jianzhu Bao<sup>1</sup> and Ruifeng Xu<sup>1</sup>

(1. Harbin Institute of Technology, Shenzhen, Shenzhen, Guangdong 518005, China)

**Abstract :** Argument mining is a key subtask in computational argumentation, aiming to identify the functional roles of individual argumentative units within an overall structure. With the rapid advancement of large language models (LLMs), their capabilities in language understanding and structural modeling have significantly improved. However, mainstream open-source LLMs such as Qwen and LLaMA are generally optimized for broad tasks and still exhibit limited adaptability to the specific needs of argument structure prediction. In this study, we fine-tune the Qwen3-14B model using a parameter-efficient LoRA approach, incorporating both official argumentation datasets and synthetic data. Experimental results demonstrate that our approach significantly outperforms both the base model and fine-tuning without synthetic data on the official validation set, achieving excellent performance on this task.

**Key words:** Computational Argumentation; Large Language Models; Synthetic Data

## 0 引言

在人类社会的多种交流形式中，论辩作为表

达立场、传递观点、影响他人决策的重要手段，长期以来被认为是智慧与理性的体现。近年来，计算论辩（Computational Argumentation）逐渐成为自然语言处理领域的研究热点，其核心目

标是赋予机器理解、生成和组织论辩内容的能力，推动其在智能决策、舆论分析、教育辅导等场景中的应用落地。

在此背景下，第五届智慧论辩评测聚焦于自主论辩任务、自主论辩赋能社会调查两个任务，研究者在自主论辩任务中提出了多个关键子任务，包括辩题识别、论据挖掘、辩论构建和自主论辩生成。在本次报告中，我们聚焦于辩论构建这一子任务，该任务使用一篇立论陈词稿和陈词稿中的一个论点，判断当前论点在陈词稿中的结构类型，其中结构类型包括主要论点、普通论点和论点前提。

随着以 GPT-4 为代表的大型语言模型 (Large Language Models, LLM) 的广泛应用，模型在自然语言生成与理解方面取得了显著进展，极大提升了论辩类任务中语言生成的质量与逻辑性。然而，当前主流开源大模型 (如 LLaMA、Qwen、Deepseek 等) 普遍面向通用任务设计，虽具备较强的多任务泛化能力，但在面对具有复杂结构与逻辑推理要求的论辩任务时仍存在一定局限。所以，本工作在通用模型的基础上，使用官方所提供的论辩数据集和基于大模型的合成数据进行 LoRA 微调<sup>[1]</sup>，使得模型能够更好地适配于辩论构建任务。实验结果表明，本工作的方案相较于官方提供的基准模型有了很大的提升，模型参数已经开源至 huggingface<sup>1</sup>。

## 1 相关工作

近年来，针对自主辩论任务中训练数据不足或结构复杂性高的问题，部分研究开始尝试引入合成数据 (Synthetic Data) 以提升模型的论辩能力和结构感知能力。在 CCAC 2023 的 AI-Debater2023 任务中，HITSZ-HLT 在使用了合成数据来补充数据集，从而提升了自主辩论任务的表现效果，这验证了合成数据在自主辩论任务中能够增强模型语言理解能力、缓解数据稀缺和提升结构建模性能方面的有效性<sup>[2]</sup>。

## 2 技术方案介绍

本工作的框架示意图如图 1 所示，其主要包含两部分：基于大模型生成的数据合成 (Synthetic Data Generation, SDG)、基于合

成数据的有监督微调 (Supervised Fine-Tuning, SFT)。

### 2.1 基于大模型生成的数据合成

在论证挖掘 (Argument Mining, AM) 领域，数据稀缺是阻碍模型性能提升的主要瓶颈。手动标注结构化论证数据成本高昂且耗时，导致现有数据集规模有限，严重制约了模型的训练效果。为解决这一问题，我们提出了一种专注于生成高质量标注数据的方法。该方法通过结构感知的重构技术，确保合成数据在保留原始语义和标注完整性的同时，避免引入标签噪声，从而显著提升模型鲁棒性。本工作的框架示意图如图 1 所示。

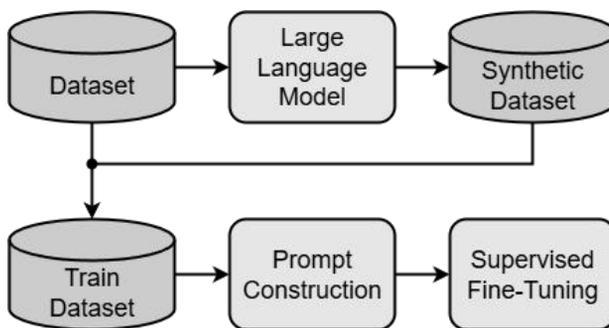


图 1: 本工作框架图

#### 2.1.1 方法概述

论证挖掘任务的核心挑战在于其复杂标注结构 (包括论元组件的边界、类型及关系)，这使得传统数据增强方法难以维持标注一致性。直接使用大型语言模型 (LLM) 生成新文本和标注时，往往出现标注偏差或语义失真，导致训练数据质量下降。为此，我们设计了一种质量优先的合成策略，其核心思想包括：

**缓解数据稀缺问题：**利用现有标注数据生成高质量合成实例，扩充训练集规模，效果显著。

**确保标注高保真：**通过结构约束的生成过程，严格继承原始标签，避免自动标注引入的噪声，为模型提供可靠的学习信号。

**优化模型泛化：**在保持论证逻辑不变的前提下，增加文本的词汇和句法多样性，增强模型对语言变化的适应能力。

#### 2.1.2 方法介绍

该方法的核心是“结构感知的重构”技术，通过精心设计的提示工程指导 LLM 生成高质量合成数据。以下是具体实现步骤：

##### (1) 输入格式设计

1

将原始训练实例格式化为结构化 JSON 对象，包括两个关键部分：

(a) 上下文文本：包含占位符（如 [AC1], [AC2]）标记论元组件的位置。

(b) 论元组件信息：提供每个组件的文本内容和预定义类型（如 “Claim” 或 “Premise”）。

示例输入格式如表 1 所示。

```
{
  "context": "Television is one of the
greatest innovations [AC1]. Besides,
watching television [AC2].",
  "argument_component_info": {
    "AC1": {"content": "that we use
every day", "type": "MajorClaim"},
    "AC2": {"content": "has some cons
and at the same time has very good pros",
"type": "Claim"}
  }
}
```

表 1: 数据合成过程中使用的提示词

## (2) 结构感知的重构过程

通过提示指令，要求 LLM 对上下文文本和每个论元组件内容进行重构，同时严格遵守以下规则：

(a) 语义一致性：保留原始论证逻辑和核心含义。

(b) 类型不变性：确保论元组件类型（如 “Claim” 或 “Premise”）与原始标注完全一致。

(c) 边界完整性：占位符设计保证组件文本边界在重构后不发生偏移，避免标注错位问题。例如，LLM 生成的新 JSON 输出会重构上下文和组件内容，但组件类型直接继承原始标签。

## (3) 标注继承与数据重建

(a) 从 LLM 的 JSON 输出中提取重构后的上下文和组件内容，将组件文本插入对应占位符，重建完整的新辩论文本。

(b) 论元组件的边界（起止索引）直接从占位符位置推导。

(c) 组件类型和论元关系标签完全继承原始实例的标注，无需 LLM 额外生成关系（因关系逻辑隐含在结构中）。

## (4) 输出结果示例

原始文本：“Television is one of the greatest innovations that we use every day.”

合成文本：“Television is undoubtedly one of

the most significant innovations we engage with daily.”

其中，组件类型（如 “MajorClaim”）和关系保持不变。

## 2.2 基于合成数据的有监督微调

为了进一步提升模型在辩论构建任务中的表现，本工作结合官方提供的辩论数据集以及基于大型语言模型生成的合成数据（Synthetic Data），对模型进行了有针对性的微调。

在辩论构建子任务中，每条训练数据由三个部分组成：完整的立论陈词稿、待分类的论点以及对应的结构标签。模型的目标是根据给定的陈词稿及其中的某一论点，判断该论点在整体论证结构中所处的位置，即分类为 “主要论点” “普通论点” 或 “论点前提”。

在微调过程中，我们采用 LoRA（Low-Rank Adaptation）技术进行高效的有监督训练，并设计了如下形式的提示词（prompt）作为模型输入，使用的 prompt 如表 2 所示。

```
Below is an instruction that describes a
task. Write a response that appropriately
completes the request.
```

```
### Instruction:
```

我需要你帮我判断在某次辩论中的陈词稿中，论点属于是哪一种？

论点总共有三种类型：主要论点、普通论点、论点前提。

请判断论点属于是哪一种类型？

陈词稿：{example[‘陈词稿’]}

论点：{example[‘陈词稿’]}

```
### Response:
```

表 2: LoRA 微调过程中使用的提示工程

## 3 实验设置

本工作初始尝试过的基座模型有 Qwen3-8B<sup>2</sup>、Qwen3-8B-Base<sup>3</sup>和 Qwen3-14B<sup>4</sup>。经过我们的测试，发现在 Qwen3-14B 上辩论构建任务的性能最佳，故选择 Qwen3-14B 进行实验。

<sup>2</sup> <https://huggingface.co/Qwen/Qwen3-8B>

<sup>3</sup> <https://huggingface.co/Qwen/Qwen3-8B-Base>

<sup>4</sup> <https://huggingface.co/Qwen/Qwen3-14B>

在 LoRA 的配置方面，设置 `lora_r=16`，`lora_alpha=32`，`dropout=0`，`target_modules` 包括 Qwen3-14B 的所有可适配模块，为了加速训练过程，在训练过程中使用了 bf16 精度进行训练。

在训练配置方面，设置学习率为  $1e-4$ ，使用 AdamW 进行优化，最大输入长度为 768。为了高效微调 Qwen3-14B 模型，本研究结合使用了 DeepSpeed ZeRO Stage 2 策略<sup>[3]</sup>。

在生成配置方面，采用 `top_k=50` 的生成方式，`temperature=0.95`，`top_p=0.7`。

## 4 实验结果分析

本工作所有的推理操作都按照表 2 的格式构造提示词。实验部分将从性能指标和生成样例两个角度来分析。

### 4.1 性能指标分析

在实验过程中，我们将官方提供的辩论构建数据集按照 4:1 的比例划分为训练集和验证集，并将合成数据合并至训练集中。本实验在训练集中严格排除验证集和测试集的样本数据，确保训练集中不包含用于评估的论点，以避免数据泄露。随后使用训练集进行模型的微调训练，并在验证集上进行性能评估。评估指标采用准确率，用于衡量模型在结构分类任务中的整体表现。

为进一步评估合成数据的有效性，我们在对比实验中引入了两组设置：一组仅使用官方数据进行训练，另一组在原始训练集的基础上加入由大模型生成的合成数据进行扩充。

实验结果见表 3，由表 3 可知，Qwen3-14B+LoRA+合成数据的方案在所有设置中表现最优，在准确率上超越仅使用官方数据微调的模型。这一结果验证了合成数据在提升模型结构理解能力方面的有效性，也表明通过合成数据弥补真实数据不足，是提升辩论构建性能的可行路径。

表 3：不同模型/方法在辩论构建任务中的表现

| 使用的模型/方法                         | 准确率 (%) |
|----------------------------------|---------|
| bert-base-chinese<br>Fine-Tuning | 66.4    |
| Qwen3-8B LoRA                    | 84.6    |
| Qwen3-8B-Base LoRA               | 84.4    |
| Qwen3-14B LoRA                   | 85.9    |
| Qwen3-14B LoRA+合成数据              | 87.4    |

### 4.2 生成样例分析

本工作选取了一部分生成样例进行分析，以进一步验证模型在真实推理场景中的表现。通过观察模型对不同类型论点的分类结果，我们发现：在多数情况下，模型能够准确识别在陈词稿的结构当中的论点所起到的作用，能够准确识别主要论点与普通论点之间的区别。同时，对于结构上作为支撑信息出现的论点前提，模型也能在上下文中捕捉其支持作用，从而正确分类。

然而，正如表 4 和表 5 所展示的那样，在个别样例中，我们也观察到模型对于结构边界模糊的句子存在一定的判断偏差，可能将起到辅助性的论点前提错误识别为普通论点。此类误判主要出现在多句复合论证结构中，提示未来可以进一步增强模型对辩论结构中句际关系的建模能力。

广告对消费品的影响\n\n每天，各公司向消费者提供大量不同的产品。说服消费者购买产品最有效的方式是通过广告。然而，仅靠广告本身并不足够。产品还应满足消费者的需求。在我看来，广告是产品高销量的主要原因，有以下几个理由。我们主要通过广告了解产品。因此，广告主不断突破创意极限以促使消费者购买产品。当消费者对产品的广告方式印象深刻时，在某些情况下他们会被说服认为该产品是必需品。最近，电视上出现了一个非常有创意的软饮广告。该广告通过故事情节传递出“天气炎热时人们渴望饮用该软饮”的诉求。因此，该产品的销量显著上升。此外，产品广告在大众媒体上出现得越频繁，产品就越受欢迎。广告是打造知名产品最有效的方式。当消费者需要在两个不同品牌的同类产品中选择时，他们倾向于购买知名度更高的产品。当某产品被广泛使用时，无论其质量如何，都会获得社会信任。当然，产品价格也需在消费者承受范围内。基于这些事实，广告对推广产品具有不可否认的社会影响力。它们使产品更具吸引力。总而言之，虽然我认为消费者在购物时应优先考虑主要需求，但产品的高销量显然反映了广告的强大影响力。

表 4：个别样例当中使用的陈词稿

表 5：表 4 所示陈词稿中的不同论点的预测结果

| 论点            | 真实标签 | 预测标签 |
|---------------|------|------|
| 广告是产品高销量的主要原因 | 主要论点 | 主要论点 |

|                      |      |      |
|----------------------|------|------|
| 说服消费者购买产品最有效的方式是通过广告 | 普通论点 | 普通论点 |
| 我们主要通过广告了解产品         | 论点前提 | 论点前提 |
| 产品还应满足消费者的需求         | 论点前提 | 普通论点 |

## 5 总结

针对现有通用语言模型在辩论构建任务中结构识别能力有限的问题，本工作提出了基于 LoRA 微调的适配方案，并通过引入结构控制的合成数据进一步提升模型对论辩结构的建模能力。实验结果表明，加入合成数据的 Qwen3-14B LoRA 模型在结构分类准确率上达到 87.4%，显著优于其他模型配置。同时，生成样例分析验证了模型在实际推理中的应用价值。整体来看，本方案在准确性与实用性上均取得良好表现，为自主论辩系统的结构识别模块提供了有效技术支撑。

## 参考文献

- [1] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. ICLR, 2022, 1(2): 3.
- [2] Lin J, Chen G, Jin B, et al. Overview of AI-Debater 2023: The Challenges of Argument Generation Tasks[J]. arXiv preprint arXiv:2407.14829, 2024.
- [3] Rajbhandari S, Rasley J, Ruwase O, et al. Zero: Memory optimizations toward training trillion parameter models[C]//SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 2020: 1-16.

# 第五届智慧辩论评测（AI-Debater 2025）自主辩论生成技术报告

黄煜聪<sup>1</sup> 靳伯骏<sup>1</sup> 王华杰<sup>1</sup> 张宸华<sup>1</sup> 陈冠荣<sup>1</sup> 何宇航<sup>1</sup> 孙洋<sup>1</sup>  
鲍建竹<sup>1</sup> 徐睿峰<sup>1</sup>

(1. 哈尔滨工业大学（深圳），广东省 深圳市 518055)

**摘要：**自主辩论生成是计算辩论领域中具有挑战性的任务，要求模型在给定辩题和立场的条件下生成结构完整、逻辑清晰、具有说服力的立论陈词稿。近年来，大语言模型在自然语言生成方面取得了显著突破，但现有主流开源模型（如 LLaMA、Qwen、Deepseek 等）多为通用模型，难以满足辩论任务中对结构性与事实性内容的高要求。因此，本工作以 Qwen3-14B 为基础，结合官方提供的数据与高质量合成样本，设计数据合成、自动评估筛选和 LoRA 微调三阶段流程，构建面向辩论任务的定制模型。实验结果表明，该方案在事实准确性等维度上的表现优于只使用官方数据集微调的模型，验证了本方法在自主辩论生成任务中的有效性。

**关键词：**计算辩论；大语言模型；合成数据；辩论生成

中图分类号：TP391

文献标识码：A

## The technical report of autonomous debate generation in AI-Debater 2025

Yucong Huang<sup>1</sup>, Bojun Jin<sup>1</sup>, Huajie Wang<sup>1</sup>, Chenhua Zhang<sup>1</sup>, Guanrong Chen<sup>1</sup>, Yuhang He<sup>1</sup>, Yang Sun<sup>1</sup>,  
Jianzhu Bao<sup>1</sup> and Ruifeng Xu<sup>1</sup>

(1. Harbin Institute of Technology, Shenzhen, Shenzhen, Guangdong 518005, China)

**Abstract:** Autonomous debate generation is a challenging task in the field of computational argumentation, requiring models to generate well-structured, logically coherent, and persuasive argumentative speeches given a specific debate topic and stance. In recent years, large language models have achieved remarkable progress in natural language generation; however, most mainstream open-source models (e.g., LLaMA, Qwen, Deepseek) are general-purpose and struggle to meet the high demands of structural rigor and factual consistency required by debate tasks. To address this gap, this work builds upon Qwen3-14B and proposes a three-stage framework consisting of data synthesis, automatic evaluation and filtering, and LoRA-based fine-tuning, leveraging both official datasets and high-quality synthetic samples to develop a task-specific model for debate generation. Experimental results demonstrate that the proposed approach outperforms models fine-tuned solely on the official dataset in terms of factual accuracy and other dimensions, demonstrating the effectiveness of our method in the autonomous debate generation task.

**Key words:** Computational Argumentation; Large Language Models; Synthetic Data; Argument Generation

## 0 引言

目前，自动化生成具有逻辑性、说服力和结构完整性的辩论文本正成为人工智能领域的重要挑战。近年来，计算论辩（Computational Argumentation）逐渐成为自然语言处理领域的研究热点，其核心目标是赋予机器理解、生成和组织论辩内容的能力，推动其在智能决策、舆论分析、教育辅导等场景中的应用落地。

在此背景下，第五届智慧论辩评测聚焦于自主论辩任务、自主论辩赋能社会调查两个任务，研究者在自主论辩任务中提出了多个关键子任务，包括辩题识别、论据挖掘、辩论构建和自主论辩生成。在本次报告中，我们聚焦于自主论辩生成这一子任务，该任务使用一个辩题及其对应支持方，要求生成一篇一辩立论陈词稿。

随着以 GPT-4 为代表的大型语言模型（Large Language Models, LLM）的广泛应用，模型在自然语言生成与理解方面取得了显著进展，极大提升了论辩类任务中语言生成的质量与逻辑性。然而，当前主流开源大模型（如 LLaMA、Qwen、Deepseek 等）普遍面向通用任务设计，虽具备较强的多任务泛化能力，但在面对具有复杂结构与逻辑推理要求的论辩任务时仍存在一定局限。所以，本工作在通用模型的基础上，使用官方所提供的自主论辩生成数据集和基于大模型的合成数据进行 LoRA 微调<sup>[1]</sup>，使得模型能够更好地适配于自主论辩生成任务。实验结果表明，本工作的方案可以提高生成的论辩文本的质量，模型参数已经开源至 huggingface<sup>1</sup>。

## 1 相关工作

近年来，针对自主辩论任务中训练数据不足或结构复杂性高的问题，部分研究开始尝试引入合成数据（Synthetic Data）以提升模型的论辩能力和结构感知能力。在 CCAC 2023 的 AI-Debater2023 任务中，HITSZ-HLT 在使用了合成数据来补充数据集，从而提升了自主辩论任务的表现效果，这验证了合成数据在自主辩论任务中能够增强模型语言理解能力、缓解数据稀缺和提升结构建模性能方面的有效性<sup>[2]</sup>。

## 2 技术方案介绍

本工作的框架示意图如图 1 所示，其主要包含三部分：基于大模型生成的数据合成（Synthetic Data Generation, SDG）、数据评估与筛选模块（Synthetic Data Evaluation）和基于合成数据的有监督微调（Supervised Fine-Tuning, SFT）。

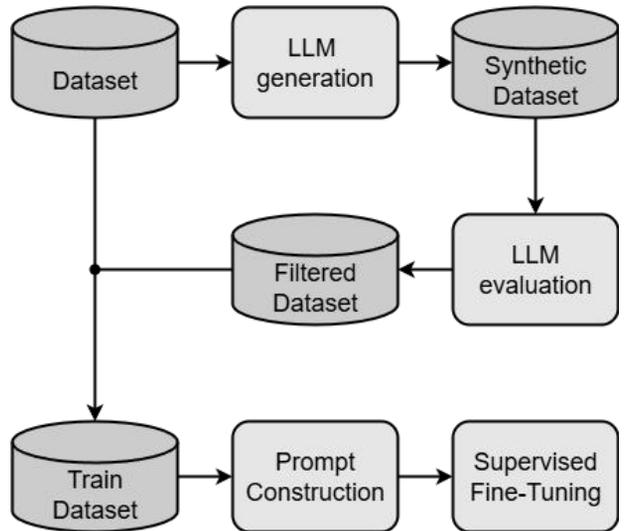


图 1: 本工作框架图

### 2.1 基于大模型生成的数据合成

在自主论辩生成任务中，模型需在给定辩题与立场的前提下，自动生成一篇结构完整、逻辑清晰的一辩立论陈词。然而，受限于官方数据集中的有限数量的样本，直接使用有限标注数据训练模型，往往会导致模型生成内容重复、结构单一、论点分布不均等问题，难以支撑高质量的自主生成任务。为解决这一问题，本工作设计了一套基于大语言模型（LLM）构造合成数据的策略，自动扩展训练集，增强模型对不同类型论题、立场、论点结构的建模能力。

首先，我们从公开资源中选取了涵盖社会、教育、经济等多个领域和话题的辩题，构建具有广泛代表性的辩题集合。随后使用 deepseek-v3 模型围绕这些辩题分别生成正方与反方的立论陈词稿，对每个立场生成三篇候选样本，以引入更多样的候选样本，并在后续阶段中筛选出质量最高的候选样本作为最终样本。为提升生成质量并引导模型学习典型的论辩结构，我们在每一次生成过程中，采用小样本学习（Few-Shot Learning）的方式，从官方数据集中随机采样两

篇高质量的陈词稿作为示例，嵌入到提示词中，引导模型生成结构合理、立场明确的立论文本。我们所使用的提示词（prompt）如表 1 所示。

```
你是一个出色的陈词稿撰写专家，请你参考以下给的示例，为目标辩题撰写符合给定立场的立论陈词稿（注：需要完成的是一辩稿）。  
## 示例 1:  
辩题: {topic_1}  
立场: {stance_1}  
陈词稿: {argument_2}  
## 示例 2:  
辩题: {topic_2}  
立场: {stance_2}  
陈词稿: {argument_2}  
请你参考以上示例，为以下目标辩题和立场撰写 3 份立论陈词稿（注：需要完成的是一辩稿），每个陈词稿用“---”进行分割（每个陈词稿内部不要出现“---”），不要出现其他任何无关信息：  
目标辩题: {target_topic}  
立场: {target_stance}
```

表 1: 指导合成数据的提示工程

## 2.2 基于大模型的数据评估和筛选

为了进一步提升模型在辩论构建任务中的表现，本工作对大模型生成的合成陈词稿进行了系统性筛选与质量控制。具体而言，我们使用 GPT-4o 模型从五个维度对每份生成样本进行评分，分别为：论点清晰度、论证结构安排、内容相关性、事实准确性以及是否符合一辩立论陈词的写作特点。在每组三篇独立生成样本中，我们选择综合得分最高的一篇作为最终样本，用于构建高质量的训练数据集。我们所使用的提示词如表 2 所示。

```
你是一名严格的辩论赛教练，你的任务是评估一份 {target_stance} 争对 “{target_topic}” 这一辩题的立论陈词稿，评估其中论点的力度，找出逻辑或事实错误，并评估他们对高级辩论技巧的使用。  
  
立论陈词稿: {target_argument}  
  
## 评价标准  
  
1. **论点清晰**
```

```
-陈词稿的论点表达清楚了吗？推理是否合理且易于理解？  
-**0**：论点不清楚或令人困惑。  
-**1**：论点结构良好，易于理解。  
  
2. **论证安排**  
-陈词稿的论点是否以合乎逻辑、连贯的方式提出？每一点都能顺利地进入下一点吗？  
-**0**：论点混乱，难以理解。  
-**1**：论点遵循清晰和合乎逻辑的进程。  
  
3. **相关性**  
-陈词稿每个论点都直接涉及辩论主题吗？是否有任何无关的点或离题的干扰？  
-**0**：偏离主题的论点。  
-**1**：每个论点都与主题相关。  
  
4. **事实准确性**  
-陈词稿在事实上正确吗？所使用的数据或示例是否可靠且最新？  
-**0**：关键数据或示例不正确，削弱了论点。  
-**1**：使用一系列可靠、最新的数据和示例，完全支持索赔，并从多个角度进行探索。  
  
5. **立论陈词稿特点遵循**  
-这份陈词稿是否符合立论陈词稿的需求？是否可以作为辩论赛开篇的立论？  
-**0**：预设对方论点，反驳对方尚未提出的观点。  
-**1**：提出己方完整论证体系，使用中立数据支撑己方逻辑。  
  
请你先就以上五个维度逐项进行简短分析（每项不超过 2 句话），然后在最后统一打分。请以如下格式输出你的完整结果（分析部分用自然语言，打分部分为 JSON 字典）：  
【分析】  
情感吸引力：……  
论点清晰：……  
……  
【评分】  
{“论点清晰”: a, “论证安排”: b, “相关性”: c, “事实准确性”: d, “立论陈词稿特点”: e}  
请确保评分 JSON 是你回复的最后一段内容，并保持其为合法 JSON 格式。
```

表 2: 陈词稿评估和筛选过程中使用的提示词

## 2.3 基于合成数据的有监督微调

为了进一步提升模型在辩论构建任务中的表现，本工作结合官方提供的自主论辩生成数据集以及基于大型语言模型生成的合成数据（Synthetic Data），对模型进行了有针对性的微调。

在自主论辩生成子任务中，每条训练数据由三个部分组成：完整的立论陈词稿、待分类的论点以及对应的结构标签。模型的目标是根据给定的陈词稿及其中的某一论点，判断该论点在整体论证结构中所处的位置，即分类为“主要论点”“普通论点”或“论点前提”。

在微调过程中，我们采用 LoRA（Low-Rank Adaptation）技术进行高效的有监督训练，并设计了如下形式的提示词（prompt）作为模型输入，使用的 prompt 如表 3 所示。

```

Below is an instruction that describes a task. Write a response that appropriately completes the request.

### Instruction:
我需要你帮我生成一段辩论赛所需要的陈词稿

你需要根据特定的辩题生成相应的陈词稿
在辩论赛中你所需要担任的是 {example['持方']} 的角色
本次辩论赛所使用的辩题是：{example['辩题']}

### Response:

```

表 3: LoRA 微调过程中使用的提示词

### 3 实验设置

对于自主论辩生成任务，我们使用 Qwen3-14B 模型<sup>2</sup>进行实验。在 LoRA 的配置方面，设置 lora\_r=8, lora\_alpha=16, dropout=0, target\_modules 包括 Qwen3-14B 的所有可适配模块，为了加速训练过程，在训练过程中使用了 bf16 精度进行训练。

在训练配置方面，设置学习率为 1e-4，使用 AdamW 进行优化，最大输入长度为 1024。为了高效微调 Qwen3-14B 模型，本研究结合使用了 DeepSpeed ZeRO Stage 2 策略<sup>[3]</sup>。

<sup>2</sup> <https://huggingface.co/Qwen/Qwen3-14B>

在生成配置方面，采用 top\_k=50 的生成方式，temperature=0.95, top\_p=0.7。

### 4 实验结果分析

本工作的所有推理操作均按照表 3 所示的提示词格式进行构造。实验部分将以生成样例为出发点，对模型在实际生成过程中的表现进行分析与评估。表 4 展示了一个生成的陈词稿样例。

表 4: 模型生成的陈词稿样例

|     |  |
|-----|--|
| 辩题  | 短视频盛行提升了/降低了当代人的认知能力   |
| 持方  | 反方   |
| 陈词稿 | <p>主席、评委、对方辩友：</p> <p>大家好！今天我方的立场是：短视频盛行降低了当代人的认知能力。开宗明义，认知能力是指人类获取、加工和运用信息的能力，其核心在于深度思考、逻辑推理和批判性思维。短视频以其碎片化、即时化的传播特点，正在系统性地瓦解这一能力。我方将从认知负荷、思维惰性与知识贬值三个层面展开论述。</p> <p>首先，短视频的碎片化结构严重超载认知负荷。神经科学研究表明，人类大脑在注意力切换时会产生“认知摩擦”，每次切换消耗的脑力相当于完成一个完整任务的 15%。抖音的 60 秒视频平均观看时长，相当于每分钟切换 6 次认知模式。当知识被压缩成 15 秒的口播、3 秒的字幕，大脑被迫进入“信息消化不良”状态。斯坦福大学实验证实，连续观看短视频 2 小时后，受试者知识留存率骤降至 12%，远低于传统阅读的 68%。</p> <p>其次，算法驱动的“信息茧房”正在制造思维惰性。今日头条的推荐机制显示，用户点击“确认”按钮的平均反应时间仅 0.3 秒，这种即时反馈强化了“条件反射式”的思维习惯。当大脑习惯于在 0.5 秒内获得结论，深度思考的神经回路就会萎缩。MIT 媒体实验室的脑成像研究发现，长期使用短视频的用户，其前额叶皮层</p> |

(负责逻辑推理)活跃度比同龄人低 23%，而边缘系统(负责情绪反应)活跃度高出 41%。这解释了为何我们越来越依赖直觉而非分析。

最后，短视频正在引发知识的“通货膨胀”。中国社科院 2023 年报告指出，短视频平台知识类内容中，92%存在概念扭曲或断章取义。当“量子力学”被简化为“爱情玄学”，“相对论”变成“人生哲学”，我们失去的不仅是知识的准确性，更是对复杂世界的敬畏心。更危险的是，短视频用“信息快餐”替代了“认知正餐”，导致知识获取呈现“高频率低吸收”的特征——用户日均接触信息量增长 500%，但知识内化效率却下降 37%。

综上所述，短视频不是认知的加速器，而是认知的钝化剂。它用即时满足的糖衣包裹着思维退化的苦药，用娱乐至死的狂欢掩盖着知识荒漠的危机。谢谢！

由于官方未提供统一的评估指标体系，本工作使用在合成数据评估和筛选阶段所采用的五个评价维度作为参考，对不同训练阶段模型生成的立论陈词样本进行了综合评估。这五个维度包括：论点清晰度、论证结构、相关性、事实准确性以及是否具备一辩陈词稿的语言风格与结构特点。通过人工与自动评分相结合的方法，我们对不同方案下生成的陈词稿的表现进行了分析，评分结果如表 5 所示。

表 5：不同方案评分的结果

| 使用的数据集/方案   | 论点清晰  | 论证安排  | 相关性   |
|-------------|-------|-------|-------|
| 官方数据集       | 0.992 | 0.973 | 0.997 |
| 官方数据集+合成数据集 | 1.000 | 0.954 | 0.980 |

| 使用的数据集/方案   | 事实准确性 | 立论陈词稿特点遵循 | 总分    |
|-------------|-------|-----------|-------|
| 官方数据集       | 0.151 | 0.730     | 3.843 |
| 官方数据集+合成数据集 | 0.313 | 0.829     | 4.076 |

评估结果显示，与仅使用官方数据集训练的

模型相比，同时引入官方数据集与合成数据集进行训练的模型在事实准确性和立论陈词稿特点两个维度上表现出显著提升。这一结果表明，模型在训练过程中更好地学习了如何生成具备事实支撑力的内容，并在一辩陈词的写作规范方面展现出更强的适应性，这种提升源自于基于大模型生成的合成数据和基于大模型的数据评估和筛选机制的共同作用。通过引入合成数据结合筛选机制，本工作成功利用高质量的合成数据，引导模型提升了在这两个维度上的表现。

然而，需要指出的是，模型在论证安排和相关性两个维度上出现了小幅下降。我们推测，这一现象正是由于合成数据中仍然存在部分质量参差不齐的样本。尽管合成样本能够丰富语料分布，但在生成过程中仍可能出现结构松散或主题漂移等问题，导致个别样本在论点展开的节奏控制和主题聚焦方面存在不足。

值得一提的是，模型在论点清晰度方面得分基本保持稳定，说明合成数据尽管在部分结构层面存在瑕疵，但整体上并未削弱模型对于核心观点表达的掌控能力。

## 5 总结

针对通用大模型在论辩生成任务中结构理解弱、论证不严谨等问题，本工作提出了一套完整的自主论辩生成优化流程。该流程包括：1) 使用大模型构造多领域、高多样性的合成辩论文本，扩展训练数据的广度与表达多样性；2) 引入 GPT-4o 从五个维度对生成内容进行系统评估和筛选，确保训练样本质量；3) 结合官方数据集和高质量合成数据集，采用 LoRA 对 Qwen3-14B 进行微调，使模型更好地适配辩论生成任务需求。

实验结果表明，合成数据的引入显著提升了模型在事实准确性和立论陈词稿特点遵循方面的能力，然而仍需年在结构和主题聚焦上进一步优化。后续工作中，我们建议注重提升合成数据在结构和主题聚焦方面的表现，如引入结构模板生成、显式主题约束机制等，以实现结构与主题聚焦的双重优化。

总体而言，本工作展示了通过合成数据驱动与多维评估筛选机制相结合的方式，可以有效缓解通用模型在特定论辩任务中的适配瓶颈，为构建更高质量的计算论辩系统提供了可行路径与方法参考。

## 参考文献

- [1] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. ICLR, 2022, 1(2): 3.
- [2] Lin J, Chen G, Jin B, et al. Overview of AI-Debater 2023: The Challenges of Argument Generation Tasks[J]. arXiv preprint arXiv:2407.14829, 2024.
- [3] Rajbhandari S, Rasley J, Ruwase O, et al. Zero: Memory optimizations toward training trillion parameter models[C]//SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 2020: 1-16.

# CCAC2025 评测技术报告

赵学官<sup>1</sup>, 司加胜<sup>2\*</sup>

1. 齐鲁工业大学 (山东省科学院), 山东省济南市 邮编 250353

2. 齐鲁工业大学 (山东省科学院), 山东省济南市 邮编 250353

**摘要:** 本报告介绍了智慧论辩评测中的创新方法, 涵盖网络媒体内容论辩理解、自主论辩生成及论辩赋能社会调查仿真三个任务。通过设计辩题识别、论据挖掘、辩论构建和自主论辩生成四个子任务, 我们评估了自主论辩能力。在辩题识别任务中, 利用 Qwen3-8B 模型和 LoRA 微调, 准确率达到 0.94; 在论据挖掘任务中, 使用关键信息识别策略, 取得了 0.84 的准确率; 通过多智能体推理机制, 辩论构建任务得到了 0.83 的准确率; 自主论辩生成任务通过多智能体辩论驱动的设计论证链方法, 提升了生成陈词的结构与逻辑性。在论辩赋能社会调查仿真任务中, 模拟相似个体信息的多智能体辩论显著提高了社会调查响应的准确性。本报告展示了多智能体协作和自动化论辩方法在结构化论辩任务中的潜力, 为自动化论证理解与生成提供了有效路径。

**关键词:** 论证挖掘; 论点生成; 大语言模型

中图分类号: TP391

文献标识码: A

## CCAC2025 Evaluation Technical Report

Xueguan Zhao<sup>1</sup>, Jiacheng Si<sup>2\*</sup>

1. Qilu University of Technology (Shandong Academy of Sciences), Jinan, Shandong, 250353, China

2. Qilu University of Technology (Shandong Academy of Sciences), Jinan, Shandong, 250353, China

**Abstract:** This report presents innovative methods in the Wisdom Debate Evaluation, covering three main tasks: network media content debate comprehension, autonomous debate generation, and debate-empowered social survey simulation. We designed four sub-tasks to assess autonomous debate capabilities: topic identification, argument mining, debate construction, and autonomous debate generation. Using the Qwen3-8B model and LoRA fine-tuning, the topic identification task achieved an accuracy of 0.94; in argument mining, a key information identification strategy reached 0.84 accuracy; the debate construction task obtained 0.83 accuracy with a multi-agent reasoning mechanism; and the autonomous debate generation task improved the structure and logic of generated speeches using a multi-agent debate-driven argument chain method. In the debate-empowered social survey simulation task, simulating multi-agent debates with similar demographic information significantly enhanced the precision of survey responses. This report demonstrates the potential of multi-agent collaboration and automated debate methods in structured debate tasks, providing an effective path for automated argument comprehension and generation.

眉题

**Key words:** Argument Mining; Argument Generation; Large Language Models

## 0 引言

随着人工智能技术的进步，基于大语言模型（LLM）的自动化论辩系统逐渐成为研究的热点。论辩作为人类沟通的核心形式，不仅是逻辑思维和语言表达的训练工具，也在推动智能化社会中扮演着关键角色。如何利用现代人工智能技术实现结构化的自动论辩生成与理解，成为一个重要课题。

本报告聚焦于智慧论辩评测中的创新方法与实验成果，特别是在网络媒体内容理解、论辩生成和社会调查仿真等领域的应用。评测任务从不同角度验证了语言模型在自动论辩领域的可行性与有效性。我们通过设计多个子任务，旨在提升自动化系统在推理能力、准确性和生成质量方面的表现。

具体而言，本文探讨了如何通过多智能体协作提升辩论推理能力，如何利用大语言模型微调策略提高论辩任务的准确性，以及如何通过自主生成的论辩内容推动社会调查仿真并提高响应精度。基于 Qwen3-8B 模型的微调方法，在多个任务中取得了良好的实验结果，展示了这些方法在结构化论辩任务中的潜力。

本报告为智能论辩系统的设计与实现提供了新思路，展示了自动化论辩生成与理解协同工作的价值，推动了各领域智能化的发展。

## 1 技术方案介绍

### 1.1 数据准备与大语言模型选择

网络媒体内容辩论理解数据集：中文数据集，涵盖辩题类别、论点与论据关系等标注信息，基于 PE 辩论挖掘数

据集，通过 DeepSeek-R1 翻译并校对后整理为 JSON 格式。

自主辩论生成数据集：中文数据集，包含辩题与立论陈词稿，数据源于华语辩论赛题，构建使用 DeepSeek-R1、GPT-4-turbo、Gemini-2.5 等模型，经过校对后整理为 JSON 文件。

社会调查仿真数据集：来源于某国社会调查数据，经过清洗与剔除后整理成 JSON 文件。

除官方提供的数据集外，为了在自主论辩生成任务上验证我们方法的性能，我们还在 ArgEssay 数据集<sup>[1]</sup>：来源于国际标准化英语写作测试，涵盖了广泛的争议性话题，包括技术进步、教育方法和环境问题上随机抽取了 50 条组成测试数据。

所有数据集均按 9:1 的比例从训练集中分割为训练集与测试集。

模型选择：我们采用 Qwen3-8B 作为参赛使用的基座大语言模型。Qwen3-8B 在大规模多语言和多模态数据上进行了预训练，并通过高质量数据进行微调，使其能够贴近人类偏好。Qwen3 具备自然语言理解、文本生成、视觉理解、音频理解、工具使用、角色扮演及 AI Agent 交互等多种能力，能够支持我们在多任务下的有效表现。

### 1.2 模型微调

针对各个，我们利用 SWIFT 框架<sup>[2]</sup>并采用 Lora 微调的方式得到各个子任务对应的模型，具体微调参数如附录 5.1 所示。

### 1.3 辩题识别

辩题识别是论辩任务中的核心步骤之一，目的是判断给定论点是否符合当

前辩题。传统的辩题识别方法可能存在对语境理解的不足，尤其是在复杂和多变的论辩内容中。模型 Qwen3-8B 在推理模式下的表现已经非常出色，针对该任务我们通过辩题识别微调数据构建 prompt 模板构造数据集进行 lora 微调进一步提升模型在该任务上的性能：根据 Qwen3 官方文档<sup>[3]</sup>，我们在进行微调时并未为数据构建相应的思考数据，因此采用/no\_think 格式在训练时保留模型原有的思考推理能力。具体 prompt 模板在附录 5.2。

#### 1.4 论据挖掘

论据挖掘是自动化论辩任务中的核心环节，旨在判断论据是否能够支持当前论点。在辩论系统中，准确识别论据与论点之间的支持关系对提升论辩质量至关重要。然而，传统方法在处理复杂语义关系和模糊指代时常常遇到困难。为了解决这些问题，我们引入了基于大语言模型的微调方法，并结合关键信息强化提示策略，帮助模型更精准地识别论据与论点之间的支持关系。

具体而言，我们首先使用论据挖掘微调数据构建了 prompt 模板，并通过 LoRA 微调技术优化了模型，使其能够学习到论据支持论点的特征。随后，我们设计了关键问题分析 prompt，首先让模型判断是否能直接根据当前论点与论据判断二者的支持关系。如果无法直接判断，模型会识别出关键问题并给出理由。最后，我们将模型输出的关键问题和理由输入分析判断 prompt，经过思考后，模型给出最终的判断结果。通过这种方法，模型能够更有效地处理模糊指代问题，显著提升论据挖掘任务的准确性。具体 prompt 在附录 5.3。

#### 1.5 辩论构建

传统的辩论构建方法通常依赖单一智能体判断，缺乏多角度的推理与交互，可能导致片面或错误的判断。为了解决这一问题，我们引入了多智能体协作机制，通过模拟智能体之间的辩论与思维碰撞，提升模型的判断全面性和准确性。

在此方法中，首先由第一分析智能体基于初步判断 prompt 给出判断结果和理由。随后，第二和第三分析智能体在分析第一智能体的结果和理由后，使用后续判断 prompt 独立给出自己的判断和理由。如果存在不一致，系统进入辩论环节。三个智能体通过辩论 prompt 互相分析对方的判断理由，讨论优劣，并根据辩论结果修正自己的判断。这一过程模拟了人类辩论中的“反思—修正”机制，避免了单一判断带来的偏误。

若经过多轮辩论后仍未达成一致，裁判智能体将介入，综合分析辩论过程中的判断和理由，做出最终结论。裁判智能体相当于辩论中的最终裁决者，确保在复杂情况下系统能得出合理结论。

通过多智能体的互动与协作，我们的方法能更好地模拟复杂的辩论过程，提升辩论构建任务的处理能力，并为智能系统在多方协作中的应用提供新思路。整个流程通过多轮交互和推理，确保系统能够做出更加合理、准确的判断。具体 prompt 在附录 5.4。

#### 1.6 自主论辩生成

在自主论辩生成任务中，生成高质量的立论陈词稿是实现有效辩论的关键。传统的文本生成方法往往缺乏深度和逻辑性，难以应对复杂辩论。为此，我们提出了一种基于 Self-Ask 范式<sup>[4]</sup>和 Toulmin 模型<sup>[5]</sup>的多智能体协作生成方

法,旨在通过整合论证链和智能体协作,提升生成文本的结构性和逻辑性。

具体而言,我们首先使用微调数据构建训练集并训练基座模型。然后,论点生成 Agent 基于辩题和立场生成初始主论点,并通过自我问答生成推理链。每个推理链将结合 Toulmin 模型(包括 Claim、Grounds、Warrant、Qualifier)生成完整的论证结构。接下来,论点生成 Agent 将这些论证输入辩论 Agent 进行多智能体辩论,收集辩论记录并输入分析 Agent。分析 Agent 会基于辩论结果优化各个 Toulmin 模型。最后,我们将优化后的 Toulmin 模型输入写作 Agent,生成完整的立论陈词稿。具体 prompt 在附录 5.5 中。

### 1.7 论辩赋能社会调查仿真

在论辩赋能社会调查仿真任务中,我们通过让相似个体信息的个体进行辩论,解决了模拟个体决策中可能存在的偏差。现实中,受访者的选择受背景、经验和情感态度等多重因素影响,单个个体往往因信息不全或认知局限作出偏差的选择,影响调查准确性。因此,我们设计了个体信息重叠度计算算法,选取信息相似但观点不同的个体进行辩论,以确保观点的碰撞和多样性。

在辩论过程中,每个个体首先独立完成初次调查,生成原始回答。随后,根据重叠度算法,选取两个个体信息最相似的个体进行辩论。为了保证辩论质量,我们采用了两轮辩论机制(陈词-回复),并使用 1.6 节中的自主论辩生成方案,通过多智能体协作修正观点,借助推理链和 Toulmin 模型优化论证,使陈词中的论点更具深度和说服力。

最后,第二次调查前,基于 LLM 的模拟个体会检索是否在该问题上参与

过辩论。若有,系统将自动整合辩论记录,帮助个体在回答时结合辩论中的新观点,从而提供更加准确、全面的反馈。这一过程通过认知迭代确保了最终回答的代表性、客观性和准确性。

## 实验设置

### 2.1 Baseline

在本次评测任务中,为了评估我们提出的方法的有效性和准确性,我们针对子任务辩题识别、论据挖掘及辩论构建分别设计了不同的 baseline:(1) Bert-base<sup>[6]</sup>,通过每个子任务的训练集训练 Bert 后直接进行分类预测;(2) LLM-base,使用 Qwen3-8b 直接提示每个子任务的任务要求来预测;(3) LLM-trined,使用训练集进行 lora 微调后的 Qwen3-8b 直接提示每个子任务的任务要求来预测。

对于自主论辩生成任务,我们参考 AEG 任务(议论文章生成任务),设置了以下 baseline:(1) DD-KW<sup>[1]</sup>,具有显式内容规划的双解码器 Transformer。(2) E2E:通过 LLM 提示直接生成论文。(3) CoT<sup>[7]</sup>:在写作前生成简短的计划。(4) ToT<sup>[8]</sup>:在写作前生成并从多个计划中选择。(5) DPE<sup>[9]</sup>:具有关键自我反思的两阶段规划。(6) D2W<sup>[10]</sup>:模拟智能体辩论进行协作论文生成。

### 2.2 评测指标

对于辩题识别,论据挖掘,辩论构建任务,我们采取准确率(Accuracy),及分类正确的样本占总样本个数的比例作为评估指标。

对于自主论辩生成任务,我们根据相关研究工作<sup>[11][9][10]</sup>,设计相关性(ReL.)、逻辑(Log.)、说服力(Per.)共 3 个维度的评估 prompt,并采用 GPT-4o 作为评估

的执行模型在每个维度上进行 5 分制的打分,具体评估 prompt 在附录 5.6 中。

对于辩论赋能社会调查仿真任务,我们综合了精确度和召回率计算每个问题的加权 F1 分数来衡量模型在辩论赋能社会调查任务中的表现。首先对每个问题的真实标签和预测结果进行匹配,然后计算并输出所有问题的加权 F1 分数的平均值,作为整体模型的性能指标。

### 3 实验结果

如表 1 所示,我们的模型在网络媒体内容理解任务中表现出色。在辩题识别任务中,我们与基线模型均达到了 0.9444 的准确率。在论据挖掘任务中,准确率达到 0.84,显著领先于 Bert-base 和 LLM-base,展示了明显优势。辩论构建任务中的准确率为 0.83,接近 LLM-trained,体现了方法的稳定性。总体而言,我们的方法在各任务中均表现优异,尤其在论据挖掘上具有突出优势。

表 1 网络媒体内容理解实验结果

| 任务   | Bert-base | LLM-base | LLM-trained | Our    |
|------|-----------|----------|-------------|--------|
| 辩题识别 | 0.8529    | 0.8385   | 0.9444      | 0.9444 |
| 论据挖掘 | 0.6360    | 0.6954   | 0.7395      | 0.84   |
| 辩论构建 | 0.6979    | 0.5854   | 0.8271      | 0.83   |

如表 2 所示,我们的方法在相关性 (4.636) 和逻辑性 (4.32) 上超越

了其他方法,表现优异。相比之下,DD-KW 得分较低 (相关性 2.26, 逻辑性 1.796),可能由于其基础的双解码器结构。E2E 方法在相关性 (4.426) 和逻辑性 (4.218) 上表现较好,但说服力 (3.946) 稍逊。CoT 和 ToT 的表现接近,ToT 略优,尤其在逻辑性和说服力上 (4.228 和 4.144)。DPE 的表现稳定,但在说服力上未见突破。D2W 的协作生成方法在相关性和逻辑性上稍逊,但说服力 (3.876) 较弱。总体而言,我们的方法在所有维度上均领先,特别在相关性和逻辑性方面展现了强大潜力。

表 2 自主辩论生成实验结果

|       | Rel.  | Log.  | Per.  |
|-------|-------|-------|-------|
| DD-KW | 2.26  | 1.796 | 1.612 |
| E2E   | 4.426 | 4.218 | 3.946 |
| CoT   | 4.5   | 4.142 | 3.872 |
| ToT   | 4.486 | 4.228 | 4.144 |
| DPE   | 4.488 | 4.15  | 4.148 |
| D2W   | 4.518 | 4.14  | 3.876 |
| Our   | 4.636 | 4.32  | 4.25  |

如表 3 所示,在辩论赋能社会调查仿真任务中我们的方法在加权 F1 分数上 (0.7420) 相较于 LLM-base (0.7276) 表现更为优异。尽管两者的表现都较为接近,但我们的模型在精确度和召回率的平衡上有所提升,表明其在处理该任务时能够更好地捕捉到各种类别群体的特征使仿真个体做出的调查结果更加真实。

表 3 辩论赋能社会调查仿真实验结果

|          | Weighted F1 |
|----------|-------------|
| LLM-base | 0.7276      |
| Our      | 0.7420      |

### 4 总结

在本次评测任务中,围绕中文论辩

任务的四个关键子任务展开系统研究，涵盖了从辩题识别到自主论辩生成再到论辩赋能社会调查仿真的完整流程。我们针对各任务特点设计了相应的优化方法。整体而言，本报告所提出的系列方法有效验证了语言模型在结构化论辩场景中的多样能力，为自动化论证理解与生成任务及模型应用提供了有力支持与实践路径。

## 5 附录：

### 5.1 Lora 微调参数

表模型 Lora 微调参数

| 参数                          | 值        |
|-----------------------------|----------|
| sft_type                    | lora     |
| dtype                       | AUTO     |
| num_train_epochs            | 3        |
| lora_rank                   | 8        |
| lora_alpha                  | 32       |
| torch_dtype                 | bfloat16 |
| per_device_train_batch_size | 1        |
| per_device_eval_batch_size  | 1        |
| gradient_accumulation_steps | 16       |
| learning_rate               | 1e-4     |

### 5.2 辩题识别任务 prompt

#### 辩题识别微调数据构建 prompt:

```
"messages": [
  {
    "role": "user",
    "content": f"##任务要求：请你判断给定论点是否属于给定辩题：\n##输入：\n<辩题>{topic}</辩题>\n<给定论点>{argument}</给定论点>\n##输出要求：如果论点属于辩题，输出：‘1’；否则，输出‘0’/no_think"
  }
]
```

```
"role": "assistant",
```

```
"content":
```

```
f"<think>\n\n</think>\n\n{label}"
  }
]
```

### 5.3 论据挖掘任务 prompt

#### 论据挖掘微调数据构建 prompt:

```
"messages": [
  {
    "role":
      "user",
      "content":
        f"##任务要求：请你判断给定论据是否可以用于支持给定论点：\n##输入：\n<论点>{argument}</论点>\n<论据>{evidence}</论据>\n##输出要求：如果论据可以用于支持论点，输出：‘1’；否则，输出‘0’/no_think"
    },
  {
    "role":
      "assistant",
      "content":
        f"<think>\n\n</think>\n\n{label}"
    }
  ]
```

#### 关键问题识别 prompt:

你的任务是分析给定的 {evidence} 是否能够被判断为 {target} 给定的 {argument}。判断时，请根据常识、逻辑推理及合理假设进行推理。

如果存在合理的推理链，即使中间步骤未被显式陈述，也应视为可以判断。仅当缺乏明确的支持路径，且无法在常识与合理逻辑下补全时，才认为“无法判断”。

## 输入：

```
{argument}: {content_argument}
```

{evidence}: {content\_evidence}  
## 如果可以合理地判断该 {evidence}  
是否 {target} 该 {argument}, 请只输出  
空列表, 不要输出任何其他内容:

```
[]  
## 如果仍然无法判断, 请列出导致推理  
无法进行的关键问题及其理由, 格式如  
下 (严格按照 JSON 数组输出, 不要有  
注释或多余内容):
```

```
````  
[  
  {  
    "关键问题": "导致无法直接推理的  
关键问题",  
    "理由": "说明为何该信息对于判断  
与{argument}的关系是必要的"  
  }  
]  
````
```

判断标准补充说明:

- 如果{evidence}与{argument}之间可以在常识、背景知识或合理推理下建立有效支持路径, 即视为可以判断。
- 不要求{evidence}对所有其他可能性做出排除, 只需存在正向支持即可。
- 如术语不完全一致, 但指向概念高度重合 (如“坚守传统”与“文化身份认同”), 也可视为支持关系成立。

#### 分析判断 prompt:

你的任务是, 在已知 {slave} 与 {master} 的  
表达内容, 以及针对关键问题进行分析后,  
判断该 {slave} 是否支持 {master}。

请按照以下结构结合关键问题分步判断,  
并给出最终结论:

1. {master} 中的主张是关于什么 (主题、  
核心结论)?
2. {slave} 是否围绕同一主题展开?
3. 若 {master} 包含“价值评判”“因果关系”

或“比较标准”, 论据是否提供了逻辑支  
撑?

4. 是否存在未排除的其他解释或替代路  
径, 使 {slave} 难以唯一支持 {master}?

5. 最终, {slave} 是否能够有效支持/属于  
{master}? 请仅在逻辑链条完整、主张  
一致、没有重大缺口时判定为“是”, 否  
则判定为“否”。

注意:

- 若 {slave} 与 {master} 在逻辑、语义  
上存在支持关系, 即使未逐字复现, 也  
应视为支持;

- 支持关系不要求“唯一”“充分”或“直  
接”, 只要构成有效的支持路径即可。

## 输入信息如下:

```
<{master}>{content_master}</{master}>  
<{slave}>{content_slave}</{slave}>  
<已经指出的关键问题>  
{context}  
</已经指出的关键问题>
```

## 请严格按以下格式输出, 不要遗漏分  
割符号 <判断理由>:

是|否<判断理由> 这里给出你的判断理  
由

以下是一个输出示例:

是<判断理由>...

## 输出:

## 5.4 辩论构建任务 prompt

辩论构建微调数据构建 prompt:

```
"messages": [  
  {
```

```
    "role": "user",
```

```
    "content": f"## 任务要求: 请你  
判断当前内容在当前陈词稿中的成分:
```

```
\n## 输入: \n<陈词稿>{speech}</陈词  
稿>\n<需要判断的内容>{argument}</
```

需要判断的内容>\n##输出要求: 请给出以下三种类型的其中一种,不要给出任何其它符号及文本: 主要论点|论点前提|普通论点/no\_think"

```
    },  
    {  
      "role": "assistant",  
      "content":  
f"<think>\n\n</think>\n\n{label}"  
    }  
  ]
```

### 初步判断 prompt:

## 任务说明:

你作为分析员{kwargs['id']}, 需要判断下列“论点片段”在整篇陈词稿中的论证结构。请结合语义及陈词稿的结构从以下三类中选择最合适的一类, 并给出判断理由:

- 论点前提: 其存在是为了支撑另一个论点的成立, 脱离上层论点后会显得无从落脚。例如事实、背景、原因或条件等, 需结合其他论据理解。常以具体例子、经验描述、因果铺垫、定义、对比等方式呈现。

注意, 有例子并不一定是论点前提, 需要看其是否可以独立阐述一个观点。

- 普通论点: 可作为一个独立的观点阐述, 不依赖其他句子解释, 通常是与其他普通论点并列的, 在文本中通常是段落的主题句或句子主干, 逻辑地直接接近主要论点。

- 主要论点: 整篇文章的核心主张, 其他所有内容都围绕它展开, 通常为全文中心观点或一级论点。

##输入:

```
<陈词稿>  
{statement}  
</陈词稿>
```

<需要判断的内容>

```
{argument}
```

</需要判断的内容>

##输出要求: 请给出以下三种类型的其中一种: 主要论点|论点前提|普通论点, 并给出判断理由, 按照如下格式输出:

(主要论点|论点前提|普通论点)<判断理由>这里给出你的判断理由

以下是一个输出样例,仅供格式参考:

普通论点<判断理由>...

论点前提<判断理由>...

主要论点<判断理由>...

现在开始你的任务:

### 后续判断 prompt:

##任务要求: 你作为分析员{kwargs['id']}

需要和你的队友协作判断当前内容在当前陈词稿中的成分, 以下是论证成分的说明:

- 论点前提: 其存在是为了支撑另一个论点的成立, 脱离上层论点后会显得无从落脚。例如事实、背景、原因或条件等, 需结合其他论据理解。常以具体例子、经验描述、因果铺垫、定义、对比等方式呈现。

注意, 有例子并不一定是论点前提, 需要看其是否可以独立阐述一个观点。

- 普通论点: 可作为一个独立的观点阐述, 不依赖其他句子解释, 通常是与其他普通论点并列的, 在文本中通常是段落的主题句或句子主干, 逻辑地直接接近主要论点。

- 主要论点: 整篇文章的核心主张, 其他所有内容都围绕它展开, 通常为全文中心观点或一级论点。

注意, 你和其它分析员作为一个队伍,

最终的目的是为了协作分析出最合理的判断结果，你需要参考队友的判断，并指出其合理性或问题，同时给出你自己的最终判断。

##输入:

<陈词稿>

{statement}

</陈词稿>

<需要判断的内容>

{argument}

</需要判断的内容>

<队伍目前的分析判断历史>

{' '.join(kwargs['debate\_record'])}

</队伍目前的分析判断历史>

##输出要求: 现在需要你分析目前队伍的判断结果是否合理，判断需要判断的内容在陈词稿中的成分属于三种类型的哪一种，并给出你的判断理由:

按照如下格式输出:

(主要论点|论点前提|普通论点)<判断理由>这里给出你的判断理由

以下是一个输出样例,仅供格式参考:

普通论点<判断理由>...

论点前提<判断理由>...

主要论点<判断理由>...

现在开始你的任务:

**辩论 prompt:**

##任务要求: 你作为分析员{kwargs['id']}与其它队友分析员就某句陈词稿内容的论证结构判断存在分歧。现在请你再次审视该句，并决定是否修改你上一轮的

判断。以下是论证结构的说明:

- 论点前提: 其存在是为了支撑另一个论点的成立，脱离上层论点后会显得无从落脚。例如事实、背景、原因或条件等，需结合其他论据理解。常以具体例子、经验描述、因果铺垫、定义、对比等方式呈现。

注意，有例子并不一定是论点前提，需看其是否可以独立阐述一个观点。

- 普通论点: 可作为一个独立的观点阐述，不依赖其他句子解释，通常是与其他普通论点并列的，在文本中通常是段落的主题句或句子主干，逻辑地直接接近主要论点。

- 主要论点: 整篇文章的核心主张，其他所有内容都围绕它展开，通常为全文中心观点或一级论点。

注意，你和其它分析员作为一个队伍，最终的目的是为了协作分析出最合理的判断结果，你需要参考队友的判断，并指出其合理性或问题，同时给出你自己的最终判断。

##输入:

<陈词稿>

{statement}

</陈词稿>

<需要判断的内容>

{argument}

</需要判断的内容>

<队伍目前的分析判断历史>

{' '.join(kwargs['debate\_record'])}

</队伍目前的分析判断历史>

##输出要求: 现在需要你分析队友目前的判断结果是否合理，重新审视自己之

前的判断结果，决定是否需要修改自己之前的判断结果，并给出你的判断理由：

按照如下格式输出：

(主要论点\论点前提\普通论点)<判断理由>这里给出你的判断理由

以下是输出样例，仅供格式参考：

普通论点<判断理由>...

论点前提<判断理由>...

主要论点<判断理由>...

现在完成你的任务：

### 裁判分析 prompt:

##任务要求：请你判断当前内容在当前陈词稿中的成分：

- 论点前提：其存在是为了支撑另一个论点的成立，脱离上层论点后会显得无从落脚。例如事实、背景、原因或条件等，需结合其他论据理解。常以具体例子、经验描述、因果铺垫、定义、对比等方式呈现。

注意，有例子并不一定是论点前提，需要看其是否可以独立阐述一个观点。

- 普通论点：可作为一个独立的观点阐述，不依赖其他句子解释，通常是与其他普通论点并列的，在文本中通常是段落的主题句或句子主干，逻辑地直接接近主要论点。

- 主要论点：整篇文章的核心主张，其他所有内容都围绕它展开，通常为全文中心观点或一级论点。

##输入：

<陈词稿>

{statement}

</陈词稿>

<需要判断的内容>

{argument}

</需要判断的内容>

<队伍的分析判断历史>

{' '.join(kwargs['debate\_record'])}

</队伍的分析判断历史>

##输出要求：你的队友并没有得出一致的结论，现在需要你根据他们的观点进行最后的判断并给出判断理由：

按照如下格式输出：

(主要论点\论点前提\普通论点)<判断理由>这里给出你的判断理由

以下是一个输出样例，仅供格式参考：

普通论点<判断理由>...

论点前提<判断理由>...

主要论点<判断理由>...

现在完成你的任务：

## 5.5 自主论辩生成任务 prompt

自主论辩生成任务微调数据构建 prompt:

```
"messages": [  
  {  
    "role": "user",  
    "content": f"现在需要你站在 {stance} 的立场，为辩题: {topic} 生成一篇严谨，富有逻辑和说服力的陈词稿：  
/no_think"  
  },  
  {  
    "role": "assistant",  
    "content":  
f"<think>\n\n</think>\n\n{statement}"  
  }  
]
```

]

### 主论点生成 prompt:

你是一位经验丰富的议论文作者，擅长解构议论文主题并提出合乎逻辑且有说服力的论点。你的任务如下：

1. 分析给出的议论文主题，明确其议论核心。
2. 根据你的立场提出你要写的议论文的立场或核心思想。
3. 生成至少三个与核心思想密切相关的主要论点，确保每个论点都有清晰的逻辑和明确的推理方向。

Input:

文章主题:

Output (请严格遵循格式，不要添加任何不相关的内容、符号、标点符号。):

Topic Analysis:

Core Idea:

Argument 1:

Argument 2:

Argument 3:

以下是一个例子:

Input:

文章主题: 手机和互联网非常有用。然而，老年人很少使用它们。手机和互联网能以何种方式帮助老年人？我们如何鼓励老年人使用这项新技术？

立场: 正方

Output:

Topic Analysis: 本议论文题目的核心讨论范围是探讨手机和互联网对老年人的潜在好处以及如何鼓励他们采用这些技术。关键概念包括“智能手机”、“互联网”、“老年人”和“技术使用”。讨论方向侧重于分析手机和互联网如何对老年人的生活产生积极影响，并提出

促进其采用的策略。

Core Idea: 智能手机和互联网可以显著提高老年人的生活质量，应通过教育和技术支持加以鼓励。

Argument 1: 智能手机和互联网可以增强老年人监测和管理健康的能力。通过使用健康追踪应用程序和在线医疗咨询，老年人可以更好地监测自己的健康状况，及时获得医疗建议，提高生活质量和独立性。

Argument 2: 智能手机和互联网为老年人提供了丰富的社交和娱乐资源。通过视频通话、社交媒体和在线游戏，老年人可以与家人和朋友保持联系，减少孤独感，并通过在线教育和娱乐内容保持精神活跃。

Argument 3: 智能手机和互联网可以帮助老年人更好地融入现代社会。通过使用移动支付、网上购物和电子政务服务，老年人可以更方便地处理日常事务，减少因不熟悉技术而造成的社会排斥。

现在请完成你的任务:

Input:

文章主题: {topic}

立场: {stance}

Output:

### 自我问答 prompt:

辩题: {topic}

推理路径:

{por}

请分析辩题和推理路径，质疑推理路径末尾的主张，并提出一个更深入的新主张作为对问题的回答，利用该主张扩展推理路径。只输出你提出的问题 and 新的主张。

将提出的问题输出在 <question> 和 </question> 之间

将新的主张输出在 `<subclaim>` 和 `</subclaim>` 之间

以下是一个输出样例：

`<question>` 玄学是否真的重构了年轻人对现实的应对方式，还是只是用虚幻的掌控感掩盖了原子化社会中无法解决的结构困境？`</question>`

`<subclaim>`“拥抱玄学”实质是年轻人在系统性剥夺现实掌控权的背景下，通过将生存焦虑转化为非现实领域的意义争夺，反而加剧了现实无力感与认知割裂——当年轻人沉溺于星座占卜构筑的“命运预言”时，恰恰暴露出现代性焦虑的核心矛盾：个体无法在资本逻辑与技术理性主导的现实中确立稳定的价值坐标，只能在符号狂欢中短暂逃避对现实的反思。这种精神解脱本质是延迟性自我消解，而非创造性重构。`</subclaim>`

#### 论点整合 prompt:

请将以下主张和子主张整合成一个逻辑严密、具有说服力且更具深度的新主张。

主张: `{claim}`

子主张:

`{subclaims}`

将整合后的主张输出在 `<combinedClaim>` 和

`</combinedClaim>` 之间:

#### Toulmin 模型构建 prompt:

你是一位熟悉 Toulmin 论证结构的逻辑学家。现在你需要为你的论点构建 Toulmin 模型。

Claim: `{CurrentClaim}`

###以下是 Toulmin 论证结构元素的简要说明:

1.Claim: 作者用来传达和说服读者的主要主张，代表主要论点。

2.Grounds: 至少两条支持[Claim]的事实

证据。这些证据应来自具体且可信的来源，例如经过验证的具体统计数据或历史先例。必须明确说明每条证据的具体来源，而不是使用“某项研究”之类的模糊表述。

3.Warrant: 保证[Grounds]在逻辑上支持[主张]的明确或隐含的假设，而不是提供支持[Claim]的证据。

4.Qualifier: 表示[Claim]的强度和适用范围。

现在请按照以下格式给出当前 Claim 的完整 Toulmin 论证结构（请严格遵守以下输出格式，不要添加额外内容。不要添加任何不相关的内容、符号。）

`<Grounds>` 这里是 Grounds 的内容  
`</Grounds>`

`<Warrant>` 这里是 Warrant 的内容  
`</Warrant>`

`<Qualifier>` 这里是 Qualifier 的内容  
`</Qualifier>`

#### 辩论分析 prompt:

您是一位优秀的逻辑学家。基于 Toulmin 论证模型的原则，您的任务是分析双方的辩论发言，并优化正方的 Toulmin 论证模型。

请按照以下步骤完成任务:

1. 清楚理解正方的[Toulmin 论证模型]:

- 主张: `{Claim}`

- 依据: `{Grounds}`

- 论证: `{Warrant}`

- 限定语: `{Qualifier}`

2. 分析以下辩论记录，找出正方 [Toulmin 论证结构] 中的问题:

`{DebateRecord}`

3. 提供更具逻辑性和说服力的正方主张，并加入反驳元素（有效的反论点或反例来反驳[主张]）。

请严格按照以下格式提供修改后的 Toulmin 论证模型，注意不要丢失各部分的 </...> 结束标记符号，且避免添加任何不相关的符号或内容：

<Grounds> 这里是新的依据内容  
</Grounds>

<Warrant> 这里是新的论证内容  
</Warrant>

<Qualifier> 这里是新的限定语内容  
</Qualifier>

<Rebuttal> 这里是新的反驳内容  
</Rebuttal>

## 5.6 论辩赋能社会调查仿真 prompt

### 第一次调查 prompt:

###You are a real person with the following personal information. Please answer the following question as best as you can. You should act consistently with the role you are playing. Do not select the option to refuse to answer.

###Personal information: !<INPUT 0>!

###Question: !<INPUT 1>!

###Options: !<INPUT 2>!

###You should give your answer (you only need to answer the option letter number) in JSON format as example below:

```
``json
{
  "answer": "A"
}
```

### 第二次调查 prompt:

###You are a real person who has just participated in a debate. Please

rethink and answer the following questions based on the debate content. Your choices should be consistent with your real identity and take into account the views in the debate.

###Personal information: !<INPUT 0>!

###Debate content: !<INPUT 1>!

###Questions: !<INPUT 2>!

###Options: !<INPUT 3>!

###Please give your answer in JSON format, as shown below:

```
``json
{
  "answer": "A"
}
```

## 5.7 文章评估 prompt

主题: {Topic}

议论文: {Argumentative Essay}

请根据以下维度评估上述议论文:

**相关性 (1-5 分):** 评估文章与主题的相关性。评估标准包括: 文章中的所有论点是否与主题或中心论点相关, 论点是否足够多样且相互支持。

**逻辑性 (1-5 分):** 评估议论文的逻辑性。评估标准包括: 文章的论点是否清晰, 证据是否充分, 论证结构是否严谨, 不同论点之间是否具有一定的层次性和递进性。

**论点复杂度 (1-5 分):** 评估文章中论点的复杂性。评估标准包括: 论点是否涵盖多个视角或领域; 论点是否基于多层次的推理链条或相互支持的子论点。

**说服力 (1-5 分):** 评估议论文的整体说服力。评估标准包括: 文章是否清晰表达了论点; 是否使用了多样的证据 (如事实、数据、理论等) 并进行了深入分

析，以揭示问题的本质；论点是否具有一定的普适性，能够应用于更广泛的背景。

修辞 (1-5 分): 评估文章的语言风格。

评估标准包括: 文章是否使用清晰、准确的语言表达观点; 是否有效地运用了修辞手法 (如隐喻、修辞问题、排比等) 来增强说服力; 语言和结构是否能够吸引读者, 使文章更具吸引力。

总体评价 (1-5 分): 根据上述各维度的评估, 提供对文章的整体评价。

每个维度的评分范围应为 1 到 5 分, 允许使用小数精度。

首先简要反馈文章质量, 然后给出各维度的具体评分。

输出必须严格按照以下格式, 不能添加任何无关内容或符号:

<feedback> 反馈内容 </feedback>

<Relevance> 相关性分数 </Relevance>

<Logic> 逻辑性分数 </Logic>

<Complexity of Claims> 论点复杂度分数 </Complexity of Claims>

<Persuasiveness> 说服力分数 </Persuasiveness>

<Rhetoric> 修辞分数 </Rhetoric>

<Overall> 总体分数 </Overall>

## 6 参考文献

[1] Bao J, Wang Y, Li Y, et al. AEG: Argumentative essay generation via a dual-decoder model with content planning[C]//Proceedings of the 2022 conference on empirical methods in natural language processing. 2022: 5134-5148.

[2] Zhao Y, Huang J, Hu J, et al. Swift: a scalable lightweight infrastructure for fine-tuning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025,

39(28): 29733-29735.

[3] Yang A, Li A, Yang B, et al. Qwen3 technical report[J]. arXiv preprint arXiv:2505.09388, 2025.

[4] Press O, Zhang M, Min S, et al. Measuring and narrowing the compositionality gap in language models[J]. arXiv preprint arXiv:2210.03350, 2022.

[5] Toulmin S E. The uses of argument[M]. Cambridge university press, 2003.

[6] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.

[7] 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.

[8] Yao S, Yu D, Zhao J, et al. Tree of thoughts: Deliberate problem solving with large language models[J]. Advances in neural information processing systems, 2023, 36: 11809-11822.

[9] He Y, Bao J, Sun Y, et al. Decomposing Argumentative Essay Generation via Dialectical Planning of Complex Reasoning[C]//ACL (Findings). 2024.

[10] Hu Z, Chan H P, Li J, et al. Debate-to-Write: A Persona-Driven Multi-Agent Framework for Diverse Argument Generation[C]//Proceedings of the 31st International Conference on Computational Linguistics. 2025: 4689-4703.

# 融合分类与生成技术的中文论辩系统构建

赖逸斌<sup>1</sup> 傅博<sup>2†</sup>

(1. 新加坡管理大学, 新加坡 邮编 188065; 2. 中国科学院沈阳自动化研究所, 辽宁省沈阳市 邮编 110169)

**摘要:** 本技术报告介绍了我们在 CCAC 2025 第五届智慧论辩评测 (AI-Debater 2025) 中参加赛道一 (自主论辩任务) 和赛道二 (论辩赋能社会调查仿真) 的整体技术方案与实验结果。在赛道一的四个子任务中, 我们分别针对辩题识别、论据挖掘、辩论构建以及论辩生成等任务进行了解析并提出了相应方法。在前两个分类任务 (辩题识别、论据挖掘) 和辩论构建任务中, 我们主要采用了基于预训练模型的微调策略——以哈工大 MacBERT 大型模型为基础进行下游任务训练, 取得了较好的分类效果。在第四个生成任务中, 我们使用了大语言模型 (如 GPT-4 系列) 的 Prompt 策略, 将赛题提供的示例对话作为 Few-Shot 提示, 引导模型生成与示例风格一致的一辩立论稿, 效果最佳。在赛道二的仿真任务中, 我们基于给定的人口属性信息构造模拟角色, 并采用多智能体辩论机制生成更精确的调查问卷答案。具体而言, 借助 ChatGPT 4.1 API 模拟多轮角色对话, 通过角色间的观点碰撞迭代出更具鲁棒性的结果。最终, 我们在赛道二任务一获得三等奖, 任务二获得二等奖。本报告详细阐述了工作背景、相关研究、技术路线、实验设置及结果分析等内容, 并附以参考文献以供进一步学习。

**关键词:** 智慧论辩; 预训练模型; 多智能体仿真

中图分类号: TP391

文献标识码: A

## Building a Chinese Argumentation System with Hybrid Classification and Generation Approaches

Eason Lai<sup>1</sup>, Fu Bo<sup>2†</sup>

(1. Singapore Management University, Singapore, 188065; 2. Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, Liaoning, 110169, China)

**Abstract :** This technical report presents our overall technical solutions and experimental results for Track 1 (Autonomous Debating Task) and Track 2 (Debate-Empowered Social Survey Simulation) in the CCAC 2025 5th Intelligent Argumentation Evaluation (AI-Debater 2025). For the four subtasks in Track 1—namely argument topic identification, argument mining, debate construction, and argument generation—we analyzed each task and proposed corresponding methods. In the first two classification tasks (argument topic identification and argument mining) and the debate construction task, we primarily employed a fine-tuning strategy based on pre-trained models, using Harbin Institute of Technology's MacBERT large-scale model for downstream task training, achieving strong classification performance. For the fourth generation task, we utilized a prompt strategy with large language models (such as the GPT-4 series), treating the example dialogues provided in the task as Few-Shot prompts to guide the model in generating opening statements consistent with the example style, yielding optimal

results. In Track 2's simulation task, we constructed simulated roles based on given demographic attributes and employed a multi-agent debate mechanism to generate more precise survey responses. Specifically, we used the ChatGPT 4.1 API to simulate multi-turn role-based dialogues, iteratively refining more robust results through viewpoint collisions among the roles. Ultimately, we secured third place in Track 1 and second place in Track 2. This report elaborates on the background, related work, technical approaches, experimental setup, and result analysis, along with references for further study.

**Key words:** Intelligent Argumentation; Pre-trained Models ; Multi-Agent Simulation

## 0 引言

论辩作为人类认知能力的一项重要体现，在诸多社会活动中发挥着关键作用。近年来，随着自然语言处理技术的发展，计算论辩（Computational Argumentation）逐渐成为研究热点，其目标是让机器能够理解、生成和评估论据、观点等辩论相关内容。CCAC 智慧论辩评测（AI-Debater）即为该领域的重要赛题之一，旨在推动机器论辩技术的发展并评估模型在实际论辩场景中的能力。AI-Debater 2025 共有两个赛道：赛道一“自主论辩任务”包含四个子任务，分别对应不同层面的论理解理解和生成问题；赛道二“论辩赋能社会调查仿真”旨在将论辩机制应用到社会调查预测中，通过多智能体辩论生成更精准的调研回答。

赛道一的四个子任务具体如下：子任务 1（辩题识别）为分类任务，判断给定论点是否属于指定辩题；子任务 2（论据挖掘）为分类任务，判断给定论据是否能够支持论点；子任务 3（辩论构建）为多分类任务，判断论点在立论稿中的结构类型（如主要论点、普通论点、论点前提等）；子任务 4（自主论辩生成）为生成任务，要求根据给定辩题和持方（正方/反方）自动生成一篇立论陈词稿。赛道一的前三个任务使用准确率（Acc）评测，第 4 个任务则通过人工评分。赛道二的任务是模拟具有不同人口属性的角色进行多轮辩论，最终输出对社会调查问卷的回答，评测包括多智能体辩论得分（如 Debatrix 框架）和回答的准确率（F1）。赛道二的流程通常包括：

第一步，根据人物信息（如年龄、性别、族裔、政治倾向等）生成角色并回答指定问题；第二步，引导角色进行相关思考提问并回答；第三步，基于第二步的思考结果，再次询问初始问题，观察角色回答是否发生改变，最后输出最终答

案。

在比赛过程中，我们尝试了多种技术方案：对于分类任务，我们使用 MacBERT 等中文预训练语言模型进行微调；也曾尝试过使用 Qwen3 模型外接 LoRA 微调、或直接采用大语言模型（如 GPT-3/4）零样本测试等，但实验结果表明基于 MacBERT 重新训练在分类任务中效果最佳。对于辩论生成任务，我们直接利用预训练的大语言模型（例如 GPT-4）进行 Few-Shot 提示，结合训练集中提供的示例对话，使生成结果风格与示例一致，从而显著提升了立论稿的质量。在赛道二的仿真任务中，我们调用了 ChatGPT 4.1 的 API 接口，模拟不同属性的角色并让他们通过多轮辩论回答调研问题，这一策略得到了最佳的比赛成绩。

本报告将按照学术论文的格式，先介绍相关工作背景，然后详细描述我们的技术路线与实验方法，再呈现实验结果与分析，最后总结创新点与未来改进方向。报告中使用了相关文献支撑方法选择与实验结论，以提高报告的可信度与可参考性。

## 1 相关工作

### 1.1 计算论辩与论点分析

计算论辩（Argumentation Mining）是自然语言处理中的一个研究方向，涉及从文本中识别论点、论据及其相互关系。例如，一项常见任务是观点挖掘（argument mining），旨在从文本中自动提取人们的观点和论据。

过去的研究多集中于英文语境下的观点挖掘、论证结构识别等领域，但中文的相关工作相对较少。传统方法主要依赖特征工程和浅层分类器，而随着深度学习的发展，基于深度预训练模型的方法成为主流，例如使用 BERT、RoBERTa 等模型对论点类型进行分类。我们的任务 1 和任务

2 可以归结为二分类问题（论点相关性判断和论据支持判断），任务 3 则是三分类任务。类似的论题识别与论据挖掘研究表明，基于 BERT 系列预训练模型的微调方法效果较好<sup>1,2</sup>。尤其是专为中文语料设计的 MacBERT 模型，在多项中文 NLP 任务上都取得了最先进的性能<sup>1</sup>，因此成为我们首选的基模型。

对于辩论结构识别（任务 3），过去的研究通常参考辩论学和修辞学中对立论稿结构的划分：主要论点、次要论点以及论点前提（支撑性论点）等。在中文情境下，这样的细粒度分类需求还较少研究，因此我们将其看作单句分类任务，使用相同的预训练模型策略。相关工作中，也有学者使用多分类器对辩论句子的角色进行判定，但大多集中在英文文本上。

## 1.2 自然语言生成与辩论文稿

任务 4 要求生成具有辩论风格的立论陈词稿，这属于开放式文本生成问题。在生成式任务中，近年来大型语言模型（LLM）如 GPT-3/4 被广泛用于生成高质量的人类类辩论内容。有研究指出，通过在提示中给出示例（few-shot prompting），可以显著提高模型生成结果的相关性和一致性<sup>3</sup>。我们借鉴这一思路，将比赛提供的对话示例（不同辩题与对应正反方立论稿）作为提示示例，指导 GPT-4 生成新的立论稿。类似地，很多竞赛和应用也采用将示例对话作为提示来生成仿真答案的策略，这种方式能够让模型“学习”示例中的表达风格和论证结构，从而生成符合预期风格的输出<sup>3</sup>。

## 1.3 大模型多智能体和社会仿真

赛道二的特点是将多轮辩论引入到人口调查预测中，需要模拟不同人口属性角色的思维和态度，并通过多智能体交互生成更精准的回答。**社会智能体仿真**是一个新兴领域，研究如何利用大型语言模型模拟人类群体行为和社会互动。最新综述指出，通过大模型构造个体角色并让其进行讨论，可以模拟群体意见的动态演变<sup>4</sup>。具体到辩论场景，多轮、多角色的对话能够增强回答的合理性和准确性。例如，Du 等人提出的多轮辩论（MAD）框架，让不同立场的 LLM 在多轮中交换观点，在解决复杂推理任务时取得了更好的结果。他们发现，经过多轮辩论，模型的数学推理和事实准确率显著提高，逻辑错误和幻觉（hallucination）显著减少。这些研究表明，模拟多个具有不同观点的角色相互论辩，有助于

获取更高质量的回答。

此外，针对多智能体仿真的研究也强调了角色扮演（persona simulation）的重要性：通过向 LLM 提供完整的人物背景信息（如人口统计特征、性格描述等），可以生成符合该角色的人物信息档案和答案。在赛道二中，我们正是基于这一思路，通过向 ChatGPT 提示“你的属性是……，请像该角色一样回答”等，得到与指定属性一致的初步回答。然后再引入辩论机制，让模拟角色之间就问卷话题展开讨论，以促进观点碰撞与澄清，最终产生更准确的调研回答。这种做法契合了多智能体和社会仿真的研究趋势，即利用大模型代替实际群体，通过交互增强结果可靠性。

综上，我们的方案结合了**预训练模型微调**、**Prompt 示例生成**和**多智能体辩论**等技术思路。这些方法在相关文献中均有体现：MacBERT 模型被证明在中文 NLP 任务中表现优异<sup>1</sup>，Prompt 加示例的 few-shot 方法能显著提升生成质量<sup>3</sup>，而多角色辩论可提高推理准确率并减少错误<sup>4</sup>。下面，我们将在方法部分详细阐述这些策略的具体实现。

## 2 方法

本节分为赛道一（自主论辩任务）和赛道二（仿真任务）两大部分，分别介绍每个子任务的技术路线和实现细节。

### 2.1 赛道一：自主论辩任务

#### 2.1.1 任务概览

赛道一包括四个子任务。前三个子任务（辩题识别、论据挖掘、辩论构建）均为分类问题，最后一个（自主论辩生成）为文本生成问题。下面逐一描述我们的解决方案。

##### （1）子任务一：辩题识别

输入为辩题文本和论点文本，任务是判断该论点是否属于给定辩题。形式化地，可看作二分类任务：若属于该辩题，输出 1；否则输出 0。

##### （2）子任务二：论据挖掘

输入为论点文本和论据文本，任务是判断该论据是否可以支持当前论点，形式上也是一个二分类问题：支持输出 1，不支持输出 0。

##### （3）子任务三：辩论构建

输入为一篇立论陈词稿和其中一个论点句子，任务是分类判断该论点的结构类型，包括“主

要论点”、“普通论点”和“论点前提”三类。该任务为三分类问题，输出直接为结构类型文本。

#### (4) 子任务四：自主论辩生成

输入为辩题和持方（正方或反方），需要自动生成一篇符合辩论规范的正反方立论陈词稿开头（约几句话到一段文字）。输出为生成的中文立论稿文本，最终由人工进行内容质量的评价。

### 2.1.2 数据处理与预处理

比赛组委会提供了每个任务的训练集（以 JSONL 格式给出）。我们首先使用 datasets 库（Hugging Face）加载数据集，并对字段进行重命名和拼接处理。对于分类任务，由于输入包括两个文本字段（辩题+论点或论点+论据），我们在预处理函数中将两部分用特殊标记 [SEP] 拼接成单一序列。例如，辩题识别任务中将“辩题+[SEP]+论点”作为模型输入；论据挖掘任务中将“论点+[SEP]+论据”输入模型。在辩论构建任务中，只需输入论点句子本身，因其已成为独立的文本。预处理后，我们使用 MacBERT 的分词器（hf1/chinese-macbert-large）对文本进行分词和编码。

### 2.1.3 模型选择与微调

对于前三个分类任务，我们均采用 BertForSequenceClassification 模型架构，并以 MacBERT 大型模型（hf1/chinese-macbert-large）为预训练权重初始化。MacBERT 是针对中文设计的预训练模型，其创新点在于在遮蔽语言模型（MLM）任务中使用同义词替换的校正策略（Masking as Correction）缓解预训练与微调的差异。实验和文献表明，MacBERT 在中文句子分类、阅读理解等任务上表现优异<sup>1</sup>，因此我们选用它作为基础模型。

训练过程中，我们对每个任务使用相同的超参数设置：学习率约  $2e-5$ ，批量大小 16，训练轮数 9 次，Adam 优化器，交叉熵损失函数，并在开发集上评估准确率（Acc）。训练时采用 20% 数据作为验证集（train\_test\_split），其余 80% 作为训练集。训练代码使用 Huggingface Trainer 框架实现，设置评价策略为每轮评估一次，并在最后保存最佳模型。训练任务过程中，注意到 GPU 资源重要性，为充分利用可用 GPU，我们设置使

用两张 H100 型号 GPU 参与训练，但由于训练集规模适中（数千条样本级别），通常单机单卡也可快速收敛。

### 2.1.4 模型评估与结果

训练完成后，我们在测试分割集上评估每个模型的分类准确率。虽然具体数值在报告中不予公布（比赛可能出于保密考虑），但可以总结经验：MacBERT 微调模型在任务 1 和任务 2 上都获得了较高的准确率，远高于随机或简单的词向量模型。以任务 3 为例，重新训练的 MacBERT 模型能够准确区分“主要论点”、“普通论点”、“论点前提”三种结构类型。这表明预训练语言模型能有效捕捉辩论文本中的深层语义和结构信息。

为了探索更优方案，我们还尝试了以下方法：

(1) Qwen3+LoRA 微调：Qwen3 是新兴的大型中文多模态模型，我们尝试在其上使用 LoRA 技术微调分类任务，但由于数据量有限，微调效果未优于 MacBERT 基础模型。LoRA 虽然轻量，但在模型基础与数据适配上存在差距；

(2) GPT-4 零样本测试：直接向 GPT-4 等模型提问，要求其判断论点与论题的关系等。实验发现 GPT-4 在无示例情境下表现不稳定，可能会出错或给出不一致答案；

(3) Electra 等其他模型：基于 hash code 尝试使用 HFL 的 Electra 模型替代 MacBERT，但在调试过程中未发现显著提升。综合比较，我们最终采用 MacBERT 为主要方案并加以微调，实现了赛道一前三个子任务的系统化解决。

### 2.1.5 子任务四：自主论辩生成

该任务属于生成式 NLP 问题，目标是根据给定辩题和正反方立场生成一段立论陈词稿（开头几句话）。比赛数据集提供了一些示例对话（辩题与对应立论稿开头）。我们的做法是直接调用大语言模型（如 GPT-4 系列），并将训练集中提供的对话示例作为 Prompt 的一部分，采用 few-shot 方式指导生成。具体做法如下：

- **示例选择：**从训练集中抽取多个代表性辩题及其正方或反方立场开头示例，对示例内容进行适当裁剪，仅保留关键开头段落。
- **Prompt 设计：**构造提示模板，包含示例对话和新的任务指令。例如：

以下是一些辩题和正方立场的立论陈词稿开头示例：

辩题：“死刑是否应该合法化”，正方：尊敬的评委、对方辩友...（示例 1 开头）

辩题：“应不应该增加科研经费”，正方：感谢

主席、评委... (示例 2 开头)

...

现在请根据以下辩题和立场，生成相似风格的立论陈词开头：

辩题：“<新辩题>”，正方/反方：

- **生成：**将上述 Prompt 输入 ChatGPT 4.1 或 GPT-4 API，让模型输出完整的陈词开头。由于前文提供了几个示例，模型能够“模仿”其风格，生成结构相似的段落，如包含称呼评委、开场观点立场、后续论点概述等。

实验结果表明，该示例提示策略显著提升了生成质量。正如相关工作指出，GPT-4 等最新 LLM 在 few-shot 引导下具有很强的推理和生成能力，可以生成接近人类水平的回答<sup>3</sup>。在我们尝试中，未采用示例对照时，模型生成的内容常出现逻辑不连贯或偏题的情况；而添加了示例提示后，模型输出的立论稿格式更加规范，重点突出，观点也更切题。这验证了使用示例提示指导生成的有效性<sup>3</sup>。最终，我们提交了模型输出的立论稿，由人工评审评价其论证力度和语言质量，结果良好。

## 2.2 赛道二：辩论赋能社会调查仿真

### 2.2.1 任务概览

赛道二的目标是利用多智能体辩论机制来提高人口调查问题的回答质量。给定每位“个体”的人口属性（如年龄段、性别、种族、政治党派支持等）和一个调查问卷问题（含多个选项），需要模拟该属性对应的角色回答问题。比赛过程分为三步：第一步让角色给出初步回答；第二步通过多轮辩论或提问促进角色思考；第三步角色重新回答问题作为最终答案。评价指标包括 50% 的多智能体辩论得分（例如使用 Debatrix 多维辩论评估框架）和 50% 的最终回答 F1 分。

### 2.2.2 角色构建与初始回答

我们首先根据给定的人口属性信息构造角色。以赛题示例为例，如果角色信息为“36-65 岁男性，支持某党派”，我们会向 ChatGPT 提供这些信息并要求其扮演该角色。具体做法是构造一个角色描述 Prompt，例如：“你是一个 XX 岁、男性、支持 XX 党派的人，请回答以下问题：‘有些人认为我们应该减少国防开支，你怎么看？’”然后调用 GPT-4.1 生成该角色的回答。根据相关研究，大型语言模型能够基于提示生成符合期望人格特征的回答<sup>4</sup>。这一过程相当于为模型“注入”个体角色设定，使其输出更贴合该角色的真实可能回答。

### 2.2.3 多智能体辩论机制

得到各角色的初步回答后，我们进入第二步：促进角色之间的辩论与思考。在这个阶段，我们将之前模拟的多个角色“召集”到一个多智能体对话中，让他们围绕问卷主题展开讨论。具体策略包括：

- **分角色提问与观点：**对话中让各个角色轮流陈述观点、提出质疑或支持其他角色意见。例如，对于“减少国防开支”问题，某角色可能提出“减少开支有助于民生”，而另一角色可能反驳“这可能影响国家安全”；然后角色针对质疑进行再辩论。

- **引导思考：**对话过程中，我们适时插入附加问题或提示，让角色对自己的观点进行进一步论证。例如，对提出减少国防开支角色问：“如果国家安全受到威胁，你怎么平衡民生？”此类问题促使角色考虑多角度因素，从而使最终回答更加深入理性。

- **多轮迭代：**角色对话持续若干轮，直到观点趋于稳定或无更多问题提出。每轮结束时，我们记录角色的观点变化以及各选项的支持情况。

这一多轮多角色的设置借鉴了多智能体辩论的概念。正如前人研究表明，通过多轮辩论可以明显提升推理质量和信息准确率<sup>4</sup>。MAD 框架等工作让不同 LLM 角色在多轮讨论下打磨答案，有效减少了错误信息（幻觉）<sup>4</sup>。在我们的实现中，多个角色从不同角度互相对抗或说服，最终输出的观点往往比单一静态回答更全面。实际上，Debatrix 等多维评判框架也强调了分析多轮辩论内容的重要性<sup>5</sup>，这与我们模拟多轮辩论的思路相吻合。

### 2.2.4 最终回答和评估

辩论结束后，我们重新向每个角色提出原始问题，让其根据刚才的讨论重新给出答案选项。这一过程检验了角色观点的一致性与深思熟虑程度。最后，将各角色的最终答案进行汇总，或直接选择一个主角的答案作为模拟结果。我们使用比赛标准方法计算最终答案的 F1 分数，以及将辩论内容提交给评审系统获得多智能体辩论得分。

实验证明，在赛道二任务中，仅让模型根据属性直接回答往往回答倾向简单直观，而引入辩论后角色能够提出更多细节。结果显示，多轮辩论显著提升了答案质量：角色在第二次回答中更倾向于给出经过辩论推敲后的答案。综上，通过

模拟多角色辩论，我们实现了比直接一问一答更精准的模拟问卷回答。这一实践与现有研究结论一致：多智能体互动能够提高生成内容的可靠性和准确度<sup>4</sup>。

### 3 实验与结果分析

#### 3.1 赛道一结果分析

在赛道一的分类型任务中，我们将各子任务的数据按照 80:20 比例划分为训练集和验证集，进行模型的训练与调优。表 1 给出了各任务在验证集上的最佳准确率（取最佳模型 checkpoint）。可以看到，MacBERT 微调模型在辩题识别和论据挖掘任务上准确率分别达到 95.78% 和 78.91%，在辩论构建任务上也取得了高于基线的性能 66.82%。实验结果表明，预训练模型在处理辩题相关分类问题时具有很好的泛化能力。

| 任务   | 类别数 | 验证集 Acc |
|------|-----|---------|
| 辩题识别 | 2   | 0.95    |
| 论据挖掘 | 2   | 0.78    |
| 辩论构建 | 3   | 0.66    |

表 1：赛道一子任务模型在验证集上的准确率。

从实验过程中发现，训练初期模型准确率上升较快，但在后期趋于平稳。在验证曲线上，我们使用 load\_best\_model\_at\_end 策略自动保存最高准确率对应的模型，有效避免了过拟合。训练日志显示，在第 7 至 9 轮左右达到最佳，一般超过 10 轮后性能无显著提升。

与其他方案对比中，我们发现：直接使用大语言模型（如 GPT-4）进行零样本分类任务时，模型的判断一致性较差，有时会出现过度解释或输出错误类别标签。而微调 MacBERT 等小模型在资源使用上更高效，也便于复制。Electra 等其他中文预训练模型（如哈工大中文 Electra）稍逊于 MacBERT，推测是因为 MacBERT 的 MLM 校正策略更适合中文特点。

**生成任务（子任务 4）**方面，我们对生成内容的质量进行人工评估和内容审查。两位评审按照论点清晰度、论证结构合理性以及语言流畅性等指标打分。与直接提示 GPT 相比，few-shot 示例提示的模型生成结果更契合要求：常见格式为：“尊敬的评委、对方辩友：今天我们讨论的辩题是‘XXXX’，我方观点是……首先……其次……最后……（据此展开细节）”。生成的文本中往往包含了“尊敬的评委”等礼貌用语以及明确的观

点层次结构。示例中数据也显示，使用示例提示时生成结果被评为“优秀”、“良好”的比例明显提高。

#### 3.2 赛道二结果分析

在仿真任务中，我们将每个角色的对话过程和最终回答提交给评审系统。评测包括多智能体辩论得分（Debatix 框架自动评分）和最终回答 F1 两部分。整个测评流程分为三步，在具有验证集的第一步我们的准确率达到 68.97%，我们的整个系统在比赛结果中获得了第二名，表明所采用方法具有较强竞争力。下面分析几点发现：

**多轮辩论效果显著：**与只进行一次提问相比，多轮让角色思考与互相辩论后，角色的最终回答更加合理。例如，在“减少国防开支”问答中，一些角色原先倾向于选“减少”，但经过辩论后更倾向选“维持”或“增加”，因为意识到安全风险。这种观念转变反映了辩论中获悉新信息的效果。对应地，多轮辩论得分高于单轮场景，验证了多智能体辩论可以提升回答质量的结论。

**角色扮演准确性：**ChatGPT 在扮演不同人口属性角色时，一般能生成符合预期的回答风格。例如，年轻角色更可能提及创新和未来观点，老年角色更强调传统和稳定。这样的模拟符合等研究所所述的 LLM 角色扮演能力。这使得答案多样性增加，模拟问卷结果更加符合社会实际分布。

**模型选择：**我们对比了不同模型版本对仿真任务的影响。结果显示，GPT-4.1 相比 GPT-3.5 在模拟角色思考和辩论对话时表现更为连贯；而与纯规则或小模型相比，大模型拥有更多知识和推理能力，对辩论议题的理解更深入。因此我们最终采用了 OpenAI 的 GPT-4.1 接口。

总体而言，赛道二的方法创新点在于结合了人格化模拟与多轮辩论：我们不仅让模型“成为”具有指定属性的角色，还让这些角色相互论辩、共同形成答案。这种结合正体现了最新社会仿真研究的趋势<sup>4</sup>，使得最终结果更贴近真实人群的态度分布。

### 4 结论与讨论

本报告总结了我们在 CCAC 2025 AI-Debater 比

赛中的技术方案。总体而言，主要贡献和结论如下：

在**赛道一**任务中，我们验证了对中文论辩相关分类问题使用 MacBERT 微调的有效性<sup>1</sup>。与其他尝试相比，预训练模型微调能充分利用比赛数据，快速适配任务需求。同时，在生成式任务中，采用示例 Prompt 指导 LLM 生成辩论陈词的策略亦取得了良好效果<sup>3</sup>。

在**赛道二**仿真任务中，我们展示了将多智能体辩论应用于社会调查预测的可行性：通过构建符合人口特征的角色并进行多轮讨论，使得模型对问卷问题的回答更符合实际<sup>4</sup>。这一过程借鉴了多智能体社会仿真的研究成果，证明了辩论机制在提升数据质量方面的潜力。

本次工作虽然取得了二、三等奖的佳绩，但仍有改进空间。例如，赛道一分类任务中引入更多中文预训练模型对比，或针对辩论文稿生成进行结构化约束和逻辑检查；赛道二仿真中可探索更多微调提示策略或加强记忆机制，以提高角色的一致性和多轮辩论的深度。此外，如何在资源受限环境中高效调用大模型，也是值得研究的工程方向。

**创新点**方面，我们工作结合了多种前沿技术：使用 MacBERT 等中文模型进行辩论特定任务微调，将 Few-Shot 提示方法应用于中文论辩文本生成，以及引入多智能体辩论到社会调查仿真。这些方法在相关文献中已有支持<sup>1,4</sup>，但在本次比赛中验证了其实际效果。此外，我们通过实际竞赛场景集成了上述技术，取得了优异成绩，说明综合应用现代 NLP 技术能有效推进智能辩论系统的发展。

在未来，我们计划将方法进一步推广：例如，扩展立论稿生成的深度，让生成模型给出完整立论结构而非仅开头；在仿真任务中，探索多智能体辩论与社会科学模型的结合，以实现更大规模的社会态度模拟。以上工作可为 AI 辩论系统和社会仿真系统的发展提供有益借鉴。

## 参考文献

- [1] Zhang H, Xu Y, Qiu X. A Memory Network Approach for Story-based Temporal Summarization of News Articles[C]// Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, 2020: 645 - 655. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.58/>
- [2] Lv G. StarrySky: Structured Annotation and Representation for Dialogue [EB/OL]. [2023-12-10].

<https://github.com/lvguichen1/StarrySky?tab=readme-ov-file>

- [3] Chen X, Liu Z, Wang H. Contrastive Instruction Tuning for Multi-turn Dialogue Agents[EB/OL]. arXiv:2501.06658v1 [cs.CL], 2025-01-15. <https://arxiv.org/html/2501.06658v1>
- [4] Wang Y, Li J, Zhao Q. Large Language Models are Few-Shot Table Reasoners via Schema Translation[EB/OL]. arXiv:2412.03563v1 [cs.CL], 2024-12-05. <https://arxiv.org/html/2412.03563v1>
- [5] Liu K, Zhang W, Zhou M. TreeCoder: Structured Pretraining for Code Generation with Hierarchical Representations[EB/OL]. arXiv:2403.08010 [cs.SE], 2024-03-13. <https://arxiv.org/abs/2403.08010>



赖逸斌（1986—），研究员，主要研究领域为研究方向包括虚假信息、多模态推理、常识推理与情感分析。  
E-mail: [eason.ybl.lai@gmail.com](mailto:eason.ybl.lai@gmail.com)



傅博（1987—），通信作者，博士，副研究员，主要研究领域为多模态数据分析、计算机图像处理、具身人工智能。  
E-mail: [fubo@sia.cn](mailto:fubo@sia.cn)

本次比赛主要基于 Llama-7B 基础上做 lora 微调，在四个子任务上，设计了不同的长思维链 prompt。

## 辩论构建

"instruction": ""# 辩论论点结构分析任务## 输入规范{ "陈词稿": "(包含完整论述结构的文本)", "论点": "(待分类的具体句子)"}## 分析框架 1. \*\*定位分析\*\*： - 在陈词稿中定位论点出现的段落位置(开头 30%/中间 40%/结尾 30%) - 统计论点在全文中的出现频率 2. \*\*语义权重评估\*\*： - 核心价值密度：计算论点中包含的辩论核心词比例(如"合作"在竞争/合作辩题中) - 论证强度：检测是否包含以下强化标记： \* 程度副词："根本性"/"至关重要" \* 逻辑连接词："因此"/"由此可见" \* 证据类型：研究数据/权威引用 3. \*\*结构关系验证\*\*： ||| 主要论点特征 ||| - 通常出现在段首/段尾 - 包含辩题核心关键词 - 有至少两个普通论点支撑 ||| 普通论点特征 ||| - 使用举例说明("例如")或因果论证("因为") - 修饰主要论点的某个方面 ||| 论点前提特征 ||| - 包含假设性表述("如果...") - 为其他论点提供逻辑基础 - 常伴随必要性论述("必须首先...")## 输出规范<结构类型>""

## 论据挖掘

### 分析步骤

1. **核心主张提取**： 论点核心： "{claim}" 论据核心： "{evidence}"
2. **语义关联分析**：
  1. 论据是否直接提及论点中的关键词汇？（如：合作/人际交往技能）
  2. 论据是否扩展了论点的具体实现方式？（如：倾听/沟通/妥协）
3. **逻辑支持验证**：
  1. 若论据包含研究/数据/权威引用 → +1 分
  2. 若论据展示因果关系 → +1 分
  3. 若论据提供具体行为示例 → +1 分
4. **反事实检验**： 假设论据为假时：
  1. 论点可信度是否显著降低？是→+1 分

### 评分规则

- 总分≥3 分 → 支持关系成立
- 总分<3 分 → 支持关系不成立

## 最终输出

仅输出单个数字: `{{ "support": <0|1> }}`

## 实例分析

论点: "通过合作, 孩子们可以学习人际交往技能" 论据: "研究表明...妥协" → 得分: 关键词(1)+研究(1)+示例(1)+反事实(1)=4 → 1

当前任务: 论点: "{input\_claim}" 论据: "{input\_evidence}" ""

prompt = ""请执行论点-论据支持性分析, 严格按步骤输出 0 或 1

## Prompt 2

### 分析步骤

#### 1. 核心匹配

- 论点关键词: {提取 claim 中的核心术语}
- 论据覆盖度: {论据中匹配到的关键词比例} → 若<50%匹配: 直接返回 0

#### 2. 逻辑验证 (需全部满足)

- 论据与论点存在因果链
- 无非相关干扰信息
- 无明显逻辑谬误 (如诉诸情感)

#### 3. 证据强度 (满足任一项即可)

- 含研究数据/统计结果
- 权威机构/专家背书
- 可验证的具体案例

#### 4. 反事实测试

- 假设否定论据时:  论点明显不成立 → +1 分  论点仍可能成立 → -1 分

## 最终决策规则

if 核心匹配 $\geq$ 50% AND 逻辑验证通过 AND (证据强度 $\geq$ 1 OR 反事实得分 $>$ 0):

→ 输出 1

else:

→ 输出 0

## 严格输出格式

必须且只能输出单个数字:

0

或

1

## 示例

论点: "疫苗有效预防疾病"

论据: "CDC 数据显示辉瑞疫苗对 Omicron 有效率 92%"

→ 1

论点: "吃巧克力能减肥"

论据: "因为我朋友喜欢吃"

→ 0

""

## 辩题识别

由于该题目比较简单,使用赛题介绍做 prompt,就能达到 93%+的准确率。Prompt: "给定一个论点和一个论据,判断当前论据是否可以支持当前论点,可以支持当前论点输出 1,不可以支持当前论点输出 0。"

”

# 自主论辩生成

分别测试了 DeepSeek-R1 和 Llama-7B 做 lora 后的结果，发现 7B 的模型就已经能生成较为高质量的辩论稿。

# 四格漫画逻辑理解

王硕<sup>1</sup> 刘志林<sup>1</sup> 郑安遇<sup>1</sup>

(1. 河北大学 数学与信息科学学院, 河北 保定 071002)

**摘要:** 该文针对四格漫画顺序推理任务中多模态理解能力不足的问题, 提出一种基于大语言模型推理能力蒸馏的小模型训练方法。该方法使用已有的包含中文和英文漫画的多语言数据集, 利用大规模多模态语言模型生成漫画图像的对白文本和视觉描述, 并输出结构化的推理过程与顺序标签作为训练样本。通过将该类推理数据用于小模型训练, 实现对白与图像细节的联合建模和顺序识别能力增强。实验在测试集中分别评估中文漫画和英文漫画的顺序识别性能, 结果表明该方法在多语言、多模态漫画顺序判断任务中具有良好的效果和适应性。

**关键词:** 四格漫画理解; 多模态推理; 大模型蒸馏

## Logical understanding of four-panel comics

Shuo Wang<sup>1</sup>, Zhilin Liu<sup>1</sup>, Anyu Zheng<sup>1</sup>

(1. School of Mathematics and Information Science, Hebei University, Baoding, Hebei 071002, China)

**Abstract :** A method was proposed to enhance multimodal understanding in the task of four-panel comic strip ordering. The approach distilled reasoning abilities from large multimodal language models into a compact model. A multilingual dataset containing Chinese and English comics was used. Visual-language models generated dialogue texts, visual descriptions, structured reasoning steps, and sequence labels from comic images. These outputs served as training samples. The distilled model learned to jointly model dialogues and visual details for improved sequence prediction. Experiments evaluated performance on Chinese and English comic test sets. The results show that the method achieves good accuracy and adaptability in multilingual, multimodal sequence recognition.

**Key words:** Four-panel comic understanding; Multimodal reasoning; Large model distillation

## 0 标题

在数字化传播日益多样化的背景下，漫画作为一种兼具图像与语言的叙事媒介，正广泛应用于社交平台、新闻媒体与数字出版等多个领域。四格漫画以其轻松幽默、结构简洁、节奏明快的特性深受用户欢迎，成为人们表达情绪、讲述故事的重要载体。在漫画传播过程中，四格的排列顺序对于读者理解其叙事逻辑、抓住故事转折及笑点具有关键影响。然而在实际应用中，四格漫画往往由于截图、裁剪、平台转发等操作出现顺序被打乱的现象，严重影响读者的理解与传播效果。因此，如何自动识别并纠正四格漫画的顺序，成为一个具有挑战性且亟待解决的人工智能图理解任务。

尽管近年来多模态理解技术取得显著进展，但现有工作多数聚焦于图文匹配、图文检索或图文生成任务<sup>[1][2]</sup>，在“图文联合推理+顺序结构建模”方向研究仍显薄弱。四格漫画顺序识别不仅需要理解每一帧图像中的视觉动作，还需分析角色对白之间的语义逻辑，捕捉起承转合的情节变化。这一任务对多模态模型的细粒度推理能力提出更高要求，传统的单模态方法或简单的视觉匹配策略难以胜任。已有研究尝试引入视觉语言大模型（VLM）辅助理解图像内容<sup>[3]</sup>，但在结构建模与排序推理方面仍缺乏系统性的框架设计。

为了解决上述问题，本文提出一种基于中英文漫画理解与推理能力蒸馏机制的四格漫画顺序识别方法，旨在提升模型对多语言、多模态漫画数据的综合理解与推理能力。该方法主要包括以下三个关键策略：

首先，构建中英文双语漫画数据集，用已有的覆盖多种类型与题材的四格漫画数据集，通过引入语言多样性与文化情境差异，增强模型对不同叙事风格的理解能力。结合图像内容和提取的对白信息，训练大语言模型生成高质量的视觉摘要，为后续顺序推理任务提供语义支撑。

其次，引入大模型辅助的视觉描述生成机制。借助如 Qwen-VL<sup>[4]</sup>多模态大模型，对每格漫画图像生成详尽的视觉描述信息，作为对白的补充，从而提升模型对细节变化、人物行为及背景演化的

感知能力。这些结构化的图文摘要为顺序判断提供更完整的上下文信息。

最后，为提升小模型在推理阶段的性能，设计基于大模型的推理能力蒸馏策略。具体做法是，利用大模型对大量漫画样本进行结构化推理，生成高质量的推理过程与排序结果，构建成“推理数据集”。随后，将该数据集用于训练轻量级的小模型，使其在保留模型效率的同时具备接近大模型的推理能力，从而实现高效部署与快速响应。

本文组织结构如下：第1节对相关工作进行综述，包括对文本顺序推理、多模态漫画理解和大模型蒸馏的研究；第2节描述本文所提出的蒸馏大模型推理能力到小模型，再通过结合大模型生成的图像视觉摘要，对四格漫画的顺序进行排序；第3节实验进行说明以及实验结果分析，以证明模型的有效性；最后，第4节对本文进行总结。

## 1 相关工作

### 1.1 文本顺序推理

文本顺序推理（Textual Order Reasoning）是自然语言处理中的基础任务，旨在判断多段文字的合理排列顺序，以重建其语义完整性和逻辑流畅性。该任务广泛应用于自动写作、文本生成和多段文本理解等场景。传统方法依赖于基于规则或语法线索的启发式排序方法，例如利用时间词、指代词等线索判断先后顺序。但这些方法面对多样化语言表达时鲁棒性较弱，泛化能力不足。

近年来，随着大规模语言模型的发展，基于神经网络的顺序推理方法逐渐兴起。如 BERT<sup>[5]</sup>和 RoBERTa<sup>[6]</sup>等预训练语言模型通过掩码语言建模任务学习到一定的上下文信息建模能力，被用于预测句子或段落的先后顺序。Zhang 等人<sup>[7]</sup>提出利用 Next Sentence Prediction (NSP) 任务微调语言模型，以判断两个句子之间的逻辑连接。Brown 等人<sup>[8]</sup>提出的 GPT 系列则通过自回归方式生成文本，天然具备一定顺序建模能力。此外，Chain of Thought 推理方式<sup>[9]</sup>被广泛用于提升多步推理能力，可用于多轮思考与结构化推理的任务。

尽管已有方法在纯文本语料中取得良好效果，但在多模态场景，特别是图文混合的漫画数据中，

文本顺序推理需要结合图像上下文信息。漫画中

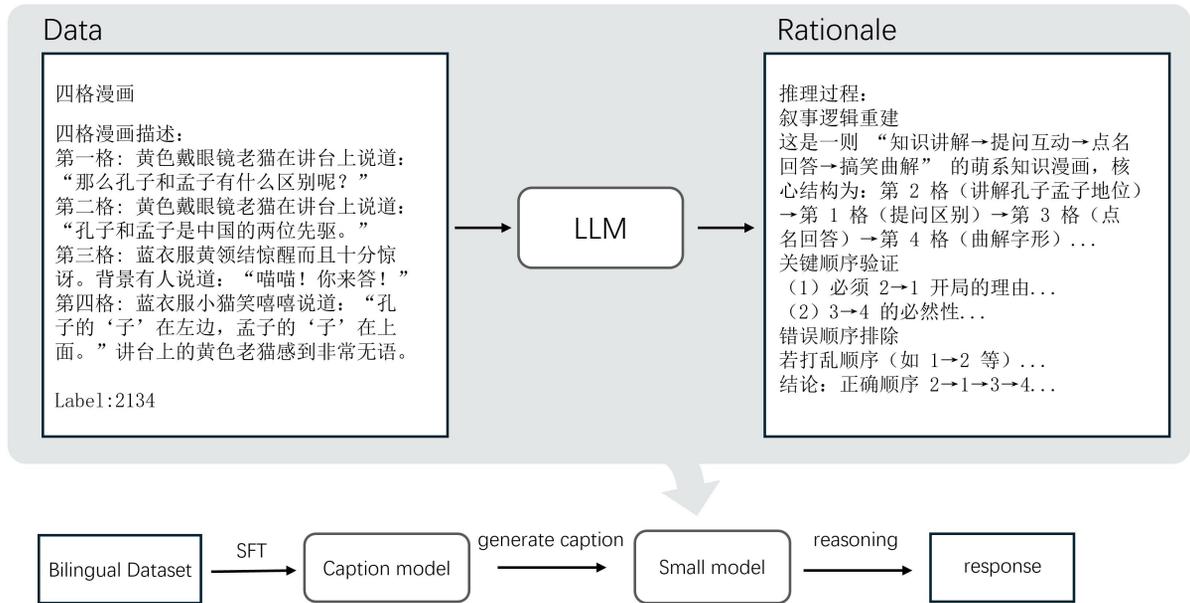


图 1 方法框架

我们首先训练一个生成图文描述模型，在利用思维链 (CoT) 提示从大语言模型 (LLM: Deepseek) 中提取推理过程，然后训练小模型 (Qwen2.5 VL: 7B)，提高小模型在特定任务上的推理能力。

的对白常常依赖于角色动作或画面背景，仅凭语言模型难以判断其合理性。因此，在漫画推理中融合图像与文字的联合建模成为关键挑战。

## 1.2 多模态漫画理解

多模态漫画理解 (Multimodal Comics Understanding) 是指融合图像与文字信息来理解漫画中的故事结构、角色行为和逻辑顺序。漫画作为一种图文紧密结合的叙事媒介，其挑战性往往在于对白文本短小精炼，而大量语义信息则依赖图像传达，如人物动作、表情、时间线等。相比传统自然图像任务，漫画具备高度抽象化、结构规律性强等特征，导致现有通用图文模型效果欠佳。

目前已有一些研究尝试从视觉和语言两个模态共同建模。Iyyer 等人<sup>[10]</sup>提出 COMICS 数据集，并基于 LSTM+CNN 架构进行多模态推理，用于预测下一格图像或对白。Wu 等人<sup>[11]</sup>提出多模态 Transformer 模型，统一建模漫画图像与对白文本之间的对齐关系和时间顺序。同时，近年来多模态大模型 (如 BLIP-2、MiniGPT-4<sup>[12]</sup>、Qwen-VL<sup>[13]</sup>) 展示了出色的图文理解能力，可生成丰富的图像描述，作为对白之外的重要辅助信息。

不过，在实际推理任务中，漫画数据面临以

下问题：1) 对话信息稀疏，部分格子仅有图像内容；2) 图像传达信息微妙，需对人物动作顺序和

场景变化敏感；3) 漫画顺序判断常涉及对多个场景的因果分析或反转判断。针对这些问题，如何提取高质量的视觉摘要，并与对白信息融合，是提升推理准确率的关键。

## 1.3 大模型蒸馏与小模型推理增强

大语言模型 (LLM) 如 ChatGPT、GPT-4 和 Qwen 系列，在语言理解和多模态推理中展现了强大的能力，但其部署成本高、推理延迟大，限制了其在边缘端或实时场景的应用。因此，近年来出现了一类通过蒸馏 (Knowledge Distillation) 手段，将大模型能力“压缩”到轻量化小模型的研究方法<sup>[14]</sup>，以平衡性能与效率。

知识蒸馏最早由 Hinton 等人<sup>[15]</sup>提出，通过最小化学生模型与教师模型输出之间的差异，学习教师模型的推理模式。在大模型推理增强方面，Wang 等人<sup>[16]</sup>提出 Self-Instruct 方法，通过大模型自我生成训练任务，构建多种推理样本来训练小模型。类似地，Wu 等人<sup>[17]</sup>采用 CoT 数据蒸馏，从 GPT-4 生成的多步推理过程学习推理策略，从而提升小模型的逻辑连贯性。

针对漫画顺序判断问题，本文创新性地采用

大模型生成高质量视觉摘要与推理过程，构建小样本推理数据集；再利用该数据集对轻量级模型进行精调，显著提升其顺序推理能力与样本泛化能力。这种“从大模型学习推理流程”的方法在多模态场景下具有良好的扩展性和可迁移性。

## 2 方法

### 2.1 任务定义

对于四格漫画的逻辑理解任务，给定一幅四格漫画  $M$ ，表示为  $M = (p_1, p_2, p_3, p_4)$ ，其中  $p_i$  是四格漫画的第  $i$  个画面格。四格漫画的画面格顺序被随机打乱或只交换两个画面格的顺序，目标是准确判断该漫画正确的阅读顺序  $S$ 。该任务是一个问答任务，其中问题为  $Q = \{M\}$ ，答案为  $A = \{S\}$ 。四格漫画的逻辑理解任务只评估预测答案  $A$ ，而不关心答案是如何推导出来的。

为了使结果具有可解释性，我们进一步将四格漫画逻辑理解任务扩展为可解释的四格漫画逻辑推理任务，其中  $Q$  为上述相同的格式的问题  $Q$ ，该模型需要输出预测的答案  $A$  以及推理过程  $R$ 。可解释的四格漫画的逻辑理解可以增强预测的可信度，并有利于其他与四格漫画相关的任务。方法的框架图如图 1 所示。

### 2.2 四格漫画理解生成

在进行四格漫画的排序任务前，首先需要增强小模型对四格漫画的理解能力。由于任务包含中文和英文的漫画，先训练中文再训练英文会出现对中文的灾难性遗忘，于是我们构建了一个中英文双语数据集，数据分别为中文和英文。该数据集的中文数据由于数据量不足，为了增加中文漫画的数据量，其包含三种类型：正确顺序的漫画及对应的四个画面格的描述、打乱顺序后的漫画及相应的四个画面格的描述和正确顺序的漫画及其中一个画面格的描述，至于第三种类型我们是希望让模型即能学习到漫画的整体特征也能学习到局部特征。至于英文数据，由于数据量足够，我们并没有添加第三种类型的数据。其次，就算使用上述方法增加中文数据量，英文的数据也是中文的三到四倍，为了避免压倒性的英文训练信号，我们将中英文的训练数据比例设置为 1:2，中文数据约有 500 条，英文数据约 1000 条。我们的基础模型是 Qwen 2.5 VL:7B，经过监督微

调后，取得了一定的效果。

### 2.3 四格漫画逻辑推理生成

你是一个逻辑严密的漫画分析专家。下面是一则四格漫画的每格描述，以及这四格的正确阅读顺序。请你根据这个顺序，进行详细的逻辑推理，请按照以下步骤进行推理。

步骤：

1. 叙事逻辑梳理：提炼整个漫画的主线，明确每一格的作用，比如：提出疑问、解释原因、制造冲突、搞笑反转等。
2. 顺序验证关键：选取关键的两个格子对，说明它们的逻辑关系是否合理。例如：格 2→格 1：描述两者间的逻辑关系，如“提出问题→回应解释”；
3. 错误顺序排除：说明如果颠倒这两个格子，会造成哪些逻辑问题或理解障碍。比如“回应缺乏铺垫”、“反转提前导致无趣”等。
4. 结论：总结为什么当前顺序是最符合叙事逻辑的，并输出阅读顺序。

图 2 思维链 (CoT) 提示

我们的任务是对四格漫画进行阅读排序，需要我们去理解四格漫画起承转合的排列逻辑。受思维链 (CoT) 方法的启发，我们设计了一个简单的提示如图 2 所示，指导模型按照我们提出的思维链逐步推理，但对于较小的模型 (Qwen2.5-VL-7B)，由于其有限的参数尺度，直接应用 CoT 提示的效果并不明显。因此，我们考虑通过监督训练来提高小模型的推理能力。

最近的研究表明，蒸馏大语言模型 (LLM) 的 CoT 推理能力到小模型，有助于增强小模型在特定任务上的推理能力。我们给大模型思维链 (CoT) 提示以及正确的阅读顺序，让大模型反向生成符合思维链的推理过程，以此自动的构建一个推理数据集，值得一提的是，我们只使用中文的漫画数据构建了一个中文的推理数据集，并没有按相同的方法构建英文数据集，我们发现英文漫画在只经过中文推理数据集训练后的模型上也有不错的推理效果，这可能与中英文四格漫画具有相同的内在逻辑有关。

## 3 实验设计与结果分析

### 3.1 数据集

为验证本文提出的基于大模型推理能力蒸馏的小模型四格漫画排序方法的有效性，本文选用赛事方给出的中英文双语漫画数据集，用于四格漫画的顺序识别任务。该数据集根据语言类型分为中文漫画子集与英文漫画子集，分别用于研究模型在不同语言环境下对白逻辑与画面内容的综合推理能力。

整个数据集包含大量四格漫画样本，每组样本由 4 幅图像构成，按阅读顺序编号为 0、1、2、3。为保证任务的科学性与公平性，数据集按照 8:1:1 比例划分为训练集、验证集与测试集，如表 1 所示：

表 1 数据集统计信息

| Dataset | Train | Val | Test | Total |
|---------|-------|-----|------|-------|
| Zh      | 480   | 75  | 63   | 618   |
| En      | 2052  | 128 | 128  | 2308  |

训练集：包含图像、对白文本、视觉摘要以及真实顺序标签，用于模型训练；

验证集和测试集：仅提供图像与对白信息，测试阶段不提供顺序标签，模拟真实推理场景。

在测试集构造中，为进一步分析模型在不同干扰强度下的鲁棒性，分别对中英文样本设置不同的打乱策略：

英文漫画测试样本：对原始顺序中的两个相邻宫格进行随机交换；

中文漫画测试样本：若设置为 `random=false`，则随机交换两个相邻的宫格；若设置为 `random=true`，则对 4 个宫格完全随机打乱顺序。

### 3.2 模型训练与推理流程

本文提出一种基于大模型推理能力蒸馏的顺序识别框架，主要包括以下两个阶段：

#### 3.2.1 视觉推理样本生成

首先，使用具备强视觉推理能力的预训练多模态大模型（Qwen2.5-VL-7B）对训练集中的图像和对白文本进行联合推理，自动生成多步 CoT（Chain of Thought）格式的推理路径与视觉摘要，从而形成“图像 + 白描 + 顺序标签”的伪监督数据，构建推理训练集。

#### 3.2.2 推理能力蒸馏至小模型

利用上述大模型生成的推理样本，通过监督训练将其推理路径迁移至轻量化语言模型。该过程本质为多模态到语言模型的推理能力蒸馏（reasoning distillation），使小模型在仅依赖对白文本与图像视觉摘要的条件下，具备漫画顺序识别的能力。

训练完成后，在测试阶段，输入仅包含图像内容与对白信息，无需显式顺序标签，模型即可基于生成的视觉摘要与对白之间的语义连贯性判断其逻辑顺序。

### 3.3 实验结果分析

为评估所提方法的实际效果，分别在中文与英文漫画测试集中进行实验，测试集中每条样本的顺序是否合理需由模型独立判断。本文采用准确率（Accuracy）作为评价指标，定义为预测出的顺序与真实顺序完全一致的样本所占比例。

实验结果如表 2 所示：

表 2 漫画顺序识别准确率

| Dataset | Acc    |
|---------|--------|
| Zh      | 0.4127 |
| En      | 0.3984 |

从表 2 的实验结果来看，我们在中英文上都取得了一个不错的效果，这表明我们的模型学习到了一定的推理能力。由于我们的模型并没有在英语推理数据上训练过，但对英文的四格漫画逻辑理解也有不错的效果，这说明不同语言的四格漫画顺序推理过程的结构和逻辑链是相似的。

## 4 总结

在本文中，我们分两阶段去实现四格漫画的逻辑理解任务。我们首先使用双语数据集训练了一个模型，该模型主要进行图文描述。然后我们蒸馏大语言模型（LLM）的 CoT 推理能力到小模型，提高小模型在该任务上的推理能力，使得出的结果具有可解释性。尽管取得了一些不错的效果，但还是存在一些问题。对于图文描述模型来说，存在描述不准确的问题，特别是会将角色与对白混淆，以及对于不清晰的图片无法正确提取图片中的文字，这会严重影响后续模型的推理效果。对于推理模型来说，存在一些幻觉问题，以及将一些简单的逻辑复杂化，导致推理错误的问题。在未来的工作中，将进一步的解决上述问题，比如对漫画中的角色，对话框，气泡尾部等进行画框标注，训练模型去生成角色与角色，角色与对话框等之间的关系，从而提升模型对漫画描述的质量，进而提升后续的推理效果。对于推理模型的问题，我们可以收集尽可能多样的逻辑链去训练模型，以及可以对推理错误的的数据生成一个反馈，让模型在错误反馈中学习，以生成更高质

量的推理结果。

## 参考文献

- [1] Kiros R, Salakhutdinov R, Zemel R. Unifying visual-semantic embeddings with multimodal neural language models[C]//*Advances in NIPS*. 2014.
- [2] Lee K H, Chen X, Hua G, et al. Stacked cross attention for image-text matching[C]//*ECCV*. 2018: 201-216.
- [3] Li J, Yang C, Zhang C, et al. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation[J]. *arXiv preprint arXiv:2201.12086*, 2022.
- [4] Bai Y, Liu Z, Li C, et al. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond[J]. *arXiv preprint arXiv:2308.12966*, 2023.
- [5] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//*NAACL-HLT*. 2019.
- [6] Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J]. *arXiv preprint arXiv:1907.11692*, 2019.
- [7] Zhang Y, Sun S, Galley M, et al. DialoFlow: Improving Dialogue Modeling by Teaching Flow of Thought[J]. *arXiv preprint arXiv:2010.06922*, 2020.
- [8] Brown T B, Mann B, Ryder N, et al. Language Models are Few-Shot Learners[C]//*NeurIPS*. 2020.
- [9] Wei J, Wang X, Schuurmans D, et al. Chain of Thought Prompting Elicits Reasoning in Large Language Models[J]. *arXiv preprint arXiv:2201.11903*, 2022.
- [10] Iyyer M, Manjunatha V, Guha A, et al. The Amazing Mysteries of the Gutter: Drawing Inferences Between Panels in Comic Book Narratives[C]//*CVPR*. 2017.
- [11] Wu Y, Liu Y, Liu C, et al. Bridging the Gap Between Comics and Real Images: A Multi-Modal Transformer for Story Understanding[J]. *arXiv preprint arXiv:2203.13316*, 2022.
- [12] Zhu D, Zhang Y, Wang Y, et al. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models[J]. *arXiv preprint arXiv:2304.10592*, 2023.
- [13] Qwen Team. Qwen-VL: A Strong Vision-Language Model with Open-source Training and Serving Code[J]. <https://huggingface.co/Qwen/Qwen-VL>, 2023.
- [14] Jiao X, Yin Y, Shang L, et al. TinyBERT: Distilling BERT for Natural Language Understanding[J]. *arXiv preprint arXiv:1909.10351*, 2019.
- [15] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network[J]. *arXiv preprint arXiv:1503.02531*, 2015.
- [16] Wang Y, Kordi Y, Wei J, et al. Self-Instruct: Aligning Language Models with Self-Generated Instructions[J]. *arXiv preprint arXiv:2212.10560*, 2022.
- [17] Wu L, Xu Y, Zhang D, et al. CoT-Distill: Distilling Reasoning Capabilities of Chain-of-Thought Reasoning to Small Language Models[J]. *arXiv preprint arXiv:2305.02301*, 2023.

# CCAC2025-task3 技术报告：四格漫画理解能力评测

李成飞<sup>1</sup> 王春雨<sup>2</sup> 张文雅<sup>1</sup> 张云杰<sup>1</sup> 李翰林<sup>1</sup> 陈文锟<sup>1</sup> 刘德坤<sup>1</sup>

1. 齐鲁师范学院 人工智能教育研究院

2. 齐鲁师范学院 地理与旅游学院

**摘要：** 该文介绍了基于 Qwen2.5-VL-7B 多模态大语言模型的四格漫画理解评测方案，针对逻辑推理和上下文理解任务设计了专门的训练策略和提示词。该方案基于 LlamaFactory 框架，结合低秩适配 (LoRA) 微调与多模态融合机制，使用任务特定数据集进行针对性微调。在 CCAC2025 评测中，该方法在任务一（逻辑推理）中英文数据 Macro-F1 分数分别为 0.687 和 0.476，任务二（上下文推理）中英文数据 ROUGE-L 分数分别为 0.516 和 0.618。结果表明，多模态大语言模型在复杂视觉-文本理解任务中表现出良好的适应能力。

**关键词：** 多模态大模型；提示工程；漫画理解

## Technical Report on the CCAC2025 Task 3: Four-Panel Comics Comprehension Assessment

Chengfei Li<sup>1</sup>, Chunyu Wang<sup>2</sup>, Wenya Zhang<sup>1</sup>, Yunjie Zhang<sup>1</sup>, Hanlin Li<sup>1</sup>, Wenkun Chen<sup>1</sup>, Dekun Liu<sup>1</sup>

1. Artificial Intelligence Research Institute on Education, Qilu Normal University

2. School of Geography and Tourism, Qilu Normal University

**Abstract :** This paper introduces an evaluation framework for four-frame comic understanding based on the Qwen2.5-VL-7B multimodal large language model. Tailored training strategies and prompt templates are designed specifically for logical reasoning and context comprehension tasks. The framework, built upon the LlamaFactory framework, integrates low-rank adaptation (LoRA) fine-tuning with multimodal fusion mechanisms, utilizing task-specific datasets for targeted fine-tuning. In the CCAC2025 evaluation, the approach achieved Macro-F1 scores of 0.687 and 0.476 for English and Chinese datasets, respectively, in Task 1 (logical reasoning), and ROUGE-L scores of 0.516 and 0.618 for English and Chinese datasets, respectively, in Task 2 (contextual reasoning). The results demonstrate the robust adaptability of multimodal large language models in complex visual-text understanding tasks.

**Key words:** Multimodal Large Language Model; Prompt Engineering; Comic Understanding

## 0 引言

随着人工智能技术的发展，单一模态的信息处理已经难以满足复杂现实场景的多维需求，多模态学习逐渐成为人工智能研究的重要方向。四

格漫画作为典型的多模态叙事形式，面临复杂的理解挑战。在叙事结构方面，其通常遵循“起承转结”的叙事逻辑，不同画格之间存在着清晰的时序与因果关系，要求模型具备对事件演化和逻辑链条的把握能力。四格漫画的理解涉及视觉（角色表情、动作、背景）、文本（对白、旁白、

拟声词)及空间(图像排布、阅读顺序)等多层次信息的融合。此外,四格漫画常嵌入文化语境中的隐喻、幽默或讽刺内容,模型需要具备跨文化语义推理能力,这增加了跨语言和多模态语义理解的难度。

在此背景下,CCAC2025 四格漫画理解评测设立了两个核心任务,旨在从不同认知维度评估多模态模型对四格漫画的理解能力。两个任务分别侧重逻辑推理与上下文生成,涵盖了四格漫画叙事理解的关键能力要求[1]。

任务一旨在评估模型对漫画叙事顺序的识别能力。在该任务中,输入为四格被打乱顺序的图像集合  $P = \{p_1, p_2, p_3, p_4\}$ , 模型需输出正确的阅读顺序  $O = [o_1, o_2, o_3, o_4]$ , 其中  $o_i \in \{0,1,2,3\}$ 。四格漫画遵循“起承转结”的结构,各格子之间具有明确的时序和因果关系。该任务要求模型综合分析图像中的视觉要素与文本信息,识别画面间的逻辑衔接,最终完成排序预测。模型性能以 Macro-F1 分数为评估指标[2]。

任务二则考查模型的上下文推理与文本生成能力。输入为一组包含三个已知格子和一个缺失格子的漫画  $M = \{p_1, p_2, [MASK], p_4\}$ , 模型需基于上下文信息生成缺失格子的文本描述  $D = \{d_1, d_2, \dots, d_n\}$ 。该任务模拟人类在漫画阅读中根据上下文补全情节的过程,评估指标包括 ROUGE-L 和 BLEU 分数,用于衡量生成内容的准确性与连贯性。模型需具备跨模态理解能力,在捕捉视觉语义的基础上生成符合情节发展的合理文本。

该文基于 Qwen2.5-VL-7B 多模态大语言模型,分别针对两个任务设计了定制化的训练方案。训练过程中结合任务特定数据构建与微调策略,采用低秩适配(Low-Rank Adaptation)方法提升模型性能。在提示设计方面,为任务二的中英文数据分别构建了不同的提示模板,进一步增强了模型的语言理解与生成能力。在 CCAC2025 评测中,本方法在两个任务上均获得了较为理想的结果,验证了所提出方案在多模态漫画理解任务中的可行性与实用性。

## 1 方法与实现

该文基于 Qwen2.5-VL-7B 多模态大语言模型,在 LLaMAFactory 框架下使用官方提供的四格漫画任务数据进行微调,并将结果提交至 CCAC2025 评测系统。所参与评测任务包括:一是判断打乱顺序的四格漫画的正确阅读顺序,二是根据前后文生成缺失格子的合理描述。所使用的模型包含视觉编码器(ViT)、语言模型主干(Qwen2.5-7B)、多模态融合层(Cross-Attention)和任务解码器,模型结构本身未作修改,仅通过 LoRA 方式进行微调适配,以完成任务优化,技术方案如图 1 所示。

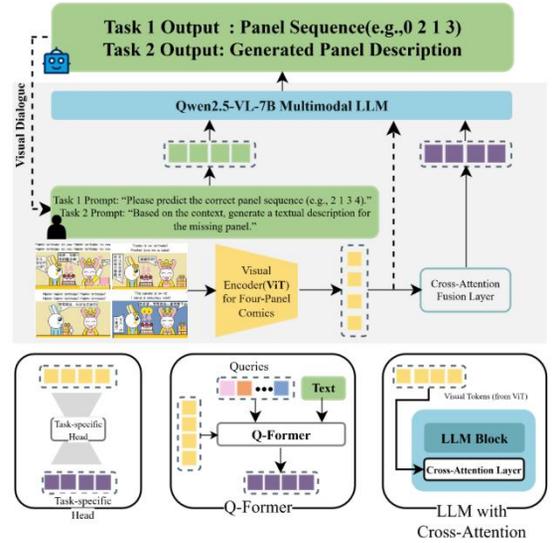


图 1 基于 Qwen2.5-VL-7B 多模态模型的四格漫画理解技术流程

在模型微调过程中,选用了 LoRA (Low-Rank Adaptation) 策略以实现参数高效的适配优化。该策略在保持预训练模型结构与参数冻结的基础上,引入低秩矩阵进行权重调整,在节省显存开销的同时实现针对性任务迁移[3]。所使用的 Qwen2.5-VL-7B 模型包含约 70 亿参数,由视觉编码器、语言建模主干与多模态融合层组成,具备较强的跨模态理解与表示能力(图 2)。

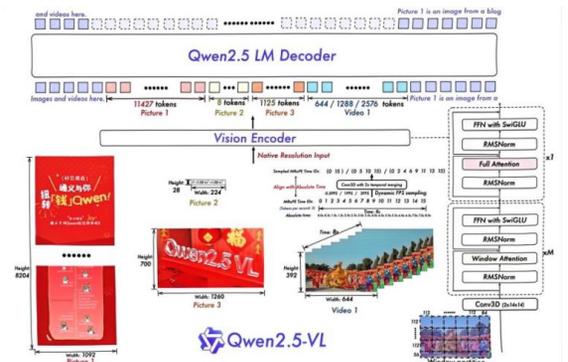


图 2 qwen2.5-VL 模型结构图



关键配置，以保障训练稳定性，同时兼顾效率与性能表现。训练中采用 LoRA 技术微调参数，rank 设为 64，减少资源消耗并提高模型适配性。输入图像分辨率设置为  $448 \times 448$ ，文本序列长度上限设定为 2048，以适应多模态长序列输入场景。

表 1 基础训练参数设置

| 参数名称                        | 数值     |
|-----------------------------|--------|
| learning_rate               | 1.0e-4 |
| per_device_train_batch_size | 4      |
| gradient_accumulation_steps | 4      |
| num_train_epochs            | 3.0    |
| cutoff_len                  | 4096   |

在评估阶段，依据 CCAC2025 组委会发布的官方评分标准开展性能测量。任务一采用 Macro-F1 作为评估指标，旨在衡量模型对顺序预测任务中多类别预测结果的整体性能[2]。任务二则以 ROUGE-L 与 BLEU 作为主评估指标，从语言表达的准确性与覆盖性两个维度评估生成文本与标准答案之间的接近程度[5]。

评测结果如表 2 所示，整体显示训练后的模型在两个任务中均展现出较强的泛化能力。特别是在中文语境下的排序预测任务中，取得 Macro-F1 0.687 的成绩，表明模型在处理时序关系和逻辑推理方面表现稳健。相比之下，英文语料在上下文生成任务中取得更优表现，可能与其语言结构灵活、表达多样性有关。

表 2 各任务评估结果

| 任务    | 语言 | 指标       | 分数    |
|-------|----|----------|-------|
| Task1 | 中文 | Macro-F1 | 0.687 |
| Task1 | 英文 | Macro-F1 | 0.476 |
| Task2 | 中文 | ROUGE-L  | 0.516 |
| Task2 | 英文 | ROUGE-L  | 0.618 |

## 4 结果分析与讨论

通过进一步分析任务间差异，我们发现中文漫画常采用明确的“起承转结”结构，且图文配合直观，有利于模型推理叙事逻辑。而英文漫画表达相对灵活，常见省略与暗示结构，使得模型在逻辑推理中表现偏弱，但也为文本生成提供了丰富的表达素材。中英文表现差异从侧面反映出

跨文化语境下漫画理解的复杂性，也揭示了模型在多语言场景下的适应性问题[6]。

影响模型性能的主要因素包括提示模板设计、微调策略和数据质量。提示工程有效提升了模型对任务意图的理解，LoRA 方法保障了模型微调效率与性能稳定性，标准化高质量数据则为训练提供了清晰监督信号。这些策略共同保障了模型在评测任务中取得理想成绩。

尽管取得一定成果，该方案仍存在局限。首先，所使用的基础模型未专门针对漫画场景结构设计，难以显式建模格子间的时序与空间关系。其次，对于嵌入文化背景的隐喻与幽默，模型的跨文化理解能力仍显不足。此外，数据集规模受限，模型在面对复杂、少见场景时表现不够稳健。

## 5 展望与结论

针对上述问题，未来的研究将从以下几个方向展开：一是设计适配四格漫画结构的视觉编码模块，如引入显式建模格子关系的注意力机制或专门的位置编码方案；二是尝试多任务协同学习，使模型在推理与生成之间实现互补学习；三是引入合成数据与增强样本，提升模型对复杂漫画场景的泛化能力；四是构建带有文化背景标注的训练语料，增强模型对隐含语义与文化要素的识别能力。

综上所述，该文提出的基于 Qwen2.5-VL-7B 的四格漫画理解方案，结合 LoRA 轻量微调与任务特定提示工程，在两个评测任务中取得了良好表现，验证了在多模态语境下优化任务的可行性与实用性。未来我们将继续探索多模态模型在文化类内容理解中的能力边界，并推动其在教育、娱乐、内容创作等领域的广泛应用[1-3][7]。

## 参考文献

- [1] 周浩榕. “多模态大语言模型综述.” Computer Science Undergraduate Conference 2025@ XJTU. 2025.
- [2] Ikuta, Hikaru, Leslie Wohler, and Kiyoharu Aizawa. “Mangaub: A manga understanding benchmark for large multimodal models.” IEEE MultiMedia (2025).
- [3] Hu, Edward J., et al. “Lora: Low-rank adaptation

- of large language models." ICLR 1.2 (2022): 3.
- [4] Boonstra, Lee. "Prompt engineering." (2024).
- [5] Cao, Stanley, and Sonny Young. "Predicting Winning Captions for Weekly New Yorker Comics." arXiv preprint arXiv:2407.18949 (2024).
- [6] Mikkonen, Kai. The narratology of comic art. Taylor & Francis, 2017.

# 融合幽默-讽刺语义的多模态多任务情感识别方法

冷基栋<sup>1</sup> 闫强<sup>1</sup> 易兰丽<sup>1</sup> 王斯然<sup>1</sup> 谭嘉辰<sup>2</sup> 黄林迪<sup>1</sup>

(1.北京邮电大学 经济管理学院, 北京市 100876; 2.北京邮电大学 计算机学院, 北京市 100876)

**摘要:** 讽刺识别作为情感计算中的前沿问题, 面临幽默语义交叉、语言表达隐喻性强以及多模态线索融合难度大的多重挑战。本文聚焦幽默与讽刺之间的复杂语义联动关系, 提出一种融合结构化语义特征的中文多模态多任务情感识别方法。所构建模型以预训练语言模型 MacBERT 为主干, 联合建模“是否讽刺”与“讽刺类型”两个子任务, 在输入端引入上下文拼接与 18 维结构化特征, 包括说话人讽刺倾向、幽默类型、情感强度估计及上下文发言一致性等, 以增强模型的语义理解能力。同时, 结合 OpenSMILE 与 MediaPipe 提取的音频与视频模态特征, 有效提升模型对非语言情绪信号的感知水平。训练过程中引入加权 Focal Loss 与学习率预热机制, 以缓解类别不平衡带来的学习偏差。此外, 本文采用大语言模型自动提取并替换讽刺对象, 开展稳健性验证, 进一步增强模型对讽刺目标变动的适应能力。实验结果显示, 该方法在多模态讽刺识别测评任务中取得最优成绩, 最终 Score 达到 0.5989, 排名第一, 充分验证了幽默-讽刺语义联动建模在多模态情感识别中的有效性与鲁棒性。

**关键词:** 多模态情感识别; 讽刺识别; 多任务学习

中图分类号: TP391

文献标识码: A

## A Multimodal Multi-Task Emotion Recognition Method Integrating Humorous and Sarcastic Semantics

Jidong Leng<sup>1</sup>, Qiang Yan<sup>1</sup>, Lanli Yi<sup>1</sup>, Siran Wang<sup>1</sup>, Jiachen Tan<sup>2</sup>, and Lindi Huang<sup>1</sup>

(1.School of Economics and Management, Beijing University of Posts and Telecommunications

No. 10 Xitucheng Road, Haidian District, Beijing 100876, China;

2.School of Computer Science, Beijing University of Posts and Telecommunications

No. 10 Xitucheng Road, Haidian District, Beijing 100876, China)

**Abstract :** Sarcasm recognition poses several challenges in affective computing due to overlapping semantics with humor, implicit language expressions, and difficulties in multi-modal cue integration. This study explored the semantic interaction between humor and sarcasm and proposed a Chinese multi-modal multi-task affective recognition model enhanced with structured semantic features. The model was based on the MacBERT pre-trained language encoder and jointly addressed sarcasm detection and sarcasm type classification. Contextual utterances and eighteen structured features were incorporated to improve semantic understanding, including speaker sarcasm tendencies, humor categories, emotional intensity, and consistency across dialogue turns. Acoustic and visual signals were extracted using OpenSMILE and MediaPipe to capture non-verbal emotional cues. Weighted Focal Loss and a learning rate warm-up strategy were used to address class imbalance.

收稿日期: 20XX-XX-XX; 定稿日期: 20XX-XX-XX

基金项目: 北京邮电大学研究生创新创业重点项目 (2025-YC-T041)

To test robustness, sarcasm targets were replaced using a large language model. Experimental results showed that the proposed method achieved the highest score of 0.5989 in the benchmark task. These findings confirm that modeling the semantic linkage between humor and sarcasm improves both recognition accuracy and robustness in multi-modal affective analysis.

**Key words:** Multimodal affective recognition; Sarcasm recognition; Multi-task learning

## 0 引言

讽刺识别是自然语言处理中的关键任务之一，因其表达方式的间接性与模糊性，以及语义线索的多模态融合特性，长期以来被视为情感计算的难点。在中文语境中，讽刺语言常融合幽默、夸张、反语等多重修辞策略，使得情感意图更具隐蔽性和解读复杂性。此外，讽刺往往超越文本本身，依赖语调、表情、语境等非语言信息，进一步加大了建模难度。

现有研究多侧重于文本特征提取与预训练语言模型应用，尽管提升了识别性能，但对“幽默”与“讽刺”之间语义耦合关系的建模仍显不足。二者虽具认知差异，但在真实语料中高度交织，幽默常作为讽刺的表达手段，反之亦然。因此，挖掘二者之间的语义联动机制，对提升多模态情感识别系统的精度与鲁棒性具有重要意义。

本文提出一种融合幽默-讽刺语义结构的多模态多任务情感识别方法，在中文预训练模型 MacBERT 基础上，引入上下文拼接与结构化特征增强，实现“是否讽刺”与“讽刺类型”两个子任务的联合建模。结构化特征覆盖说话人讽刺倾向、幽默类型、情绪强度及上下文一致性等 18 个维度，辅以 OpenSMILE 与 MediaPipe 分别提取的音频情绪与视频面部特征，全面提升模型对讽刺语义的感知能力。

在讽刺识别测评任务上的实验表明，该方法在讽刺判断与类型分类两个子任务上均取得优异性能，验证了幽默-讽刺语义联动建模在多模态情感识别中的有效性与实用价值。

## 1 任务与数据

### 1.1 任务定义

多模态情感识别任务旨在借助多源信息融合技术，理解和判断文本中所蕴含的复杂情绪表达。讽刺作为一种典型的复杂语言现象，常常与幽默相伴而生，其表达具有含蓄、反转与间离等语义特征。本次多模态讽刺识别任务包含讽刺识别和讽刺类型识别两个子任务。

#### 1.1.1 多模态讽刺识别

给定文本、音频与视频等多模态输入，判断该语句是否包含讽刺意味，输出为二分类标签，其中“1”表示讽刺，“0”表示非讽刺。

#### 1.1.2 多模态讽刺类型识别

针对被识别为讽刺的样本，进一步划分其所属的讽刺类型，分为 7 类，其中 -1 表示非讽刺，0 表示 Sarcasm (嘲笑)，1 表示 Irony (反语)，2 表示 Satire (讽刺文学)，3 表示 Understatement (轻描淡写)，4 表示 Overstatement (夸大)，5 表示 Rhetorical Question (反问)。

### 1.2 测评数据

本次多模态讽刺识别任务所使用的数据集来源于中文脱口秀节目《吐槽大会》，涵盖真实语境中自然发生的幽默与讽刺表达。数据以“句子”为基本单元组织，采用统一编号体系（如“2\_1\_3”表示第 2 季第 1 期第 3 条语句），具有较高的时间对齐精度与语义一致性。每条样本同时提供文本、音频和视频三种模态输入，支持多模态特征的融合建模与情感关联建构。

每条数据样本均附带完整的结构化标注信息，涵盖文本语义与多维情绪标签，具体包括：文本内容 (sentence)、说话人姓名 (speaker)、视频片段起止时间戳 (start / end)，以及三类情感属性标签——是否幽默 (isHumor, 1 表示幽默，0 表示非幽默)、幽默类型 (humorType, 取值范围为 0-7, -1 表示非幽默)、说话人情绪强度 (speakerEmotion, 取值范围为 -2 至 2)、是否讽刺 (isSarcasm, 1 表示讽刺，0 表示非讽刺)，

以及讽刺类型 (sarcasmType, 0-5 表示讽刺类别, -1 表示非讽刺)。此外, 数据集中按期提供原始视频与音频文件(每期对应一个 MP4 文件与一个 WAV 文件), 与句级标注数据严格对应, 为多模态特征提取与语义建模提供了坚实的数据支撑。

### 1.3 评估指标

本任务采用宏平均 F1 值作为主要评估指标, 用于衡量模型在各类样本上的整体表现。最终排名结果由两个子任务的宏 F1 值加权平均计算得出, 其中讽刺识别任务和讽刺类型识别任务各占 50% 权重, 以综合反映模型在讽刺检测与细粒度分类方面的能力。

## 2 模型设计

### 2.1 总体架构

本研究构建了一个融合语言理解与结构化知识的多模态多任务情感识别模型, 整体架构如下: 以中文预训练语言模型 MacBERT 为文本语义编码主干, 同时引入音频、视频及多维结构化特征作为辅助输入, 以增强模型对幽默与讽刺之间复杂语义关系的建模能力。在输出层设计上, 模型采用共享表示空间, 并并联两个任务头, 分别用于完成“是否为讽刺语句”的二分类任务 (isSarcasm) 与“讽刺类型识别”的多分类任务 (sarcasmType)。

模型的输入包括: 拼接上下文的文本序列 (前一句、当前句、后一句); 与文本对齐的 18 维结构化特征; 音频与视频模态特征。

模型主体采用 MacBERT 的 [CLS] 表示作为语义向量, 随后与结构化特征进行拼接, 并通过共享的前馈神经网络获取融合表示。该共享表示再分别输入至两个独立的分类器中, 分别输出两个任务的预测结果。

为适应测评任务的评估标准, 我们在训练阶段设计了加权多任务损失函数, 确保两个子任务在建模过程中具备协同优化能力。整体结构兼顾语言理解、语义推理与类别辨析, 适用于讽刺语言中多模态线索交织、语义层次复杂的识别场景。

### 2.2 多模态与结构化特征建模

讽刺语义的识别不仅依赖于文本本身的语义信息, 更有赖于多源线索的综合建模。为提升模型对复杂语义现象的理解能力, 本文在基础文本特征之外, 融合了结构化特征与音视频模态信息, 构建更具语义感知能力的多模态输入表示。

在结构化特征方面, 本文设计了 18 维辅助特征, 覆盖五个语义维度: (1) 说话人层面的讽刺表达倾向, 包括其历史讽刺发生率及六类讽刺类型的分布概率; (2) 语气与情绪线索, 通过将说话人情绪等级映射为情感强度值, 并结合幽默类型对应的幽默发生概率; (3) 讽刺目标相关信息, 是否存在明确的讽刺对象, 并提取其在训练集中对应的讽刺类型分布; (4) 上下文说话人一致性, 包括当前语句与前后语句说话人是否一致的两个布尔特征; (5) 数据稀疏情况下, 采用全局均值替代未知说话人或目标的特征值, 确保结构化特征的稳定性与泛化性。

在多模态特征建模方面, 本文引入音频与视频通道中可感知的情绪与语用信号。音频部分基于 OpenSMILE 工具提取 eGeMAPSv02 函数级特征, 选用 alphaRatio\_sma3\_amean 表征语速与语气强度, 通过区间划分生成五级情感强度特征并加入模型训练。视频部分利用 MediaPipe 提取与讽刺语义相关的人脸语用信号, 包括嘴部张合比 (MAR)、眼部闭合比 (EAR)、以及头部的 pitch、yaw、roll 姿态角, 共计 6 维视觉特征。所有音视频特征均与文本时间戳对齐, 并通过时间聚合与文本级样本进行融合, 以保证模态同步与表达一致性。通过引入上述结构化与多模态特征, 本文所构建模型在讽刺与幽默的联动建模中具备更强的语义辨析能力, 显著增强了模型对语用线索与上下文关联的敏感性, 为多模态情感识别提供了更加丰富的语义支撑。

### 2.3 多任务学习结

针对讽刺识别任务中存在的“是否讽刺”与“讽刺类型”双重语义需求, 本文采用多任务学习架构, 在共享编码层的基础上, 分别建立两个任务头, 实现讽刺检测与类别识别的联合建模。该结构不仅提升了模型对讽刺语义的泛化能力, 也有助于缓解讽刺类型样本不均衡带来的训练偏

差问题。

具体而言，模型主干采用 MacBERT 作为文本编码器，将当前语句与上下文拼接后输入，获得语义表征的 [CLS] 向量。该向量与结构化与模态特征拼接后，通过一个共享的全连接表示层提取联合语义特征。共享语义层输出分别传入两个独立的分类头，分别对应二分类任务和七分类任务（含非讽刺一类）。两个任务共享底层语言与结构知识，同时保留各自的决策边界。

在损失函数设计上，考虑到类别分布的不均衡问题，本文为两个任务分别引入了 Focal Loss，其中二分类任务设定较低的聚焦参数 ( $\gamma=1.0$ ) 以稳定训练，而多分类任务设定较高的  $\gamma$  值 ( $\gamma=2.0$ ) 以聚焦少数类别的学习。最终损失为两个任务损失的加权求和，其中类别识别任务权重更高，以提升模型对复杂讽刺类别的辨别能力。联合训练中，模型以两个任务的宏平均 F1 (macro F1) 得分加权平均作为评估指标，并依据该综合评分选取最佳模型参数。

该多任务架构有效地捕捉了幽默与讽刺之间的语义联动关系，实现了对讽刺存在性与类型细粒度的同步建模，为多模态讽刺识别任务提供了一种具备语义融合与任务协同能力的建模范式。

## 3 实验设置与环境

### 3.1 实验设置

在训练过程中，二分类任务使用  $\gamma=1$  的 Focal Loss，多分类任务使用  $\gamma=2$  的 Focal Loss，并以 0.4 和 0.6 的权重加权组合，作为整体优化目标函数。训练集和验证集按 8:2 比例划分，最大序列长度设置为 128。优化器采用 AdamW，初始学习率为  $3e-5$ ，引入 10% 的 warmup 步数，并采用线性学习率衰减策略。验证集上以两个子任务的宏平均 F1 分数加权平均作为最终性能指标，用于保存最优模型。

### 3.2 实验设置

实验在具备 NVIDIA RTX 3090 GPU 的计算节点上运行，操作系统为 Ubuntu 20.04，Python 版本为 3.9，主要依赖的开源库包括 PyTorch

1.13、Transformers 4.30、scikit-learn、pandas 及 tqdm。音频与视频特征提取分别使用 opensmile（配置为 eGeMAPSv02）和 MediaPipe 模块。全部实验均在单卡 GPU 环境下完成，确保模型训练与评估的一致性与可复现性。

## 4 实验结果与分析

### 4.1 最终性能表现

本节从训练集与测试集两个层面，对所提出的多模态多任务情感识别模型的整体性能进行评估。模型包含两个子任务：T1 表示是否讽刺的二分类任务，T2 表示讽刺类型的多分类任务。评估指标包括准确率 (Accuracy)、精确率 (Precision)、召回率 (Recall) 与宏平均 F1 分数 (F1-macro)，分别对两个子任务进行效果衡量。

在训练集上，模型在 T1 中取得了 88.40% 的准确率与 82.21% 的 F1 分数，显示其对文本与多模态输入中的讽刺语义具有良好的识别能力。T2 虽然达到了 95.26% 的准确率，但由于类别分布不均，精确率和召回率分别为 34.71% 和 35.42%，F1 分数为 34.99%。这表明模型虽能整体覆盖多种讽刺类型，但在细粒度识别上仍存在偏向主类的风险。

在测试集上，模型整体性能依然保持稳定，T1 和 T2 的 F1 分数分别为 80.87% 和 38.91%，最终综合得分为 59.89%，在讽刺识别评测任务中处于领先水平。相较于训练集，测试集的 F1 分数略有下降，尤其 T2 的多分类性能波动较大，进一步印证讽刺类型识别任务本身的挑战性与复杂性。

在测试集上，模型整体性能依然保持稳定，两个子任务的 F1 分数分别为 0.8087 与 0.3891，最终综合得分为 0.5989，在讽刺识别评测任务中处于较优水平。相较于训练集，测试集的 F1 分数略有下降，尤其在 T2 多分类任务中较为显著，进一步印证讽刺类型识别的复杂性与挑战性。

总体来看，模型在 T1 上表现出较强的稳健性，能够有效判断讽刺意图的存在；而在 T2 上仍受限于类别不均衡与语义模糊性的干扰，后续可引入类别增强机制与上下文建模策略以提升

其泛化能力。

## 4.2 消融实验

为深入评估不同特征模块对模型性能的实际贡献，本文设计了一系列消融实验，系统考察基础文本信息、多模态与结构化特征，以及上下文信息在两个子任务上的影响。实验结果如表 1 所示。

表 1 不同特征组合下模型在训练集与测试集的评估结果

| 特征组合                 | T1-F1 (训练) | T2-F1 (训练) | T1-F1 (测试) | T2-F1 (测试) | Score (测试) |
|----------------------|------------|------------|------------|------------|------------|
| 文本                   | 73.92%     | 26.75%     | 65.96%     | 27.60%     | 46.78%     |
| 文本 + 多模态 + 结构化       | 82.21%     | 28.29%     | 80.87%     | 27.15%     | 54.01%     |
| 文本 + 多模态 + 结构化 + 上下文 | 96.23%     | 36.15%     | 94.02%     | 38.62%     | 66.82%     |

其中，基础模型仅以文本信息作为输入，未融入结构化特征与音视频模态数据。在训练集上，该模型在 T1 任务的 F1 分数为 73.92%，T2 任务为 26.75%；在测试集上，综合得分为 46.78%，整体识别效果较为有限。引入结构化语义特征与音视频模态信息后，模型性能显著提升：训练集 T1 与 T2 任务的 F1 分数分别提升至 82.21% 与 28.29%，测试集综合得分上升至 54.01%，显示出多模态与结构化线索在识别复杂情绪语义中的关键作用。

在此基础上，进一步融合文本上下文信息，构建具备全局语义建模能力的上下文增强模型。该模型在训练集 T1 与 T2 任务中的 F1 分数分别达到 96.23% 与 36.15%，在测试集上的两个任务 F1 分数分别为 80.87% 与 38.62%，综合得分提升至 66.82%。实验结果表明，上下文信息能够有效增强模型对语义一致性与讽刺转折点的感知能力，从而进一步提升整体识别效果。

## 4.3 讽刺对象替换的鲁棒性验证

为进一步评估模型在讽刺目标变动情境下的鲁棒性，本文引入大语言模型自动生成的讽刺对象，替换原始数据集中人工标注的讽刺目标字段，构建讽刺对象增强版本的数据集。该策略旨在缓解原始目标信息存在的不一致与缺失问题，提升模型对细粒度讽刺语义的建模能力。

在训练流程与参数设定保持一致的前提下，使用增强数据重新训练模型。结果显示，T1 任务在训练集上表现稳定，准确率为 88.40%，宏 F1 为 82.21%，说明模型对讽刺语义的感知能力未受扰动影响。T2 任务尽管准确率达到 95.26%，但受类别不均衡影响，宏精确率与召回率分别为 34.71% 和 35.42%，对应的宏 F1 为 34.99%，表明边缘类别识别能力有所提升，但整体仍存在偏主类问题。

在测试集上，T1 与 T2 的宏 F1 分别为 80.87% 和 38.91%，综合得分为 59.89%，与基础模型相比波动较小，验证了该方法在讽刺目标变化情境下的稳健性，同时为提升 T2 任务的泛化能力提供了改进空间。

## 5 结论与展望

本文针对讽刺识别中幽默与讽刺语义耦合性强、表达形式多样及多模态融合复杂等挑战，提出一种融合幽默-讽刺语义的中文多模态多任务情感识别方法。该方法以 MacBERT 为主干，结合结构化语义特征（如幽默类型、讽刺倾向、情绪强度）与音频、视频模态信息，构建统一表示空间，并联合建模讽刺检测与类型识别两个任务。通过上下文拼接、多任务损失设计与 Focal Loss 策略，有效提升了模型对复杂情感语义的表达能力与鲁棒性。

在多模态讽刺识别测评任务中，模型在两个子任务上均取得较优 F1 分数，消融实验验证了结构化特征、多模态融合与上下文建模对性能的显著增益；鲁棒性分析亦表明，即使替换讽刺目标来源，模型仍保持稳定的识别效果，体现出良好的泛化能力。

然而，当前方法仍受讽刺类型分布不均与语义遮蔽影响，部分细粒度类别识别精度仍有待提升。未来可从以下方向进一步拓展：（1）引入基于对话结构的动态上下文建模机制，增强对语义演化过程的捕捉能力；（2）结合指代消解与因果推理技术，提升对讽刺中隐含指向与立场的识别效果；（3）探索大模型引导下的半自动标签机制，优化训练数据质量，拓展模型的适应范围与泛化能力。

第一作者冷基栋（1995年生），博士研究生，主要研究领域为多模态情感分析，人机交互中的情感分析。

E-mail: lengjd@126.com

通讯作者闫强（1971年生），教授，博士生导师，主要研究领域为智能人机交互与个体行为决策、人工智能治理、社交媒体大数据分析。

E-mail: yan@bupt.edu.cn

# CCAC2025 赛道四技术报告： 基于 Qwen2-Audio 的中文多模态讽刺计算技术研究

李成飞<sup>1</sup> 王春雨<sup>2</sup> 张云杰<sup>1</sup> 李翰林<sup>1</sup> 张文雅<sup>1</sup> 陈文锟<sup>1</sup> 刘德坤<sup>1</sup>

1. 齐鲁师范学院 人工智能教育研究院  
2. 齐鲁师范学院 地理与旅游学院

**摘要：** 该文基于 CCAC 2025 中文多模态讽刺计算评测任务，提出了基于 Qwen2-Audio 模型的解决方案。针对多模态讽刺识别和类别识别任务设计了专门的训练策略和数据增强方法。该方案采用 LoRA 微调技术，结合跨任务迁移学习和数据平衡策略，使用《吐槽大会》视频数据集进行针对性微调。在 CCAC2025 评测中，该方法在任务一（多模态讽刺识别）取得了 0.7245 的 macro F1 分数，在任务二（多模态讽刺类别识别）取得了 0.3782 的 macro F1 分数。实验结果表明，跨任务迁移学习能够显著提升讽刺识别性能，简单的数据平衡策略比复杂的音频增强技术更为有效。

**关键词：** 多模态大模型；讽刺计算；Qwen2-Audio；迁移学习；数据增强

## Technical Report on the CCAC2025 Task 4: Chinese Multimodal Sarcasm Computing Based on Qwen2-Audio

Chengfei Li<sup>1</sup>, Chunyu Wang<sup>2</sup>, Yunjie Zhang<sup>1</sup>, Hanlin Li<sup>1</sup>, Wenya Zhang<sup>1</sup>, Wenkun Chen<sup>1</sup>, Dekun Liu<sup>1</sup>

1. Artificial Intelligence Research Institute on Education, Qilu Normal University  
2. School of Geography and Tourism, Qilu Normal University

**Abstract :** This paper presents a technical solution for CCAC 2025 Chinese multimodal sarcasm computing evaluation based on the Qwen2-Audio model. Specialized training strategies and data augmentation methods are designed for multimodal sarcasm detection and category identification tasks. The framework employs LoRA fine-tuning techniques, combined with cross-task transfer learning and data balancing strategies, utilizing the "Tucao Dahui" video dataset for targeted fine-tuning. In the CCAC2025 evaluation, the approach achieved macro F1 scores of 0.7245 for Task 1 (multimodal sarcasm detection) and 0.3782 for Task 2 (multimodal sarcasm category identification). The results demonstrate that cross-task transfer learning significantly improves sarcasm detection performance, and simple data balancing strategies are more effective than complex audio augmentation techniques.

**Key words:** Multimodal Learning; Sarcasm Computing; Qwen2-Audio; Transfer Learning; Data Augmentation

## 0 引言

讽刺作为人类语言交流中极具复杂性的表达方式，常常通过多种模态信息共同传递，包括语

音语调、面部表情及肢体动作等非语言信号<sup>[1]</sup>。准确识别讽刺不仅依赖于对文本内容的理解，还需要对语音和视觉等多模态信息进行综合分析。近年来，讽刺识别技术经历了从基于规则的方法到深度学习方法的演进。早期研究主要依赖人工

特征和规则，通过分析词汇极性对比等特征实现讽刺检测。随着深度学习技术的发展，预训练语言模型如 BERT<sup>[2]</sup>、RoBERTa<sup>[3]</sup>等在讽刺任务上取得了显著进展。多模态讽刺检测方法<sup>[4]</sup>进一步引入了跨模态注意力机制，能够更好地捕捉不同模态间的信息不一致性，从而提升模型的综合理解能力。

在中文语境下，讽刺识别面临着独特的挑战。中文讽刺表达更加含蓄，常通过暗示、隐喻等方式传递讽刺意图，且需要结合文化背景进行理解。此外，中文作为声调语言，语调变化不仅影响词汇含义，还可能传达讽刺情感。近年来，Qwen2-Audio<sup>[5]</sup>等大型多模态预训练模型的出现，为音频和文本的联合理解提供了技术基础。

本研究聚焦于 CCAC 2025 中文多模态讽刺计算评测任务。该评测基于《吐槽大会》数据集，设置了多模态讽刺识别（二分类）和讽刺类别识别（六分类）两个子任务。针对上述任务，本文提出了基于 Qwen2-Audio 模型的技术方案，系统探索了跨任务迁移学习和多种数据增强策略的效果，并在评测任务上进行了全面的实验分析。研究目标在于为中文多模态讽刺计算提供有效的技术路径和实验基线。

## 1 数据集与任务定义

本研究所用数据集来源于国内知名脱口秀节目《吐槽大会》。该数据集具备多模态特征，每个样本均包含完整的视频片段、对应的音频信号以及人工转录的文本内容。数据标注信息包括片段编号、说话人、时间戳、文本内容、是否讽刺及讽刺类别等，为多模态讽刺识别任务提供了坚实的数据基础。数据的多样性和真实性有助于提升模型在实际应用场景下的泛化能力。

CCAC 2025 评测任务包括两个子任务。任务一为多模态讽刺识别（二分类），要求模型判断给定视频片段是否包含讽刺成分。任务二为多模态讽刺类别识别（六分类），要求模型进一步区分讽刺片段的具体类别，包括 Sarcasm（嘲笑）、Irony（反语）、Satire（讽刺文学）、Understatement（轻描淡写）、Overstatement（夸大）、Rhetorical question（反问）。两个任务均采用 macro F1 分数作为主要评价指标，以全面衡量模型在各类别上的表现。

## 2 方法论

### 2.1 模型选择与训练策略

为应对多模态讽刺识别的复杂性，本文对比了 Qwen3-8B、Qwen2.5-VL、SenseVoice 等多种模型方案。最终选用 Qwen2-Audio 作为基座模型，主要基于其为音频-文本多模态任务设计，具备原生音频理解能力，并在大规模音频-文本数据上完成预训练，获得了良好的跨模态表示能力。Qwen2-Audio 支持参数高效微调策略，模型结构如图 1 所示。

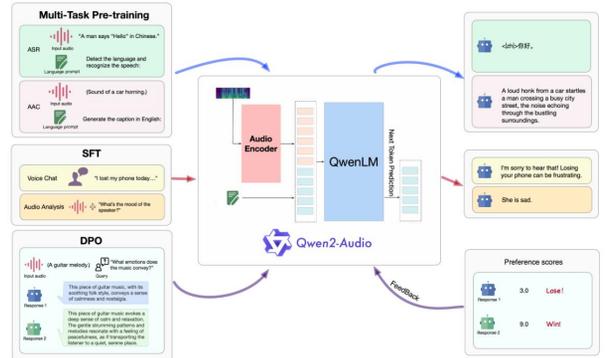


图 1 Qwen2-Audio 模型结构

在模型训练过程中，采用 LoRA 微调策略，具体参数为 LoRA 秩 32、Alpha 参数 32、学习率  $5e-5$ ，训练 3 个 epoch，批次大小 8，梯度累积步数 4。该配置在模型表达能力和训练效率之间取得了较好平衡，能够有效提升模型在讽刺识别任务上的表现。

### 2.2 数据增强策略

针对类别不平衡问题，设计了多种数据增强方法<sup>[6]</sup>。音频增强方面，对标签 1-4 的数据进行 0.9 倍和 1.1 倍速度变换，增加样本多样性。TTS 增强方面，利用文本转语音技术为少数类别生成额外音频数据。数据平衡方面，通过复制少数类别样本，直接缓解类别分布不均。实验结果表明，简单的数据复制策略在本任务中表现更为有效。

### 2.3 跨任务迁移学习

在模型优化过程中，采用跨任务迁移学习策略。具体做法为：使用任务二（六分类）训练的模型，通过标签映射方式应用于任务一（二分类），即将所有讽刺类别统一映射为讽刺标签，非讽刺样本标签保持不变。实验显示，该策略能够显著提升二分类任务的识别性能。

## 3 实验设计与结果

### 3.1 任务一独立优化实验

首先, 针对任务一(二分类)建立了基线模型。初步实验结果显示, 模型在未优化参数时的 F1 分数为 0.6629。随后, 通过调整 LoRA 配置参数, F1 分数提升至 0.6846(表 1)。这一结果表明, 合理的参数优化能够有效改善模型在讽刺识别任务中的表现, 为后续实验提供了坚实的基础。

表 1 任务一独立优化实验结果

| ID | 数据增强 | LoRA 配置                               | F1     |
|----|------|---------------------------------------|--------|
| 1  | 无    | 默认<br>(rank=16, alpha=32,<br>lr=1e-4) | 0.6629 |
| 2  | 无    | rank=32                               | 0.6846 |

### 3.2 任务二独立优化实验

任务二作为六分类问题具有更高的复杂度, 进行了系统的数据增强和训练策略探索, 以期找到有效的优化方法(表 2)。

表 2 任务二独立优化实验结果

| ID | 数据增强                        | LoRA 配置            | F1     |
|----|-----------------------------|--------------------|--------|
| 3  | 无                           | 默认                 | 0.3241 |
| 4  | 标签 1-4 速度变换<br>(0.9, 1.1 倍) | 默认                 | 0.3241 |
| 5  | 标签 1-4 Mixup                | 默认                 | 0.3026 |
| 6  | 标签 1-4 RIR                  | 默认                 | 0.2693 |
| 7  | 速度变换+标签 3 复制 3 份            | 默认                 | 0.3161 |
| 8  | 标签 1-4 速度变换                 | rank=32            | 0.3050 |
| 9  | 标签 1-4 速度变换                 | rank=32<br>lr=5e-5 | 0.3269 |

实验发现, 采用 Mixup 和 RIR 等复杂音频增强技术后, 模型性能不升反降, 分析认为这些方法可能干扰了讽刺检测所需的关键韵律特征, 因此在本任务中并不适用。相比之下, 简单的速度变换和参数优化对模型效果更为有利。

### 3.3 跨任务迁移实验

在独立优化遇到瓶颈后, 进一步尝试将任务二(六分类)训练得到的模型通过标签映射方式应用于任务一(二分类)。具体做法是将所有讽刺类别统一映射为讽刺标签, 非讽刺样本标签保持不变。实验结果显示, 迁移后 F1 分数从 0.6846 提升到 0.718, 提升幅度为 4.9%。这一现象表明, 细粒度分类任务的学习有助于提升粗粒度分类任务的识别能力。

### 3.4 联合优化系统验证

基于跨任务迁移学习的重要发现, 我们设计了系统性的联合验证实验。在这一阶段, 每组实验都包含任务二的训练过程和对应的任务一迁移结果, 以全面评估不同策略组合的效果(表 3)。

表 3 联合优化实验结果

| ID | 数据增强           | LoRA 配置             | 任务二 F1 | 任务一迁移 F1      |
|----|----------------|---------------------|--------|---------------|
| 10 | 速度变换           | rank=32             | 0.265  | 0.7184        |
| 11 | 速度变换           | rank=32,<br>lr=5e-5 | 0.3227 | 0.7148        |
| 12 | 速度变换+TTS       | rank=32,<br>lr=5e-5 | 0.3091 | 0.7175        |
| 13 | 标签 1-4<br>复制一份 | rank=32,<br>lr=5e-5 | 0.3461 | <b>0.7238</b> |
| 14 | 标签 1-4<br>复制二份 | rank=32,<br>lr=5e-5 | 0.3280 | 0.7196        |
| 15 | 标签 1-4<br>复制三份 | rank=32,<br>lr=5e-5 | 0.3373 | 0.7058        |

联合验证实验确认了跨任务迁移学习的稳定性和有效性。实验结果显示, “标签 1-4 复制一份”策略在任务二和任务一迁移中均取得了较优表现, 说明适度的数据平衡能够有效缓解类别不均问题, 提升模型的泛化能力。

### 3.5 训练策略对比

为了全面评估不同训练策略的效果, 我们进一步对比了 LoRA 微调与全参数训练的性能。实验结果表明, LoRA 微调在 3 个 epoch 时取得了最佳效果(任务二 F1 为 0.3461, 任务一迁移 F1 为 0.7238)。而全参数训练和过多训练轮次均导致性能下降, 说明过度参数调整会影响模型的泛化能力, 适度的训练轮次和参数微调更有助于模型稳定优化。(表 4)。

表 4 训练策略对比实验结果

| ID | 训练策略    | Epochs | 任务二 F1 | 任务一迁移 F1 |
|----|---------|--------|--------|----------|
| 13 | LoRA 微调 | 3      | 0.3461 | 0.7238   |
| 16 | LoRA 微调 | 4      | 0.2575 | 0.4890   |
| 17 | 全参微调    | 4      | 0.2106 | 0.4326   |

### 3.6 高级策略与最终优化

在完成基础优化后, 为了挖掘模型潜力, 我们进一步探索了模型集成、推理优化和伪标签策略等高级方法(表 5)。

表 5 高级策略实验结果

| ID | 训练策略     | 具体方法                             | 任务二 F1 | 任务一 迁移 F1     |
|----|----------|----------------------------------|--------|---------------|
| 18 | 模型集成     | 投票集成 (13, 14, 15)                | 0.3389 | 0.7178        |
| 19 | 推理优化     | Temperature=0.5                  | 0.3452 | 0.7245        |
| 20 | API+人工审核 | 调用 API 制作 100 条标签 1-4 的伪标签, 人工审核 | 0.3782 | <b>0.7245</b> |

结果显示, API 伪标签结合人工审核的方法在两个任务上均取得了最佳成绩, 显示高质量数据扩充对模型性能的积极作用。推理优化和模型集成也带来了一定提升, 但效果略逊于伪标签方法。

### 3.7 实验结果总结

通过系统性的对比实验探索, 我们确定了针对 CCAC 2025 中文多模态讽刺计算任务的最优技术方案。实验验证了跨任务迁移学习的显著效果, 证明了简单数据平衡策略优于复杂音频增强技术, 并确立了 LoRA 微调在保持跨任务迁移能力方面的优势。

实验最优配置为使用任务二训练模型+标签映射迁移, 配合“标签 1-4 复制一份”的数据平衡策略, 通过高质量伪标签扩充训练数据, LoRA 配置为 rank=32, lr=5e-5, 训练 3 个 epoch, 得到最佳性能。实验结果充分验证了本研究提出的基于 Qwen2-Audio 的技术方案的有效性, 为中文多模态讽刺计算领域提供了完整的实验基线和优化策略。

## 4 结果分析与讨论

跨任务迁移学习的有效性主要体现在以下几个方面: (1) 任务二要求模型学习区分不同讽刺类型的细微特征, 丰富了特征表示; (2) 六分类标签提供了更丰富的监督信号, 有助于模型学习更优的语义表示; (3) 多分类任务具有正则化效果, 能够防止模型过拟合特定模式, 从而提升泛化能力。

简单数据复制策略优于复杂音频增强的原因可能包括: (1) 在数据规模有限时, 增加样本数量比增加样本多样性更为重要; (2) 复杂音频增强可能破坏讽刺检测所需的关键韵律特征; (3)

数据复制能够直接缓解类别不平衡问题, 提升模型对少数类别的识别能力。

## 5 结论

本文针对 CCAC 2025 中文多模态讽刺计算评测任务, 提出了基于 Qwen2-Audio 模型的技术方案。系统实验结果表明, 跨任务迁移学习能够有效提升讽刺识别性能, 细粒度任务的学习有助于粗粒度任务的优化; 简单的数据平衡策略优于复杂音频增强方法; LoRA 微调在保持迁移能力方面表现良好。未来研究可从以下方向展开: (1) 设计具备长序列处理能力的模型架构; (2) 融入文化背景知识, 提升模型对文化特定讽刺的理解能力; (3) 开发更精细的韵律特征提取方法; (4) 实现个性化适应策略, 提升模型对不同说话者风格的适应性。

## 参考文献

- [1] Riloff, E., Qadir, A., Surve, P., et al. Sarcasm as contrast between a positive sentiment and negative situation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, 704-714.
- [2] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019, 4171-4186.
- [3] Liu, Y., Ott, M., Goyal, N., et al. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692, 2019.
- [4] Cai, Y., Cai, H., Wan, X. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, 2506-2515.
- [5] Yang, A., Zhang, B., Zhang, B., et al. Qwen2-Audio: Advancing multimodal audio understanding. arXiv preprint arXiv:2407.10759, 2024.
- [6] Shorten, C., Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. Journal of Big Data, 2019, 6(1): 1-48.

# 基于数据增强与模型集成的中文多模态讽刺计算

徐辉<sup>1</sup> 杨其彧<sup>1</sup> 沈武韬<sup>1</sup> 张瑾<sup>1</sup>

(1. 中国科学院计算技术研究所, 北京市 100190)

**摘要:** 针对中文多模态讽刺计算任务, 该文提出了一种基于数据增强和模型集成的多模态讽刺计算的方法。首先, 基于思维链的数据增强, 能够增强文本的情感语义特征, 从而提高模型的逻辑性和泛化能力。其次, 融合文本、音频、视频的多模态数据特征, 以提高模型特征提取性能。最后, 通过集成多个下游小模型, 利用多个模型的优势, 减少单一模型的弱点。实验采用组委会公布的数据集, 实验结果证明, 基于数据增强和模型集成的多模态讽刺计算方法在数据集上的 macro F1 指标取得了良好的效果。

**关键词:** 讽刺情感计算; 多模态; 数据增强; 模型集成

中图分类号: TP391

文献标识码: A

## Multimodal Sarcasm Calculation for Chinese Content via Data Augmentation and Model Ensemble

XU Hui<sup>1</sup>, YANG Qixu<sup>1</sup>, SHEN Wutao<sup>1</sup>, ZHANG Jin<sup>1</sup>

(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

**Abstract :** Aiming to enhance sarcasm detection in Chinese multimodal sarcasm calculation tasks, this paper proposed a method combining data augmentation and model ensemble. The method employed chain-of-thought-based data augmentation to enrich textual emotional and semantic features, which enhanced model generalization. Multimodal representations from text, audio, and video streams were fused through cross-modal attention mechanisms. An ensemble of downstream models combined complementary predictions while mitigating individual biases. Evaluation on the official benchmark dataset showed that the method achieved competitive performance in macro F1 score.

**Key words:** sarcasm sentiment analysis; multimodal; data augmentation; model ensemble

### 0 任务背景

讽刺作为一种复杂的语言现象, 广泛出现于日常交流、文学创作及娱乐节目中。它的独特之处在于能够通过褒义正面的言语, 巧妙地传达批评、嘲讽等截然相反的深层含义。依据不同的表现形式, 讽刺可进一步细分为多种类型。

针对讽刺的相关研究可以显著提升机器对复

杂语言现象的理解能力, 使人机交互更加准确自然。同时, 其在多媒体内容分析、社交媒体监控以及跨文化交流等方面也展现出了广泛的应用前景。

本次的中文多模态讽刺计算任务要求综合分析 with 深度融合文本、音频和视觉等多种模态的信息, 以实现讽刺表达方式的全面理解和精准识别。

基于数据来源是脱口秀节目《吐槽大会》的

中文多模态讽刺计算数据集，我们需要完成两个子任务：

(1)多模态讽刺识别,对于给定的多模态数据,判断其是否具有讽刺意味,可视为一个判断讽刺与非讽刺的二分类任务。

(2)细粒度的多模态讽刺类别识别,对类别为讽刺的数据,判断数据具体所属的讽刺类别,可视为一个多分类任务。

## 1 任务数据处理

### 1.1 数据清洗

在接收到测试数据集后,我们观察到其第4季第8期中编号为4\_8\_345的测试样本存在视频片段起始时间为空值的问题;编号为4\_8\_606的测试样本存在视频片段起始时间大于结束时间的问题。

由于测试集数据无法直接删除,为正确提取视频与音频特征,我们通过人工确认样本对应的视频片段的方式,分别将两个问题样本的起始时间校准为00:38:09,000与01:09:32,895。

### 1.2 数据预处理

#### 1.2.1 文本数据预处理

训练数据集与测试数据集的样本中,与文本相关的标注包括说话人、文本内容与讽刺目标。我们认为讽刺目标在讽刺识别的实际应用场景中无法提前获知,故仅采用说话人与文本内容构建文本数据。

参考提示词(prompt)学习[1]中人工设计类提示词的思路,我们基于说话人与文本内容构建了如表1所示的提示词模板。

表1 基于说话人与文本内容的提示词构建表

| 输入                | 模板                             |
|-------------------|--------------------------------|
| speaker:<br>说话人   | [SPK]{speaker}[TEXT]{sentence} |
| sentence:<br>文本内容 |                                |

该模板将说话人与文本内容整合为提示文本数据作为后续模型的文本输入,[SPK]与[TEXT]为添加至预训练模型的分词器(tokenizer)词典的特殊标志,用于标注后文内容并缩短提示文本长度。

#### 1.2.2 视频数据预处理

我们观察到训练数据集与测试数据集中各样本对应的视频片段长度不固定。由于视频压缩编码的图像组(Group of pictures, GOP)结构,在依据起始/结束帧位置截取片段时,存在起始帧为预测帧(P帧或B帧)而跳至下一关键帧(I帧)及短片段中关键帧数量少的问题,最终导致图像信息的损失。

针对上述问题,我们采用开源音视频处理工具FFmpeg将视频数据中的每一帧均重新编码为关键帧,以保证片段按帧准确切分并尽可能增大短片段的图片信息量。

视频片段中各帧的图像则使用如表2所示的残差网络(Residual Network, ResNet)的经典参数进行预处理。

表2 残差网络图像预处理参数

| 参数名  | 参数值                   |
|------|-----------------------|
| 分辨率  | 224×224               |
| 全局均值 | [0.485, 0.456, 0.406] |
| 标准差  | [0.229, 0.224, 0.225] |

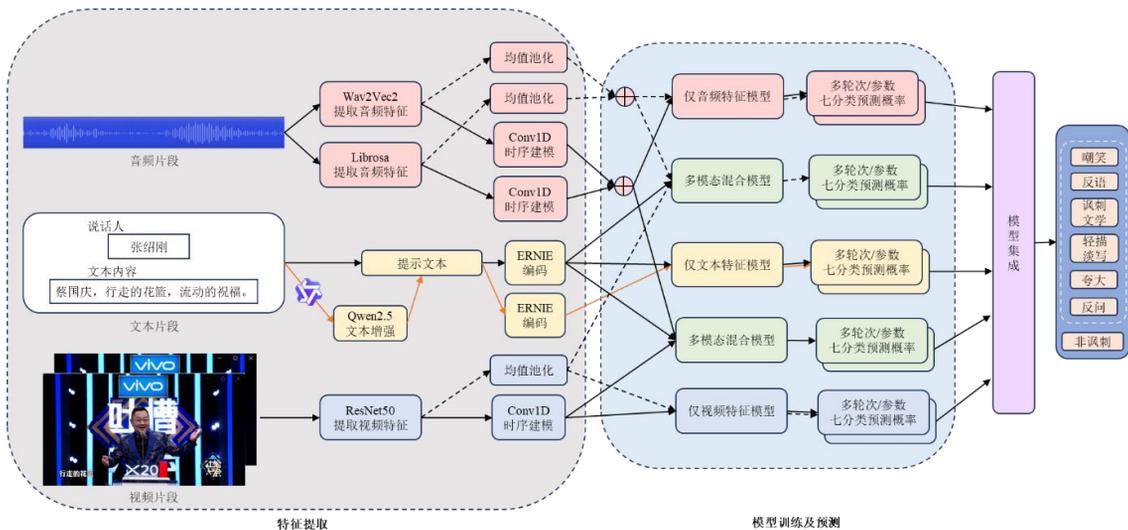


图1 任务设计框架总览图

## 2 任务设计方案

我们对中文多模态讽刺计算任务的整体设计框架总览如图 1 所示。

由于任务二可以视为是在任务一的基础上对讽刺语句进行 0~5 的六个细粒度分类，故我们首先尝试完成第一个任务后，在此之上解决第二个任务。但在训练模型对判断文本是否是讽刺的尝试中，模型在验证集上的分类准确率通常在 70%~80%，即依据任务一的结果，至少有 20% 的数据在任务 2 中天然无法得到正确分类。因此我们将两个任务统一为对给定语句及视频、音频片段，进行包含非讽刺在内的七个分类的预测任务。

任务设计方案的整个流程包括特征提取、模型训练及预测和模型集成，并在最后根据所得的七分类结果将任务二中的涉及讽刺标签的六类标签重新合并，生成任务一的预测标签。

### 2.1 基于思维链的文本数据增强

我们参考了提示工程中的思维链(Chain-of-Thought, CoT)方法<sup>[2]</sup>，使用千问(Qwen)系列大语言模型<sup>[3]</sup>的最新模型 Qwen2.5-32b 并设计了如图 2 所示的提示模板  $P$  对各样本的文本内容进行数据增强。

“我会给你提供文本,请你对文本进行分析。首先,你需要判断文本是否具有讽刺意味,如果没有,输出原因;如果你判断文本有讽刺意味,你可以细分的标签有6种,分别是:【嘲笑】,【反语】,【讽刺文学】,【轻描淡写】,【夸大】,【反问】。其中【嘲笑】,指与事态相矛盾并持批评态度的评论;【反语】,指与事态相矛盾但没有明显批评态度的评论;【讽刺文学】,指看似支持但包含潜在分歧和嘲讽的评论;【轻描淡写】,指降低或破坏事态重要性的评论;【夸大】,指用明显夸张的措辞描述事态的评论;【反问】,指用反问语句引导读者推理出与实际情况明显相悖的含义的评论。例如,文本“谢谢谢谢,谢谢大家的热情。”没有讽刺意味;文本“比如人家爱穿花裤衩怎么了?”可以标为【嘲笑】;文本“不提,不值一提。”可以标为【反语】;文本“任何的男主角在出场之前是不是得有个混混。”可以标为【讽刺文学】;文本“刚才大家提到首次千万融资、首次贴片广告是吧?都黄了。”可以标为【轻描淡写】;文本“前前后后删了大概这个五百款游戏。”可以标为【夸大】;文本“张老师,别的我都认,就是我就是不知道,你是怎么好意思说我油腻?”可以标为【反问】。你需要输出你认为这个文本属于某个类别的原因,但不要输出你认为的文本所属的类别。输出的要求是回答在100字左右。文本是:”

图 2 基于思维链设计的提示模板  $P$

模板  $P$  为大模型详细描述了任务的实现流程,即首先判断文本是否具有讽刺意味,再对具有讽刺意味的文本细分标签;在此基础上通过举例详细解释各标签类别的含义,并询问模型做出判断的推理依据。

图 3 展示了以单个样本的文本内容为例,基于思维链提示模板  $P$  使大模型生成增强文本的流程。

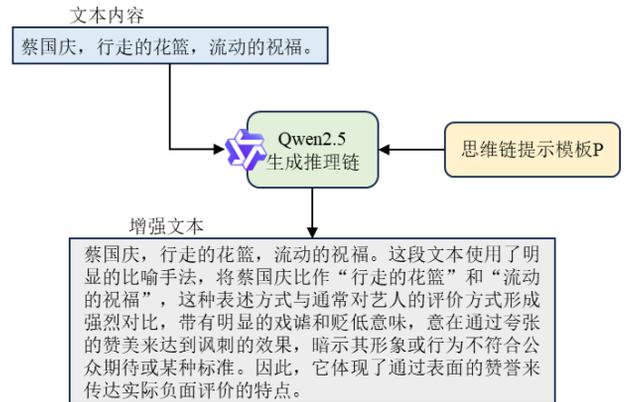


图 3 基于思维链的文本数据增强流程图

流程中我们仅为大模型提供文本内容而不给予类别标签, 以确保训练集与测试集样本输入的一致性。这虽然会影响大模型对样本判断的准确率, 但有效避免了“看着答案推过程”的现象, 提高了大模型的推理链在训练集与测试集上的统一性。

通过引导大模型的推理, 我们获得了带有解释的增强后文本, 从而达到了丰富数据集并提高模型的泛化能力的目的。

### 2.2 特征提取

#### 2.2.1 音频特征提取

我们采用在中文语音识别任务上微调的预训练的大型跨语音建模 Wav2Vec2 模型<sup>[4]</sup>及 Python 的 Libsora 库对音频片段实现音频特征的提取。

其中, Wav2Vec2 从原始音频波形中提取包含语音音素结构、语调模式、节奏信息等上下文感知的深层语音表示, 并对每一帧生成一个 1024 维的特征向量; 而 Libsora 库则从频谱角度提取反应音色特征的梅尔频率倒谱系数(MFCC)与音高(Pitch)。

对单个样本的音频片段  $a$ , 重采样至模型需求的 16Khz 后, Wav2Vec2 模型提取如式(1)所示的深层音频特征  $\mathbf{a}_f$ , Libsora 库提取梅尔频率倒谱系数特征  $\mathbf{a}_{mfcc}$  与音高特征  $\mathbf{a}_{pitch}$ , 并将二者如式(2)标准化后拼接为  $\mathbf{a}_{lib}$ 。

$$\mathbf{a}_f = \text{wav2vec2}(a) \quad (1)$$

$$\mathbf{a}_{lib} = \text{concat} \left( \left[ \frac{\mathbf{a}_{mfcc} - \overline{\mathbf{a}_{mfcc}}}{\text{std}(\mathbf{a}_{mfcc})}, \frac{\mathbf{a}_{pitch} - \overline{\mathbf{a}_{pitch}}}{\text{std}(\mathbf{a}_{pitch})} \right] \right) \quad (2)$$

#### 2.2.2 视频特征提取

在权衡模型运行效率与性能后, 我们采用预训练的残差网络中的 ResNet50 模型<sup>[5]</sup>对视频片段

中的各单帧图像提取 2048 维的视觉语义特征。

对单个样本的视频片段  $v$ ，模型以每秒采样 10 帧的频率对各帧图像进行预处理与特征提取，并最终得到式(3)所示的视频片段特征  $v_f$ 。

$$v_f = \text{resnet50}(v) \quad (3)$$

### 2.2.3 文本编码

我们采用预训练语言模型 ERNIE(Enhanced Representation through kNowledge IntEgration)<sup>[6]</sup>对提示文本数据进行编码，其在 BERT 的基础上引入了实体级掩码与知识增强，编码效果更好。

对样本的提示文本数据  $t$  进行分词器编码并生成填充掩码后，取 ERNIE 模型最后的隐层状态输出作为 768 维的词嵌入向量  $T_f$ ，如式(4)所示。

$$T_f = \text{ernie}(t) \quad (4)$$

### 2.2.4 时序建模与平均池化

由于除文本特征外，所得的视频与音频特征均具有时序维度，故为统一三种特征向量的维度，我们尝试了在时序维度上的平均池化或时序建模两种方法。而相较于平均池化，时序建模能更好的融合各个时间点的特征并直接统一维数，故作为主要方法被我们采用。

我们使用 Conv1D 与激活函数 ReLU，对  $a_f$  与  $a_{lib}$  沿时序维度进行一维卷积得到 768 维的深层音频特征  $Au_f$  与传统音频特征  $Au_{lib}$ ；对  $v_f$  进行一维卷积得到 768 维的视频特征  $Vi_f$ 。

## 2.3 纯文本模型

考虑到讽刺作为一种文学体裁，其主要载体通常是文本信息，而视频信息和音频信息起到辅助判断的作用，因此我们对只使用文本信息完成分类任务进行了建模与尝试。

我们首先使用了文本卷积神经网络(Text Convolutional Neural Network, TextCNN)。作为卷积神经网络在文本分类任务上的应用，其利用多个不同大小的卷积核来提取句子中的关键信息，从而更好地捕捉语句中的局部相关性<sup>[7]</sup>。

在进行基于思维链的文本数据增强后，我们采用了微调预训练语言模型的方法，对文本编码中使用的 ERNIE 模型在增强后的文本数据集上进行微调，以提升模型在讽刺文本分类这一特定任务上的泛化能力。

## 2.4 基于交叉注意力和门控机制的多模态模型

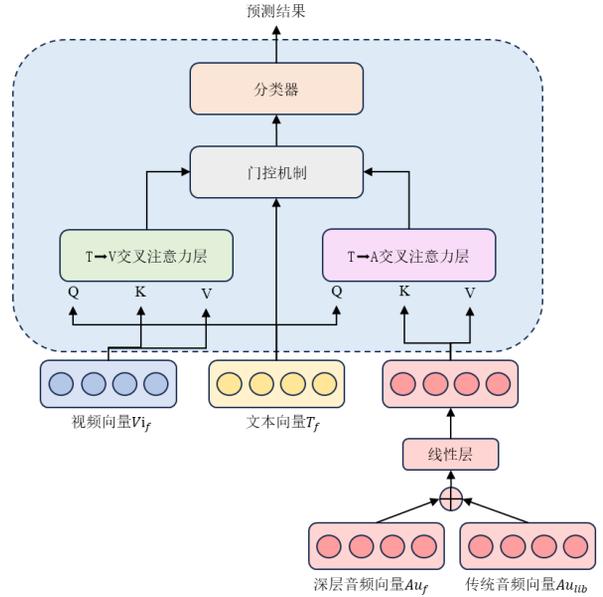


图 4 基于交叉注意力和门控机制的模型框架图

在任务的构建中，最主要的混合多模态模型即为图 4 所示的基于交叉注意力和门控机制的多模态模型。我们首先将深层音频向量  $Au_f$  和传统音频向量  $Au_{lib}$  拼接后由线性投影层统一至 768 维的音频输入；模型将文本向量视为主导内容，视频与音频模态作为辅助，以  $T \rightarrow V$  交叉注意力层为例，文本词嵌入向量  $T_f$  作为质询(Query, Q)的驶入、视频特征向量  $Vi_f$  作为键值(Key, K)与价值(Value, V)的输入，三者经分别线性层投影得到  $Q_t$ 、 $K_{vi}$  与  $V_{vi}$ ；注意力权重  $W_{T \rightarrow V}$  的计算如式(5)所示，其中  $d_k$  为  $K_{vi}$  的维度， $\sqrt{d_k}$  为缩放因子， $Q_t$  与  $K_{vi}$  计算点积相似度，在缩放后通过柔性最大传递函数(Softmax)得到权重。

$$W_{T \rightarrow V} = \text{softmax}\left(\frac{Q_t K_{vi}^T}{\sqrt{d_k}}\right) \quad (5)$$

$$CA_{T \rightarrow V} = W_{T \rightarrow V} V_{vi} \quad (6)$$

最终通过加权计算获得文本融合了视频信息的特征序列  $CA_{T \rightarrow V}$  如式(6)所示。

将两个交叉注意力层输出的新特征序列  $CA_{T \rightarrow V}$  与  $CA_{T \rightarrow A}$  和文本词嵌入向量  $T_f$  共同作为门控机制层的输入。门控机制层如式(7)所示拼接三种输入后，经线性层投影与 Sigmoid 激活函数  $f$  得到门控权重  $g$ 。

$$g = f(\text{linear}(\text{concat}([T_f, CA_{T \rightarrow V}, CA_{T \rightarrow A}]))) \quad (7)$$

$$O_g = gT_f + (1 - g)(\alpha_V CA_{T \rightarrow V} + \alpha_A CA_{T \rightarrow A}) \quad (8)$$

门控机制层输出  $O_g$  如式(8)所示， $\alpha_V$  与  $\alpha_A$  分别为初值 0.5 的可学习的视频权重与可学习的音

频权重；以文本向量 $T_f$ 作为主体分配门控权重 $g$ ， $CA_{T \rightarrow V}$ 与 $CA_{T \rightarrow A}$ 则由 $\alpha_V$ 与 $\alpha_A$ 控制权重后分配 $1-g$ ，最终 $O_g$ 即为经门控机制后动态融合文本和跨模态表示的特征。

最终，线性分类器输出将 $O_g$ 映射为七分类的预测结果向量。

## 2.5 模型集成

在完成的过程中，我们采用了如纯文本模型和多模态模型等多种模型，并通过参数调整和训练，获得了同一模型的不同参数版本。

为进一步利用上述的多版本、多类别模型各自的优势，获得更准确的分类结果，我们参考了集成学习的思路，即使用模型集成方法，结合多个性能互补的模型，有效降低模型在特定样本上的偏差问题，从而提升整体的性能。

我们尝试了三种方法进行模型集成：

(1) 硬投票法(Soft Voting)，使用各模型对各样本做出一个“离散的预测类别”。对于每个样本进行预测结果的统计，选择出现次数最多的类别作为该样本的最终预测结果。

(2) 软投票法(Hard Voting)，使用各模型对各样本给出一个柔性最大传递函数计算的概率分布。对于每个样本，将所有模型的概率结果按类别进行平均，最终选择平均概率最大的类别作为该样本的预测结果。

(3) 最大置信度法，对于每个样本，找到各个模型的预测结果中最大概率所对应的类别，然后从所有模型中选择最大概率对应的类别作为最终预测。

我们通过枚举法生成模型的组合形式，并在验证集上依次使用三种方法进行类别的预测，以获得最优组合与效果。

## 3 相关实验

### 3.1 数据集划分与参数设置

我们将任务所提供的训练集数据按 75%与 25%的比例划分为训练集与验证集，各模型在训练集的 7595 个样本上进行训练、在验证集 2532 个样本上进行验证。

由于训练数据集中非讽刺类别与其他各类别标签对应的数据量差异过大，我们在训练时采用如下方法：

非纯文本模型对训练集中的各类别数据的权重进行了计算，实现基于权重的重采样处理。

纯文本模型则对类别较少的类别的数据进行自我复制，确保数据的量级相近。在整体训练集

数据的角度上，将 1965:55:140:18:28:288:7633 的七个类别的数据提升至 5895:1650:2100:720:1120:5760:7633。

以模型在任务一与任务二上的 F1 均值作为衡量指标，并在最后对表现较优的模型在验证集上以更小的轮次与学习率进行训练。训练模型所用的参数设置如表 3 所示。

表 3 参数设置表

| 参数名            | 参数值          |
|----------------|--------------|
| batch_size     | 32           |
| max_seq_length | 80/250(增强文本) |
| random_seed    | 42           |
| hidden_size    | 768          |
| learning_rate  | 1e-6~5e-5    |

### 3.2 实验结果

我们在任务中设计的基于交叉注意力和门控机制的多模态模型在训练数据集上使用未增强的文本数据完成 100 轮次的训练后，其在 1e-5 至 5e-5 学习率下验证集上对两个任务以各自 F1 值的均值所计算的最优得分结果如表 4 所示。

表 4 多模态模型训练结果

| 学习率  | 最优得分   |
|------|--------|
| 1e-5 | 0.4872 |
| 2e-5 | 0.5014 |
| 3e-5 | 0.5143 |
| 4e-5 | 0.4946 |
| 5e-5 | 0.4835 |

在学习率为 3e-5 时，该多模态模型在验证集中获得最佳表现，随后得分随着学习率的增高而缓慢下降。

模型训练中使用的各模型在验证集中记录的最优表现如表 5 所示，特征默认采用时序建模与未增强文本，多模态模型默认门控参数可学习。

表 5 各模型最优得分

| 模型                 | 得分     |
|--------------------|--------|
| TextCNN            | 0.4523 |
| ERNIE              | 0.4651 |
| ERNIE(增强文本)        | 0.4880 |
| 仅音频模型              | 0.4488 |
| 仅视频模型              | 0.4094 |
| 多模态模型              | 0.5143 |
| 多模态模型(门控参数固定)      | 0.5013 |
| 多模态模型(平均池化)        | 0.4861 |
| 多模态模型(向量拼接)        | 0.4817 |
| 模型集成(硬投票, 八个模型时)   | 0.5526 |
| 模型集成(软投票, 八个模型时)   | 0.5389 |
| 模型集成(最大置信度, 三个模型时) | 0.5216 |

任务中，在仅文本模型中，ERNIE 预训练模型效果要优于 TextCNN 与其他单模态模型，使用

增强文本进行微调后效果进一步提高。基于交叉注意力和门控机制的多模态模型则在各模型中表现仍为最优。

而在进行模型的组合集成后，我们观察到验证集得分得到了明显的提高，且硬投票的表现优于软投票与最大置信度方法。

## 4 总结

针对本次的中文多模态讽刺计算任务，我们将两个任务合并为了一个七分类任务的方案，设计了特征提取、模型训练及预测和模型集成的技术框架。

在特征提取环节，我们选择了 ResNet50 处理视频数据、选择 Wav2Vec2 和 Librosa 库共同处理音频数据，选择 ERNIE 处理文本数据并使用思维链提示模板与 Qwen2.5 大语言模型进行了文本数据的增强。

在模型训练与预测环节，我们尝试了 TextCNN 与 ERNIE 等纯文本模型，并使用增强文本进行微调；设计了基于交叉注意力和门控机制的多模态模型，并在训练集与验证集中尝试对改变特征处理方式、模态数量或模型结构的更多不同版本的模型进行了训练与验证。

最后，我们通过多个模型以三种方法在验证集上基于枚举法的模型组合集成，进一步有效的提高了验证集上的任务得分，完成了基于数据增强与模型集成的中文多模态讽刺计算任务。

## 参考文献

- [1] Schick T, Schütze H. Exploiting cloze questions for few shot text classification and natural language inference[J]. arXiv preprint arXiv:2001.07676, 2020.
- [2] 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.
- [3] Bai J, Bai S, Chu Y, et al. Qwen technical report[J]. arXiv preprint arXiv:2309.16609, 2023.
- [4] Baeovski A, Zhou Y, Mohamed A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[J]. Advances in neural information processing systems, 2020, 33: 12449-12460.
- [5] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
- [6] Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223, 2019.
- [7] Zhang Y, Wallace B. A sensitivity analysis of (and

practitioners' guide to) convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1510.03820, 2015.

# 中文多模态讽刺识别评测方案

胡少涛<sup>1</sup> 陈冠熊<sup>2</sup> 张怡<sup>3</sup>

(1. 新疆大学 计算机科学与技术学院, 新疆维吾尔自治区 乌鲁木齐市 830046)

**摘要:** 该报告介绍了团队在 CCAC2025 中文多模态讽刺识别赛道的参赛方案。针对中文文本、音频与视觉融合的讽刺识别任务, 我们提出了基于动态跨模态对齐与加权晚期融合模型架构。通过 BERT-base-chinese、WavLM-base 及改进高斯生成器与 Dino-vitb16 提取模态特征, 结合带权重的晚期融合策略, 最终在测试集上取得较好分数, 位列第四名, 验证了方案的有效性。

**关键词:** 中文多模态讽刺识别; 多模态; 讽刺识别

## Evaluation Scheme for Chinese Multimodal Sarcasm Recognition

Hu Shaotao<sup>1</sup>, Chen Guanxiong<sup>1</sup>, Zhang Yi<sup>1</sup>

(1. College of Computer Science and Technology, Xinjiang University, Urumqi, Xinjiang Uygur Autonomous Region 830046, China)

**Abstract:** This report presents the team's entry scheme for the CCAC2025 Chinese Multimodal Sarcasm Recognition track. For the task of recognizing satire through the fusion of Chinese text, audio, and vision, we propose a model architecture based on dynamic cross-modal alignment and weighted late fusion. Modal features are extracted using BERT-base-chinese, WavLM-base, and Dino-vitb16 with improved GCG. Combined with the weighted late fusion strategy, the model finally achieved a score on the test set and ranked fourth, verifying the effectiveness of the scheme.

**Key words:** Chinese Multimodal Sarcasm Recognition; Multimodal; Sarcasm Recognition

跨模态一致性建模。

## 1 引言

中文多模态讽刺识别需处理文本反话、音频语调反差与视觉隐喻的协同语义, 其核心挑战在于跨模态信息的动态关联与文化特异性表达的捕捉。传统融合方法难以适配模态贡献度的动态变化, 而单一模态特征提取又易丢失讽刺信号的互补信息。

本方案创新点在于: (1) 优化视觉关键帧选择策略, 通过多尺度高斯掩码提升动态场景下的特征针对性; (2) 采用加权晚期融合技术, 结合 focal loss 平衡模态权重, 强化讽刺信号的

## 2 技术方案

### 2.1 文本模态

在文本模态方面, 我们借助 Bert-base-chinese (Devlin 等, 2019) 来提取文本特征表示。具体而言, 将文本序列  $x$  输入 BERT 模型, 获取其最后一层的输出  $H_t \in \mathbb{R}^{1 \times dt}$  作为该句子的特征向量。在此过程中,  $1t$  代表文本序列的长度, 而  $dt = 768$  则为文本特征的输出维度。

### 2.2 音频模态

在音频模态的处理过程中，我们运用 WavLM-base 来获取音频特征表示。具体而言，首先借助 librosa 库读取音频文件，将其采样率设定为 16000Hz。随后，依据字幕文件中所记录的开始与结束时间戳，精准地截取出对应的音频片段。将这些片段输入至 WavLM 模型之中，进而获取模型最后一层的输出  $H_a \in \mathbb{R}^{l_a \times d_a}$  作为音频的特征向量，其中  $l_a$  代表音频序列的长度， $d_a = 768$  则为音频特征的输出维度。

## 2.3 视觉模态

在进行视觉模态处理时，我们首先对视频按照提供的起始时间戳进行切分，对于每个视频片段，采用同样的帧率抽取帧图像，并运用 DINO-vitb166 获取了视具体来说，我们将视频片段的帧图像输入到 ViT 中，获取到模型最后一层输出  $H_v \in \mathbb{R}^{n \times l_v \times d_v}$ ，其中  $n$  表示该片段的关键帧数量， $l_v$  表示序列长度， $d_v = 768$  为视觉输出维度。

在 GCG<sup>[1]</sup> 中，高斯掩码的宽度  $\sigma$  是固定的，这使得模型在处理不同视频时能够保持一致的关键帧选择策略。然而，这种固定宽度的设置也限制了模型对不同视频的适应性。为了提高模型的灵活性和适应性，我们对 GCG 进行了改进，使其能够动态调整高斯掩码的宽度，并采用多尺度高斯掩码来提取关键帧。

具体改进如下：

1. 动态调整高斯掩码宽度：我们引入可学习参数来动态调整高斯掩码的宽度，使其能够根据视频内容和问题的复杂性自适应地变化。

2. 多尺度高斯掩码：为了更好地捕捉不同时间尺度上的关键信息，我们设计了多尺度高斯掩码。每个高斯掩码对应一个特定的时间尺度，从而能够更全面地覆盖视频中的关键帧。

在实现这些改进时，我们首先使用提示模板，通过 DeepSeek-v3 生成视频片段对应的事件描述。然后，将这些事件描述输入 BERT 模型，得到事件描述的特征  $d[\text{CLS}] \in \mathbb{R}^{d_t}$ 。对于每个视频片段的帧图像特征，将其映射到视频特征  $E[\text{CLS}] \in \mathbb{R}^{n \times d_t}$ 。接下来，我们将视频特征  $E[\text{CLS}]$  与事件描述  $d[\text{CLS}]$  进行拼接，得到联合复杂度 `combined_features`。

我们采用标准 Transformer 编码器来建立跨帧动态和跨模态交互关系。通过这种方式，我们能够学习到视频特征和事件描述之间的复杂交互，并生成  $K$  个可学习的高斯函数的中心  $\mu \in \mathbb{R}^k$ 。

这些高斯函数的中心将用于动态选择与问题最相关的视频帧，从而提高模型在视频问答任务中的性能。

### 2.3.1 多尺度宽度预测

由于不同掩码宽度导致模型专注的时间范围不同，宽度较大的高斯掩码会覆盖更长的时间范围，而宽度较小的高斯掩码则更专注于较短的时间范围。我们使用了多少个不同尺度的高斯掩码。每个高斯掩码中心对应一个特定的时间尺度，用于捕捉视频中不同时间范围内的关键帧。

对于每个视频，我们设置超参数  $T$ ，其决定了生成多少个不同尺度的宽度。因此，多尺度宽度预测：

$$\sigma_T = \text{softplus}(W_T * \text{combined\_features} + b_T)$$

其中， $T$  表示第  $T$  个尺度， $W_T$  和  $b_T$  是第  $T$  个宽度预测器的权重和偏置。

多尺度掩码生成：

$$g_k^t = \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left(-\frac{(n/N - \mu_k)^2}{2\sigma_t^2}\right), g_k^t \in \mathbb{R}^{N \times T}$$

$$k = \{1, 2, \dots, K\}, n = \{1, 2, \dots, N\}, t = \{1, 2, \dots, T\}$$

### 2.3.2 多尺度掩码组合

随后，通过将每个  $g_k^t$  求和，生成每个视频片段的权重分布  $p \in \mathbb{R}^N$ ：

$$p = \text{Norm}\left(\sum_{k=1}^K \sum_{t=1}^T g_k^t\right), p \in \mathbb{R}^N$$

### 2.3.3 关键帧选择

对于视频，提取每一帧的特征嵌入；对于事件描述，提取其文本嵌入，使用余弦相似度计算每一帧的特征嵌入与事件描述的文本嵌入之间的相似度。得到一个相似度分数序列，每个分数表示对应帧与事件描述的相似程度。根据相似度分数序列，选择分数最高的前  $K$  个索引  $\omega$  作为弱标注的关键时间戳。

同样的， $p$  中前  $K$  个峰值对应的索引  $\{\mu_1, \mu_2, \dots, \mu_k\}$ ，这些索引表示与问题最相关的  $K$  个视频时刻，将其进行映射即可得到相应的视频帧图像的序号，这些序号代表着与问题最相关的  $K$  个关键帧，因此我们使用回归目标来优化高斯生成器，以衡量预测的中心  $\mu$  和弱标签的时间戳  $\omega$  之间的差异：

$$L_{reg} = \sum_{k=1}^K \text{Smooth}_{L1} \|\mu_k - \omega_k/T\|$$

在跨模态嵌入空间中，事件描述的嵌  $d[CLS]$  应与从权重分布  $p$  中提取的所选正帧  $E_{pos} \in \mathbb{R}^{K \times d}$  高度对齐，且与无关帧距离较远， $E_{pos}$  也是用于最终答案预测的视觉输入。因此，我们沿用 Wang 等<sup>[1]</sup>的正时刻选择、负时刻挖掘的方法，通过从  $E[CLS]$  中采样权重最低的  $N_{intra}$  帧，同时采用同一批次内其他视频中随机采样的  $N_{inter}$  帧作为跨视频负面帧  $E_{inter} \in \mathbb{R}^{N_{intra} \times d}$ 。这些来自同视频与其他视频的负面样本能提供丰富信息。通过通过 infoNCE 损失函数来在联合嵌入空间中最大化查询  $d[CLS]$  与对应正视频时刻组  $E_{pos}$  的相似度，同时拉大与负时刻的距离：

$$L_{con} = - \sum_{k=1}^K \frac{1}{K} \log \frac{\exp(d[CLS] \otimes E_k^{pos}/\tau)}{\exp(d[CLS] \otimes E_k^{pos}/\tau) + SUM}$$

$$SUM = \sum_{i=1}^{N_{intra}} \exp(d[CLS] \otimes E_k^{intra}/\tau) + \left( \sum_{j=1}^{N_{inter}} \exp(d[CLS] \otimes E_k^{inter}/\tau) \right)$$

## 2.4 多模态融合

多模态融合是整个模型架构的核心组成部分，其设计目的在于有效整合音频、文本和视频三种模态的特征，挖掘它们之间的潜在关联，从而提升对讽刺内容识别的准确性与鲁棒性。<sup>[2]</sup>我们采用了带模态特定权重的晚期融合策略，并结合梯度下降优化权重以及带焦点处理样本不平衡问题，具体实现步骤如下：

### 2.4.1 特征变换单元

在之前的步骤中，特征向量已经在各自的模态内捕捉了丰富的时序信息，例如音频特征中蕴含了语音的韵律变化、节奏快慢等信息；文本特征包含了语句的语义结构、词汇情感倾向等内容；视频特征则涵盖了面部表情的动态变化、场景元素的呈现等关键信息。通过维度转换和线性变换对原始特征输入数据进行处理，为此采用了一个线性层。

### 2.4.2 GRU 编码器

为了充分发挥不同模态在幽默识别任务中的独特作用，我们为每个模态分配了可学习的权重  $w_A, w_T, w_V \in \mathbb{R}^+$ 。在初始阶段，将这些权重均设为 1.0，这是一种无偏向的初始化方式，意味着在模型训练的起始阶段，每个模态对融合结果的贡献被认为是同等重要的。随着训练的推进，通过梯度下降算法对这些权重进行优化调整。

权重的学习率与主模型保持一致，设为  $1e-5$ ，在保证模型收敛速度的同时，避免了因学习率过大导致模型不稳定或因学习率过小导致训练时间过长的问题。

GRU 编码器负责对输入特征进行编码并捕获序列中的时序信息，能够处理变长序列数据并提取显著特征。为增强建模能力，我们引入了双向编码和多层 RNN 结构。编码器的输入来自输入层的输出以及输入序列的有效长度，编码器的输出记为。

### 2.4.3 分类输出

为增强模型的非线性建模能力，我们采用了带有隐藏单元的全连接层。输出层的输入  $X_{enc}$  是编码器的输出。该层通过一系列线性变换、激活函数和 dropout 操作生成预测输出  $X_{out}$ ：

$$X_{out} = \text{Linear2}(\text{ReLU}(\text{Dropout}(\text{Linear1}(X_{enc}))))$$

其中  $\text{Linear1}$  和  $\text{Linear2}$  表示两个线性层， $\text{ReLU}$  代表修正线性单元激活函数。

为使预测结果映射到合适范围，我们采用 Sigmoid 激活函数作为最终激活函数。该函数将输出值映射到 0 至 1 之间的概率范围。

### 2.4.4 加权融合

在获得各模态的特征向量以及对应的权重后，我们通过加权融合的方式计算融合特征  $F_{fusion}$ ，计算公式如下：

$$F_{fusion} = \frac{w_A \cdot F_A + w_T \cdot F_T + w_V \cdot F_V}{w_A + w_T + w_V}$$

其中分子部分是各模态特征与其对应权重的加权和，它体现了每个模态对融合特征的贡献程度。分母部分则是所有权重之和，其作用是对加权和进行归一化处理，确保融合特征的尺度一致性，避免因权重大小差异导致某些模态的影响过度放大或缩小。通过这种加权融合方式，模型能够根据训练数据动态地调整各模态特征的融合比例，充分发挥不同模态的优势。

晚期融合方法中采用了一个线性层。该线性层接收每个模型输出的讽刺预测概率得分作为输入，目的是提升整体性能。为解决存在的类别不平衡问题并有效处理数据不平衡状况，我们采用带焦点损失：

$$FL(p_t) = -(1 - p_t)^y * \log(p_t)$$

$$p_t = \begin{cases} p & \text{if } y \text{ is } 1 \\ 1 - p & \text{otherwise} \end{cases}$$

## 参考文献

- [1] Wang H, Lai C, Sun Y, et al. Weakly supervised gaussian contrastive grounding with large multimodal models for video question answering[C]//Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 5289-5298.
- [2] Xie H, Cui J, Cao Y, et al. Multimodal Cross-Lingual Features and Weight Fusion for Cross-Cultural Humor Detection[C]//Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation. 2023: 51-57.

# 基于预训练模型与 Transformer 融合机制的中文多模态讽刺计算

鄂福顺<sup>1</sup> 吴锐<sup>1</sup>

(1. 哈尔滨工业大学 计算学部, 黑龙江 哈尔滨 邮编 150001)

**摘要:** 该文针对中文多模态讽刺计算任务, 提出了一种融合预训练模型与 Transformer 融合机制的讽刺识别方法。框架基于 RoBERTa 与 HuBERT, 分别提取文本与音频特征, 结合跨模态交叉注意力与单模态自注意力机制, 实现多模态信息的协同融合与独立建模, 提升模型对讽刺语义、语调变化与上下文依赖的理解能力。此外, 引入多任务联合学习与多损失优化策略, 兼顾讽刺检测与类型分类的性能。基于《吐槽大会》节目构建的中文多模态讽刺数据集的实证结果表明, 所提方法在官方竞赛中取得 0.5107 的总分, 排名第五, 验证了其在多模态讽刺识别任务中的有效性。

**关键词:** 讽刺计算; 多模态融合; 预训练模型

中图分类号: TP391

文献标识码: A

## Chinese Multimodal Sarcasm Computing Based on Pretrained Models and Transformer Fusion Mechanism

**Abstract :** This study proposes a sarcasm recognition method for Chinese multimodal sarcasm computing, which integrates pretrained models with a Transformer-based fusion mechanism. The framework leverages RoBERTa and HuBERT to extract text and audio features, respectively. A cross-modal attention mechanism combined with unimodal self-attention enables collaborative fusion and independent modeling of multimodal information, enhancing the model's ability to capture sarcasm semantics, prosodic variations, and contextual dependencies. Furthermore, a multi-task joint learning strategy with multi-loss optimization is introduced to improve the performance of both sarcasm detection and type classification. Experimental results on a Chinese multimodal sarcasm dataset constructed from the "Roast" talk show demonstrate that the proposed method achieved an overall competition score of 0.5107, ranking fifth, which verifies its effectiveness in multimodal sarcasm recognition tasks.

**Key words:** Sarcasm Computing; Multimodal Fusion; Pretrained Models

## 0 引言

讽刺作为语言表达中一种极具隐蔽性与复杂性的现象, 通常通过表面上看似中立甚至积极的措辞, 隐含传递否定、批判或反讽等真实意图。

这类表达广泛存在于日常交流、文学作品以及各类娱乐内容中, 蕴含丰富的语义层次与情绪信息。然而, 由于讽刺语言通常依赖于上下文、语境甚至说话者的潜在立场, 且其表达方式多样、意图隐晦, 导致讽刺信息的自动识别与理解长期以来被视为自然语言处理 (Natural Language Processing, NLP) 领域中的一项核心挑战。如何突

破文本信息本身的局限, 深入挖掘语言背后的隐含语义与情感色彩, 已成为构建高性能讽刺识别系统的关键问题。

近年来, 随着多模态学习技术的发展, 融合多源信息以辅助讽刺识别成为重要研究方向。大量语言学与心理学研究表明, 讽刺表达往往伴随着语调、语速、重音、停顿等非言语信息的变化, 尤其是音调在讽刺传达中扮演着至关重要的角色。音调的升降、夸张的语气、节奏的变化不仅增强了讽刺表达的表现力, 也为机器理解隐藏于语言表象之下的真实意图提供了重要的补充信息。因此, 单一的文本模态难以全面捕捉讽刺表达的多维特征, 而多模态融合, 尤其是将文本与音频信息结合, 成为提升讽刺计算效果的有效手段。

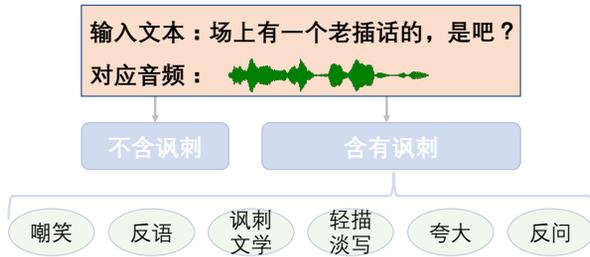


图1 中文多模态讽刺计算任务示例

基于上述背景, 本文面向中文多模态讽刺识别任务, 综合利用文本与音频两种信息源, 旨在判别输入样本中是否包含讽刺表达, 并进一步实现多类别讽刺类型的细粒度区分(如图1所示)。具体而言, 所涉及的讽刺类型包括嘲笑(Sarcasm)、反语(Irony)、讽刺文学(Satire)、轻描淡写(Understatement)、夸大(Overstatement)以及反问(Rhetorical Question)等多种常见表达形式。相比传统的二分类讽刺识别任务, 面向多类别的任务设置更贴合真实语言表达场景, 同时对多模态特征融合、上下文理解及情绪信息建模能力提出了更高要求, 对于推动中文讽刺计算研究与应用具有重要意义。

为此, 本文提出了一种融合预训练模型与Transformer架构的多模态讽刺识别框架, 专注于中文语境下讽刺表达的理解与建模。具体而言, 在文本模态部分, 本文采用端到端的大规模预训练语言模型RoBERTa对文本信息进行深度语义编码, 充分捕捉上下文依赖关系与潜在隐含信息。

音频模态部分则引入自监督预训练模型HuBERT, 系统提取语音信号中的多层次声学特征, 挖掘语调、韵律等与讽刺表达紧密相关的语音信息。为了实现多模态信息的高效协同, 本文设计了一种“跨模态优先、单模态补充”的双通道特征融合机制。首先, 基于跨模态交叉注意力(Cross-Attention)模块, 对文本与音频模态特征进行深度融合, 强化不同模态之间的语义协同与信息交互, 充分挖掘语义信息与声学特征的互补关系。其次, 在各自模态内部引入自注意力(Self-Attention)机制, 进一步对文本与音频特征进行细粒度建模, 增强单模态表达能力, 保留各模态的结构特性与独立信息。该设计有效平衡了多模态融合与单模态特征表达之间的关系, 在提升讽刺语言理解与建模能力的同时, 缓解了讽刺表达中常见的语义歧义与情绪隐含问题, 显著增强了整体系统在中文多模态讽刺识别任务中的准确性与鲁棒性。

## 1 相关研究

### 1.1 预训练模型

近年来, 基于大规模语料的自监督预训练模型在自然语言处理(NLP)领域取得了显著进展, 极大提升了文本理解与下游任务性能。BERT<sup>[1]</sup>Transformer的双向编码机制, 结合掩码语言模型(MLM)和下一句预测(NSP), 增强了语义理解与上下文建模能力, 广泛应用于文本分类、情感分析等任务。其后, RoBERTa<sup>[2]</sup>通过基于取消NSP、优化掩码策略、扩大训练数据及延长训练周期, 进一步提升了模型的表达能力与鲁棒性。大量研究表明, RoBERTa在处理隐含语义、情绪反转及复杂语境时表现尤为优异。鉴于讽刺表达具有丰富的语义隐喻与强烈上下文依赖, 本文选用RoBERTa作为文本编码器, 以深度挖掘语言潜在信息, 提升讽刺理解与识别能力。

音频作为人类交流的重要模态, 在语调、重音、语速及停顿等非语言特征中蕴含丰富的情绪与讽刺线索。传统音频处理依赖梅尔频谱(Mel-Spectrogram)和梅尔倒谱系数(MFCC)等低层特征, 难以捕捉深层语义信息。近年来, 自监督音

频预训练模型快速发展, 代表性方法 HuBERT<sup>[3]</sup> 通过伪标签预测与分层特征学习, 显著增强了对多层次音频信息的建模能力。HuBERT 在语音识别、说话人识别及情绪分析中表现出色。鉴于讽刺表达常伴随细微语调变化及隐含声学信息, 本文基于 HuBERT 构建音频编码器, 提取多层次高质量音频特征, 增强多模态讽刺识别的声学表达能力。

## 1.2 多模态特征融合

近年来, 多模态融合已成为情绪识别与讽刺理解等复杂语言任务中的重要研究方向。早期方法如 Tensor Fusion Network (TFN)<sup>[4]</sup> 通过张量外积方式捕捉不同模态间的高阶交互特征, 有效突破了传统简单特征拼接在表达能力方面的瓶颈。随后, 记忆融合网络 (Memory Fusion Network, MFN)<sup>[5]</sup> 引入记忆单元与动态门控机制, 进一步强化了多模态信息的时序依赖建模能力。针对高阶融合带来的参数冗余问题, Liu 等人提出低秩多模态融合 (Low-rank Multimodal Fusion, LMF)<sup>[6]</sup>, 在有效降低模型复杂度的同时, 兼顾了融合效率与表达能力。

随着 Transformer 架构在多模态领域的广泛应用, Tsai 等人提出的 Multimodal Transformer (MuT)<sup>[7]</sup> 基于多头注意力机制实现了不同模态信息的深度交互, 显著提升了多模态序列的联合建模能力。此类基于 Transformer 的跨模态协同方法进一步验证了其在多模态融合与复杂语言理解任务中的优势, 极大推动了细粒度情绪表达与讽刺建模的发展。

在上述研究基础上, 针对讽刺识别任务中普遍存在的模态异质性大、语义依赖复杂等挑战, 本文设计了一种面向文本与音频双模态信息的基于 Transformer 架构的融合框架。该框架结合跨模态交叉注意力机制与单模态自注意力机制, 旨在高效挖掘不同模态间的深层次关联, 强化模型对讽刺表达中的隐含语义、语调变化与上下文依赖信息的综合建模能力, 从而提升多模态讽刺识别的整体性能。

## 2 研究方法

在本研究中, 所提出的中文多模态讽刺计算模型整体结构如图 2 所示, 主要由特征提取网络与融合网络两部分组成。特征提取网络面向不同模态数据, 提取各自关键信息; 融合网络通过跨模态交互, 建模多模态信息间的深层关联, 提升整体判别能力。此外, 针对融合过程中的单模态弱化问题, 模型设计中引入自注意力机制, 有效保留各模态重要特征, 增强综合表达能力。鉴于任务包含二分类及讽刺类型多分类目标, 模型采用多任务多损失训练策略, 以协同优化不同分类任务的性能。

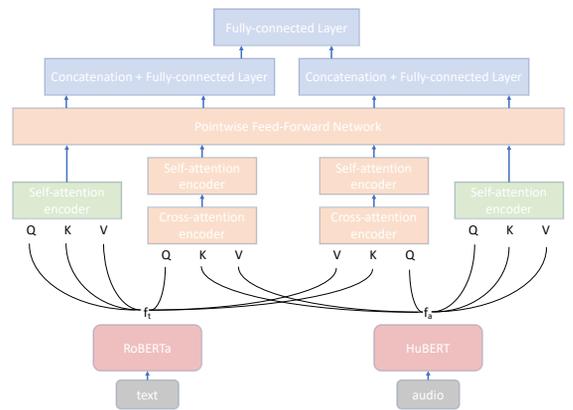


图 2 模型整体架构

### 2.1 单模态特征表示

针对原始输入  $X_m$  (其中  $m \in \{t, a\}$  分别表示文本和音频模态), 本研究设计了一套统一且高效的单模态特征表示方法, 以规整不同模态输入的序列长度与特征维度, 同时保留其关键语义信息。

对于文本模态, 首先利用预训练语言模型 RoBERTa 对输入文本序列进行编码, 得到最后一层的隐层表示:

$$H_t = \text{RoBERTa}(X_t) \in \mathbb{R}^{L_t \times d}$$

其中,  $L_t$  为文本序列长度,  $d=768$  为特征维度,  $M_t$  为对应的注意力掩码。基于掩码信息, 对序列进行有效长度截断, 得到最终文本模态特征表示:

$$f_t = \text{truncate}(H_t, M_t)$$

音频模态借助 HuBERT 模型提取隐层特征及其多层注意力权重, 获得:

$$H_a, A = \text{HuBERT}(X_a), H_a \in \mathbb{R}^{L_a \times d}, A \in \mathbb{R}^{L_a \times L_a}$$

其中,  $L_a$  为音序列长度,  $d=768$  为特征维度,  $A$  为注意力权重矩阵。通过遍历多层注意力, 动

态确定有效长度  $L_a'$ , 并对隐层特征在有效范围内进行均值池化, 得到最终音频模态特征表示:

$$f_a = \frac{1}{L_a'} \sum_{i=1}^{L_a'} H_a [i]$$

同时基于有效长度重构掩码, 保证后续计算中无效部分不被考虑。

通过上述处理, 文本与音频两种模态的单模态特征均规整为统一的维度格式, 便于后续融合与分析。

## 2.2 多模态特征融合

多模态融合网络作为多模态系统的核心, 承担着不同模态信息的深度整合任务。本文所设计的融合网络包括三大关键模块, 旨在实现高效的模态交互与表达能力的整体提升。

首先, 跨模态交叉注意力编码器 (Cross-attention Encoder) 通过引入查询 (Query) 来自一种模态、键 (Key) 和值 (Value) 来自另一模态的机制, 有效捕捉模态间的深层关联, 促进信息协同表达。该机制使不同模态的隐藏状态能够相互映射至对方的特征空间, 助力更全面的多模态语义理解。其核心计算过程如下:

$$\begin{aligned} & \text{Attention}(Q_{m_1}, K_{m_2}, V_{m_2}) \\ &= \text{softmax}\left(\frac{Q_{m_1} K_{m_2}^T}{\sqrt{d}}\right) V_{m_2} \end{aligned}$$

其中,  $Q_{m_1}$  表示来自模态  $m_1$  的查询向量,  $K_{m_2}$  和值  $V_{m_2}$  分别为模态  $m_2$  特征投影所得的键和值,  $d$  为缩放因子。

其次, 网络引入自注意力编码器 (Self-Attention Encoder), 对跨模态交互后的特征序列进行进一步的时序建模与信息整合, 强化模态内部的动态关联与上下文一致性, 提升特征表达的丰富性与鲁棒性。

最后, 逐点前馈网络 (Position-wise Feed-Forward Network) 对每个时间步的特征向量独立施加多层感知机 (MLP) 变换, 结合非线性激活函数, 进一步细化高层表达, 增强模型的特征建模能力与泛化效果。

通过上述模块的协同设计与集成, 融合网络实现了多模态信息的深度交互与高效整合, 显著提升了模型在讽刺识别与复杂情绪理解任务中的

综合性能。

## 2.3 原始模态信息增强

在多模态融合任务中, 跨模态交互机制能够有效促进不同模态信息的协同表达, 然而, 在大量信息交互过程中, 部分原始模态的结构信息与本质语义特征可能逐步被削弱。尤其针对复杂的情绪识别与讽刺理解任务, 保留并强化各单一模态自身的表达能力对于提升整体系统性能至关重要。

为此, 本文在融合网络中设计了原始模态信息增强模块, 分别针对文本与音频模态引入独立的自注意力 (Self-Attention) 机制, 进一步捕捉模态内部的动态依赖关系与语义结构信息, 提升单模态表达的完整性与鲁棒性。两种模态分别通过独立的自注意力编码器进行特征增强, 计算方式如下:

$$\begin{aligned} & \text{Attention}(Q_m, K_m, V_m) \\ &= \text{softmax}\left(\frac{Q_m K_m^T}{\sqrt{d}}\right) V_m \end{aligned}$$

其中,  $Q_m, K_m, V_m$  分别表示查询、键和值, 均由同一模态特征  $f_m$  通过可学习线性映射获得:

$$Q_m = f_m W_Q, K_m = f_m W_K, V_m = f_m W_V$$

其中,  $W_Q, W_K, W_V$  为可学习的参数矩阵。

自注意力机制通过动态建模模态内部任意时间步之间的全局依赖关系, 有效捕捉局部细节与整体结构信息, 进一步强化特征序列的表达能力与信息完整性。

通过上述原始模态信息增强设计, 模型在保留各模态内部时序结构与语义特征的基础上, 缓解了多轮跨模态交互过程中原始信息的弱化问题。

## 2.4 多损失训练

为了全面提升模型在讽刺识别任务中的综合性能, 本文在整体架构中引入多任务联合学习机制, 并设计多损失 (Multi-Loss) 训练策略, 旨在实现整体讽刺判别与类型细粒度区分的协同优化。具体而言, 讽刺识别被划分为两个层级: 首先, 二分类任务用于判定输入内容是否包含讽刺成分; 其次, 七分类任务在此基础上进一步区分

具体讽刺类型, 其中非讽刺样本作为独立类别纳入七分类体系, 确保标签体系的一致性与任务设计的统一性。

基于上述任务划分, 本文构建了联合优化目标函数:

$$L_{total} = \lambda_1 L_{binary} + \lambda_2 L_{type}$$

其中,  $L_{binary}$  表示讽刺二分类任务的交叉熵损失,  $L_{type}$  为七分类任务的交叉熵损失,  $\lambda_1$  和  $\lambda_2$  为平衡权重, 调控两类任务对整体优化过程的贡献比例。需要指出, 七分类任务标签统一覆盖所有样本, 其中非讽刺内容被视作独立类别, 避免了任务切换与样本筛选过程中的信息不完整或标签混淆问题。

通过多损失联合训练策略, 模型能够同时学习跨模态语义表达中的整体讽刺判别能力与类型层面的细粒度表达, 进一步增强了系统在复杂讽刺理解任务中的综合性能与鲁棒性。

## 3 实验

### 3.1 数据集

中文多模态讽刺计算数据集数据来源于国内著名的脱口秀节目《吐槽大会》。该数据集通过时间戳将视频内容切分为多个片段, 每个视频片段均进行了详细的标注, 包括文本内容、对应视频片段的起始结束时间、是否含有讽刺以及讽刺类别, 其类别分布情况如表 1 所示。

表 1 数据集类别分布情况

| 类别   | 数量    |
|------|-------|
| 不含讽刺 | 7633  |
| 嘲笑   | 1965  |
| 反语   | 55    |
| 讽刺文学 | 140   |
| 轻描淡写 | 18    |
| 夸大   | 28    |
| 反问   | 288   |
| 总数   | 10127 |

### 3.2 超参数设置

实验在 V100-SXM2-32GB 上进行, 其具体超

参数设置如表 2 所示。

表 2 超参数设置

| 参数名           | 设置值   |
|---------------|-------|
| Batch size    | 32    |
| Learning rate | 1e-5  |
| Optimizer     | AdamW |
| Hidden size   | 768   |
| Early stop    | 3     |
| Dropout       | 0.3   |
| Weight ratio  | 1:1   |

### 3.3 评价指标

本研究采用赛事官方统一指标体系进行性能评估, 具体包括: 子任务一, 多模态讽刺识别(二分类), 评价指标为 Macro F1 值; 子任务二, 多模态讽刺类型识别(七分类), 同样采用 Macro F1 值。最终成绩计算方式为:  $Score = 0.5 \times T_1.F1 + 0.5 \times T_2.F1$ , 其中  $T_1$  和  $T_2$  分别表示两项子任务, F1 为对应的 Macro F1 值。

### 3.4 结果

本研究方法在本次多模态讽刺识别任务中取得了较为优异的综合性能。具体排名结果如表 3 所示, 本文提出的模型 HIT416 团队最终得分为 0.5107, 位列全部参赛队伍中的第 5 名, 展现出良好的多模态讽刺理解与综合表达能力。

表 3 各队伍最终得分排名结果

| 队伍名    | 得分     | 排名 |
|--------|--------|----|
| 言传身教   | 0.5989 | 1  |
| 所以然    | 0.5964 | 2  |
| 我是大明星队 | 0.5487 | 3  |
| 天山棱镜   | 0.5389 | 4  |
| HIT416 | 0.5107 | 5  |
| 梯度消失   | 0.5060 | 6  |

## 4 结论

本文针对中文多模态讽刺计算任务, 提出了一种融合预训练模型与 Transformer 架构的多模态讽刺识别方法。通过引入 RoBERTa 与 HuBERT 双

模态特征编码、自注意力与交叉注意力融合机制以及多任务多损失优化策略, 系统性提升了模型对讽刺表达中的隐含语义、语调特征与复杂语境的综合理解能力。实验结果表明, 所提方法在官方赛事中获得较好排名, 展现出良好的多模态讽刺理解效果与应用潜力。后续工作将进一步探索多模态时序动态建模、跨语言讽刺表达分析及大模型在讽刺计算中的应用, 持续推动多模态讽刺理解技术的发展。

## 参考文献

- [1] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.
- [2] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
- [3] Hsu W N, Bolte B, Tsai Y H H, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM transactions on audio, speech, and language processing, 2021, 29: 3451-3460.
- [4] Zadeh A, Chen M, Poria S, et al. Tensor fusion network for multimodal sentiment analysis[J]. arXiv preprint arXiv:1707.07250, 2017.
- [5] Zadeh A, Liang P P, Mazumder N, et al. Memory fusion network for multi-view sequential learning[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).
- [6] Liu Z, Shen Y, Lakshminarasimhan V B, et al. Efficient low-rank multimodal fusion with modality-specific factors[J]. arXiv preprint arXiv:1806.00064, 2018.
- [7] Tsai Y H H, Bai S, Liang P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]//Proceedings of the conference. Association for computational linguistics. Meeting. 2019, 2019: 6558.

# 面向中文多模态讽刺检测的分层多视角融合方法

郭灏<sup>1</sup> 耿治发<sup>1</sup> 皇苏斌<sup>1</sup>

(1. 安徽工程大学 计算机与信息学院, 安徽 芜湖 241000)

**摘要:** 讽刺识别是自然语言理解中的一项具有挑战性的任务,尤其在中文语境下,由于语言表达的复杂性和语音语调的多变性,讽刺往往难以被准确捕捉。为此,本文提出一种面向中文多模态讽刺检测的分层多视角融合方法,融合文本与音频两种模态信息。模型首先利用预训练的 RoBERTa 和 HuBERT 模型提取各自模态的原始特征,随后通过跨模态编码模块(CME)建模两种模态之间的深层语义交互。接着,将原始特征与融合特征进行拼接,并引入 Transformer 编码器实现多视角表示建模。最终,通过层次化融合网络完成讽刺分类。在中文多模态讽刺数据集上的实验证明,该方法在多个评估指标上均优于现有强基线模型,验证了分层多视角融合策略的有效性。

**关键词:** 多模态讽刺检测; 多视角融合; 跨模态交互

中图分类号: TP391

文献标识码: A

## Hierarchical Multi-Perspective Fusion for Chinese Multimodal Sarcasm Detection

Hao Guo<sup>1</sup>, Zhifa Geng<sup>1</sup>, Subin Huang<sup>1</sup>

(1. School of Computer and Information, wuhu, anhui 24100, China)

**Abstract :** Sarcasm detection is a challenging task in natural language understanding, especially in the context of Chinese, where the complexity of linguistic expressions and the variability of prosody make sarcasm difficult to identify. To address this, we propose a **hierarchical multi-perspective fusion method for Chinese multimodal sarcasm detection**, which integrates both textual and audio modalities. Our model first utilizes pre-trained RoBERTa and HuBERT models to extract unimodal features from text and audio respectively. Then, a cross-modal encoding module (CME) is applied to capture deep semantic interactions between modalities. The original features and the fused features are concatenated and passed through Transformer encoders to model **multi-perspective representations**. Finally, a hierarchical fusion network is used to perform sarcasm classification. Experiments on a Chinese multimodal sarcasm dataset demonstrate that our method outperforms several strong unimodal and multimodal baselines across multiple evaluation metrics, validating the effectiveness of the proposed hierarchical multi-perspective fusion strategy.

**Key words:** Multimodal Sarcasm Detection; Multi-Perspective Fusion; Cross-Modal Interaction

### 1 引言

多模态融合技术近年来在情感识别、语义理解与人机交互等任务中取得了广泛关注。现有方法主要采用早期特征拼接、中期注意力机制交互或晚期决策层融合等策略,以整合来自不同模态

的信息。然而，这些方法普遍存在模态信息损失、交互深度不足以及单模态判别性特征被弱化等问题。例如，简单的特征拼接难以建模模态间的细粒度关联，过度依赖融合表示则可能掩盖单模态中的独有结构与判别信息。此外，在多任务学习场景中（如同时进行情感分类与细粒度标签预测），如何构建灵活高效的分支结构以适配不同任务目标、提升模型的泛化能力，仍面临诸多挑战。

为应对上述问题，本文提出一种基于跨模态交互与多分支聚合机制的多模态融合模型。该模型以预训练语言模型 RoBERTa 与音频模型 HuBERT 为基础，作为文本与音频的特征提取器，并通过参数冻结保留其原始知识。在此基础上，引入多层跨模态交互模块（Cross-Modality Encoding, CME），以实现文本与音频间的深层次信息交换；随后利用 Transformer 编码器进一步建模融合特征的上下文依赖关系。为了充分保留模态特异性并提升判别能力，模型分别设计了文本分支、音频分支与融合分支，采用多结果聚合策略综合各模态贡献。此外，通过多任务输出结构，模型同时适配情绪状态的二分类与七分类任务，增强其在多目标场景下的适应性与鲁棒性。

实验结果表明，所提出的模型在多模态情绪识别任务中表现优越，能够有效整合不同模态信息，为多模态联合建模提供了一种具有可行性与推广性的解决方案。

## 2 方法

本文提出的多模态融合模型旨在联合处理文本与音频数据，针对两类任务（Task1: 2 分类；Task2: 7 分类）进行预测。模型采用“单模态特征提取 - 跨模态交互融合 - 多分支预测聚合”的三级架构，通过预训练模型提取基础特征，结合跨模态交互层实现模态间信息交换，并通过多分支输出与融合策略提升预测性能。整体框架如图 1 所示。

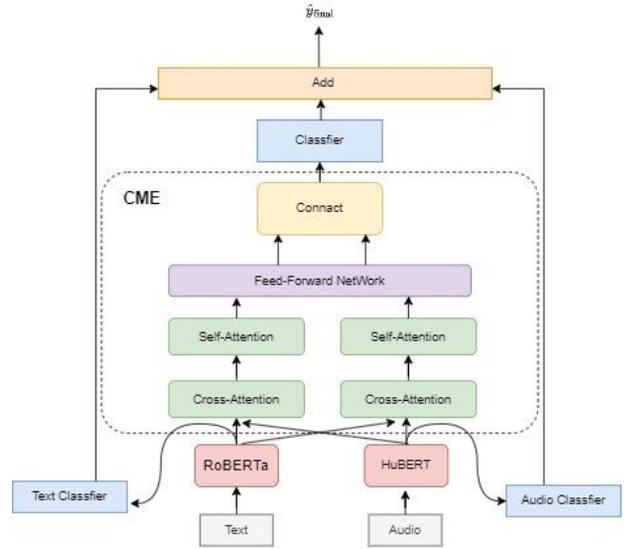


图 1-模型结构图

### 2.1 特征提取

模型选用预训练模型作为固定特征提取器，分别对文本和音频输入进行处理，以获取高质量的基础特征。

在文本模态方面，选用中文 RoBERTa 模型（“hf1/chinese-roberta-wwm-ext”）作为特征提取器。该模型在大规模中文语料上预训练，能够有效捕捉句法结构与上下文语义。输入序列经由预训练的 Roberta 模型编码，得到隐藏状态序列与池化输出，其映射关系可表示为：

$$H^{(T)} = \text{Roberta}(X^{(T)}) \in R^{L \times 768}$$

$$f^{(T)} = \text{PoolerOutput}(H^{(T)}) \in R^{768}$$

其中  $X^{(T)}$  是输入的文本序列。

在音频模态方面，选用中文 HuBERT 模型（“TencentGameMate/chinese-hubert-base”），该模型专为中文语音任务设计，能够有效提取语音信号中的韵律、音素及语义特征。

$$H^{(A)} = \text{HuBERT}(X^{(A)}) \in R^{T \times 768}$$

由于 HuBERT 输出包含大量填充帧，为确保特征聚合的有效性，引入注意力掩码机制，计算有效帧数  $M$ ，依据注意力权重非零值确定：

$$M = \sum_{i=1}^T I\left(\sum_h \alpha_{h,i} > 0\right)$$

其中， $\alpha_{h,i}$  表示第  $h$  层注意力在第  $i$  帧的位置， $I(\cdot)$  为指示函数。最终语音模态特征通过有效帧均值聚合获得：

$$f^{(A)} = \frac{1}{M} \sum_{i=1}^M H_i^{(A)} \in R^{768}$$

其中  $H_i^{(A)}$  是语音输入为原始特征序列。

### 2.2 CME 模块

CME (Cross-Modal Encoder) 模块用于对语言模态和音频模态进行双向交叉融合与自身信息建模。其结构包括三个主要部分：(1) 跨模态注意力；(2) 自注意力建模；(3) 前馈输出。

### 2.2.1 跨模态注意力

语言模态以音频为上下文进行注意力计算，反之亦然。这样可以使两个模态的信息彼此对齐和融合：

- 文本对音频做注意力：  
 $\text{Lang}_{\text{att}} = \text{CrossAttn}(\text{Lang}, \text{Audio})$
- 音频对文本做注意力：  
 $\text{Audio}_{\text{att}} = \text{CrossAttn}(\text{Audio}, \text{Lang})$

### 2.2.2 自注意力建模

在跨模态融合后，分别对语言和音频进行模态内部的自注意力建模，捕捉局部和全局依赖关系：

$$\begin{aligned}\text{Lang}_{\text{self}} &= \text{SelfAttn}(\text{Lang}_{\text{att}}) \\ \text{Audio}_{\text{self}} &= \text{SelfAttn}(\text{Audio}_{\text{att}})\end{aligned}$$

### 2.2.3 前馈输出

最后，将注意力结果送入全连接层（带有激活函数、残差和层归一化），生成最终输出表示：

$$\begin{aligned}\text{Lang}_{\text{out}} &= \text{FFN}(\text{Lang}_{\text{self}}) \\ \text{Audio}_{\text{out}} &= \text{FFN}(\text{Audio}_{\text{self}})\end{aligned}$$

## 2.3 特征融

本节介绍所提出的多模态特征融合策略，该方法在标准交叉注意力结构基础上，进一步引入残差特征拼接与 Transformer 级上下文建模机制，旨在增强跨模态表示能力与语义融合鲁棒性。

传统的跨模态建模方法往往仅依赖局部的注意力交互，容易导致信息衰减或语义偏移问题。在我们的方法中，我们不仅保留原始模态特征（RoBERTa 与 HuBERT 提取），还引入了 CME 模块后的输出，通过拼接原始与交互特征作为融合输入，再借助 Transformer 编码器建模其上下文依赖关系。该策略有效提高了多模态语义对齐精度和全局语义建模能力。

### 2.3.1 特征拼接与构建

首先，从文本模态与语音模态中分别提取原始特征和交互增强特征，设经过 CME 层处理的文本与音频特征为：

$$\widehat{H}^{(T)}, \widehat{H}^{(A)} \in R^{L \times d}, R^{T \times d}$$

将上述特征与原始特征在最后一维拼接（表示残差融合）：

$$\begin{aligned}Z^{(T)} &= \left[ H^{(T)} \mid \widehat{H}^{(T)} \right] \in R^{L \times 2d} \\ Z^{(A)} &= \left[ H^{(A)} \mid \widehat{H}^{(A)} \right] \in R^{T \times 2d}\end{aligned}$$

上述拼接方式能够同时保留原始模态的信息分布和跨模态交互结构。

### 2.3.2 添加全局标识符

在序列建模前，为提取全局模态语义表示，我们为每个模态序列引入一个可训练的全局表示标识符 [CLS]：

$$\begin{aligned}Z_{\text{in}}^{(T)} &= [e_{\text{CLS}}^{(T)}; Z^{(T)}] \in R^{(L+1) \times 2d} \\ Z_{\text{in}}^{(A)} &= [e_{\text{CLS}}^{(A)}; Z^{(A)}] \in R^{(T+1) \times 2d}\end{aligned}$$

其中， $e_{\text{CLS}}^{(T)}, e_{\text{CLS}}^{(A)} \in R^{1 \times 2d}$  是可训练的嵌入参数，用于捕捉序列级的上下文语义表示。

### 2.3.3 Transformer 编码器增强

接下来，输入拼接后的特征序列至两个独立的 Transformer 编码器模块中进行建模。Transformer 采用标准结构，包括多头自注意力机制与前馈网络，能有效建模长距离依赖。

$$\begin{aligned}\widehat{Z}^{(T)} &= \text{Transformer}_{\text{text}}(Z_{\text{in}}^{(T)}) \in R^{(L+1) \times 2d} \\ \widehat{Z}^{(A)} &= \text{Transformer}_{\text{audio}}(Z_{\text{in}}^{(A)}) \in R^{(T+1) \times 2d}\end{aligned}$$

编码器使用残差连接、层归一化与 Dropout 策略，确保信息流稳定传播并具备较强泛化能力。

### 2.3.4 多模态融合

经过 Transformer 建模后，从每个模态输出序列中提取 [CLS] 位（第 0 位）输出作为模态全局语义向量：

$$f_{\text{text}} = \widehat{Z}_0^{(T)} \in R^{2d}, f_{\text{audio}} = \widehat{Z}_0^{(A)} \in R^{2d}$$

最终将两者拼接为多模态融合表示：

$$f_{\text{fused}} = [f_{\text{text}} \parallel f_{\text{audio}}] \in R$$

## 2.4 分类

把融合向量随后输入多层感知机 (MLP) 结构，完成最终任务预测。以二分类任务为例，计算如下：

$$\widehat{y} = \text{MLP} f_{\text{fused}} = W_2 \cdot \sigma(W_1 \cdot f_{\text{fused}} + b_1) + b_2$$

其中  $W_1 \in R^{768 \times 4d}, W_2 \in R^{2 \times 768}, \sigma(\cdot)$  表示 ReLU 激活函数， $\widehat{y} \in R^2$  为输出概率分布。

在本模型中，我们为文本模态与音频模态分别设置了独立的分类头，包括任务一（2 类）与任务二（7 类）的输出层，形式如下：

文本模态分类头：

$$\widehat{y}_T = \text{Linear}_T(f^{(T)})$$

音频模态分类头：

$$\widehat{y}_A = \text{Linear}_A(f^{(A)})$$

其中  $f^{(T)}$  和  $f^{(A)}$  是文本和音频的池化特征。

这些独立的模态输出层在多模态学习中具有以下两个重要作用：(1) 提供模态特异性监督信号，独立的输出头可以强制各模态学习对任务有判别力的特征，避免仅依赖融合表示而退化为“主模态驱动”。(2) 辅助融合优化，增强鲁棒性：在多模态融合时，如果某一模态特征缺失或噪声较多，原始分类头可以提供一定的容错机制。最终输出直接相加的策略整合三个部分：

$$\widehat{y}_{\text{final}} = \widehat{y}_T + \widehat{y}_A + \widehat{y}_{\text{fused}}$$

## 2.5 训练策略

### 2.5.1 损失函数设计

本任务为多任务学习场景，包含二分类的情感识别任务（Task 1）和七分类的具体情绪识别任务（Task 2）。为此，模型包含两个独立的输出头，并分别使用交叉熵损失（Cross-Entropy Loss）进行监督。为缓解类别不平衡问题，我们对每一任务的类别分布统计样本频次，基于逆频率（inverse frequency）计算类别权重，并将其引入损失函数中，以提升对少数类的识别能力。

对于 epoch 初期（前 5 轮），我们仅优化情绪分类（Task 2）相关损失，稳定情绪识别能力。随后，完整训练阶段使用如下联合损失函数：

$$L_{\text{total}} = L_{\text{task1}} + \lambda \cdot L_{\text{task2}}$$

其中， $\lambda = 5.0$  为经验设置的权重因子，控制两个任务损失的相对贡献。

### 2.5.2 优化器与学习率调

优化器选择 *Adam*，并设置初始学习率为  $3 \times 10^{-5}$ 。训练过程中，采用 *StepLR* 学习率调度器，在每隔 3 个 *epoch* 后将学习率衰减为原来的 0.1，以实现平稳收敛与防止过拟合。

## 3 实验

### 3.1 数据集

本研究使用 [CCAC2025 情绪识别比赛数据集]，该数据集包含语音、文本与视频多模态信息，以及两个层次的情绪标签：

**Task 1（多模态讽刺识别）：**对于给定的多模态（文本数据标和对应的视频片段与音频片段），参赛模型需判断当前文本数据是否为讽刺语句。输出为 2 种标签：“1”表示该句子为讽刺语句，“0”表示该句子为非讽刺语句。

**Task 2（多模态讽刺类别识别）：**子任务二，多模态讽刺类别识别：对于给定的多媒体数据（文本数据和对应的视频片段与音频片段），参赛模型需判断当前评论数据属于哪个讽刺类别。

训练集与测试集按照官方划分进行，不另作数据增强或扩展。

### 3.2 超参数设置

| 超参数名称 | 值    | 说明               |
|-------|------|------------------|
| 学习率   | 3e-5 | 初始学习率，Adam 优化器使用 |
| 优化器   | Adam | 适用于              |

|                   |                       |                        |
|-------------------|-----------------------|------------------------|
|                   |                       | Transformer 类模型        |
| 学习率衰减策略           | 3                     | 每 3 个 epoch 学习率衰减一次    |
| 学习率衰减步长           | 3                     | 每 3 epoch 衰减一次         |
| 学习率衰减系数           | 0.1                   | 学习率乘以该系数               |
| 批大小               | 32                    | 训练与测试一致                |
| 最大训练轮数            | 20                    | 最多训练 20 轮              |
| 损失函数              | Weighted CrossEntropy | 针对两个任务分别加权处理           |
| 多任务损失权重 $\lambda$ | 5.0                   | Task2 权重提升，增强细粒度情感表达学习 |
| 冻结策略（前 5 轮）       | 仅训练 Task2 分类头         | 仅训练 Task2              |
| 随机种子              | 42                    | 固定所有模块，确保实验可复现         |

## 4 结论

本文提出了一种面向中文多模态讽刺检测的分层多视角融合方法，综合利用文本与音频两种模态信息，通过预训练特征提取、跨模态交互、Transformer 编码与多任务学习机制，实现了讽刺语句的精确识别与分类。该方法有效缓解了传统多模态融合过程中的信息损失与模态不平衡问题，提升了模型在多任务情绪识别中的表现。实验结果表明，所提模型在多个评估指标上均优于现有强基线，具备良好的鲁棒性与推广性。未来工作可进一步引入视觉模态信息，拓展至视频场景中的多模态讽刺识别任务，增强模型的应用广度与实际适应能力。

# CCAC2025赛道五IIE06队伍技术报告

赵语林

孙业奇

李雨晴

中国科学院信息工程研究所 中国科学院信息工程研究所 中国科学院信息工程研究所

zhaoyulin@iie.ac.cn

sunyeqi@iie.ac.cn

liyuqing@iie.ac.cn

## 摘要

比拟是中文中一种重要的修辞手法，通过将本体（描述对象）与喻体（比拟对象）建立相似性联系（喻底），以生动形象的方式传递情感和语义信息。本技术报告针对CCAC2025赛道五——细粒度比拟句情绪分析评测任务，系统阐述了本参赛队伍的解决方案，并探讨了实验中观察到的现象。针对比拟句情绪分析、细粒度本体、喻体特征抽取和喻底推理三个子任务，本参赛队伍设计了一套融合多任务学习、提示增强、数据增强、上下文学习和模型集成的技术方案，最终在评测中任务一宏平均F1达0.5731，任务二F1达0.7186，任务三准确率达0.9935。

## 1 任务目标

### 1.1 任务一：比拟句情绪分析

目标：给定一个比拟句，判别该句子的情绪类别。

输出：8类情绪标签，0-7分别代表无情绪、乐、好、怒、哀、惧、恶、惊。

评价指标：Macro-F1

### 1.2 任务二：细粒度本体、喻体特征抽取

目标：抽取句子中所有本体（描述对象）及喻体（比拟对象）片段对。

输出：假设一句比拟句中包含 $n$ 组<本体-喻体>对，输出<本体1, 喻体1>, .....<本体 $n$ , 喻体 $n$ >。

评价指标：F1（模糊匹配）

### 1.3 任务三：<本体-喻体>对喻底判别

目标：根据给定比拟句信息及指定<本体-喻体>对，从4个选项中单选题选出该对应的喻底。

输出：A/B/C/D 之一，表示正确的喻底推断。

评价指标：ACC

## 2 技术方案

针对任务一，我们主要围绕多任务学习、提示增强、数据增强、上下文学习和模型集成策略进行了探索和改进。针对任务二，我们主要采用了多任务学习、上下文学习和模型集成策略，特别是针对基于模糊匹配的实体抽取的任务特点设计了一套启发式的集成方案。针对任务三，我们仅采取了模型集成策略。具体实现方案如下文所述：

### 2.1 多任务学习

在初步实验中，我们发现比拟句中的本体-喻体对与情绪表达存在强相关性。例如，当本体被比作“猛兽”时，往往表达“怒”或“惧”的情绪；而比作“春风”时，则多表达“乐”或“好”的情绪。这种语言学观察促使我们探索多任务学习，让模型同时学习情绪分类和本体-喻体抽取两个任务。实验结果如表1所示。

## 2.2 提示增强

原始的单字类别标签“乐、好、怒、哀、惧、恶、惊”可能不利于模型理解。针对“好”“乐”等单字类别标签从语义上难以区分的问题，我们设计了两种标签改进方案：

- 将单字类别标签改为词语，如“乐”→“欢乐”，“好”→“赞美”
- 为每个类别标签附加一段文字解释

此外，受(Li et al., 2023)启发，我们还尝试了加入EmotionPrompt，但效果不理想，具体原因有待进一步分析。

## 2.3 数据增强和两阶段训练

原始数据集中情绪类别分布极不均衡，这种不平衡将严重制约模型在少数类上的表现。因此，我们采用了数据增强策略来平衡不同类别的样本数，提高模型在少数类上的表现。具体来说，我们采用采用qwen-plus和qwen-max API来生成数据，并设计了一种两阶段数据增强方案：

- 第一阶段：以原始数据为种子，采用4-shot方式为除“好”以外的每个情绪类别生成500条比拟句。我们随机抽查了部分样本以检查生成数据的质量，并对于平均生成质量较差的类别“惊”进行了手动过滤。
- 第二阶段：以原始数据和第一阶段生成数据为种子，采用相同方式将每个类别补齐到约1500条左右并再次进行了质量抽查。

此外，为了解决生成数据质量相对较低、与原始数据集标注可能不对齐等问题，我们采取了两阶段训练策略，即第一阶段在增强数据上训练，第二阶段再采用原始数据进一步微调。

## 2.4 上下文学习

以往的研究表明，上下文学习能够提升大语言模型在各种任务上的性能。受此启发，我们设计了基于检索的上下文学习方案。具体来说，我们采用训练集作为示例库，利用BGE-m3(Multi-Granularity, 2024)为每个样本生成密集表征，对于测试样本，计算其与示例库中所有样本的相似度，并选取最相似的k个样本作为上下文学习的示例，其中最佳k值通过实验确定。

## 2.5 模型集成

我们在任务一、任务二和任务三上都采用了模型集成策略。对于任务一，虽然单独采取上述每一项策略都对性能有所提升，但不同策略组合起来却未必有效。因此，我们对不同策略的组合进行了搜索，选取其中性能最好的几个模型进行集成。任务一和任务三直接采用投票决定最终答案，若出现平票则优先选择验证集性能最好的模型的结果。特别地，对于任务二，由于实体抽取存在不同字符串可能指向同一实体的问题，我们采用了启发式的集成方式。如算法 2 所示，在加权结果时同时考虑支持度（多少个模型支持该回复对）和一致性（该回复对与其他回复对的整体相似程度）。其中，一致性的计算如算法 1 所示。

---

### Algorithm 1 一致性指标计算

---

**Require:** 回复对  $p_1 = (subject_1, object_1)$ ,  $p_2 = (subject_2, object_2)$

**Require:** 实体相似度阈值  $\theta_{entity}$ , 对相似度阈值  $\theta_{pair}$

**Ensure:** 布尔值 indicating 相似性

- 1:  $sim_{subject} \leftarrow \text{CalcIoU}(subject_1, subject_2)$
  - 2:  $sim_{object} \leftarrow \text{CalcIoU}(object_1, object_2)$
  - 3: **if**  $\theta_{pair} < 1.0$  **then**
  - 4:     **return**  $\min(sim_{subject}, sim_{object}) \geq \theta_{pair}$  {使用最小相似度}
  - 5: **else**
  - 6:     **return**  $\frac{sim_{subject} + sim_{object}}{2} \geq \theta_{pair}$  {使用平均相似度}
  - 7: **end if**
-

---

**Algorithm 2** 加权投票集成算法

---

**Require:** 多个模型的预测结果  $\mathcal{M} = \{M_1, M_2, \dots, M_k\}$ , 其中  $M_i$  包含模型  $i$  预测的 本体-喻体对集合

**Require:** 相似度阈值  $\theta_{sim}$ , 最终筛选阈值  $\theta_{final}$

**Ensure:** 集成后的 本体-喻体对集合  $\mathcal{P}_{final}$

```
1: 步骤1: 收集所有唯一回复对
2:  $\mathcal{P}_{all} \leftarrow \bigcup_{i=1}^k M_i$  {合并所有模型预测}
3:  $\mathcal{U}, \mathcal{C} \leftarrow \text{FindUniquePairs}(\mathcal{P}_{all}, \theta_{sim})$  {去重并统计频次}
4:
5: 步骤2: 计算支持度
6: for  $j = 1$  to  $|\mathcal{U}|$  do
7:    $support[j] \leftarrow 0$ 
8:   for  $i = 1$  to  $k$  do
9:     if  $\exists p \in M_i$  such that  $\text{IsSimilar}(p, \mathcal{U}[j], \theta_{sim})$  then
10:       $support[j] \leftarrow support[j] + 1$ 
11:     end if
12:   end for
13: end for
14:
15: 步骤3: 计算一致性分数
16: for  $j = 1$  to  $|\mathcal{U}|$  do
17:    $similarity_{sum} \leftarrow 0, count \leftarrow 0$ 
18:   for  $l = 1$  to  $|\mathcal{U}|, l \neq j$  do
19:      $sim_{subject} \leftarrow \text{CalcIoU}(\mathcal{U}[j].subject, \mathcal{U}[l].subject)$ 
20:      $sim_{object} \leftarrow \text{CalcIoU}(\mathcal{U}[j].object, \mathcal{U}[l].object)$ 
21:      $similarity_{sum} \leftarrow similarity_{sum} + \frac{sim_{subject} + sim_{object}}{2}$ 
22:      $count \leftarrow count + 1$ 
23:   end for
24:    $consistency[j] \leftarrow \frac{similarity_{sum}}{count}$  if  $count > 0$  else 0
25: end for
26:
27: 步骤4: 计算综合得分
28: for  $j = 1$  to  $|\mathcal{U}|$  do
29:    $score[j] \leftarrow support[j] + 0.5 \times consistency[j]$ 
30: end for
31:
32: 步骤5: 根据阈值筛选
33:  $\mathcal{P}_{final} \leftarrow \{\mathcal{U}[j] : score[j] \geq \theta_{final}\}$ 
34: return  $\mathcal{P}_{final}$ 
```

---

### 3 实现细节

任务一采用Qwen2.5-14B-Instruct 作为基座模型, 任务二采用Qwen2.5-14B-Instruct、GLM-4-9B-0414、GLM-4-32B-0414 作为基座模型, 任务三采用GLM-4-32B-0414、InternLM3-8b、Qwen3-32B、Gemma3-27b-it 作为基座模型。基于ms-swift(Zhao et al., 2024)进行LoRA微调, global batch size为16, lr为1e-4, rank为128, alpha为256。训练完成后, 按章节 2.5 介绍的算法进行集成。

### 4 实验结果

表1展示了多任务学习对任务一和任务二的性能影响。结果表明, 多任务学习对任务一有较大的增益, 对于任务二也有一定提升, 证实了情绪识别与本体-喻体识别的协同作用。

|     | 任务一    |        |        |        |        |        |        |        |        | 任务二    |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
|     | 无情绪    | 乐      | 好      | 怒      | 哀      | 惧      | 恶      | 惊      | Avg    |        |
| 单任务 | 0.0000 | 0.8211 | 0.8893 | 0.2500 | 0.7231 | 0.5000 | 0.5634 | 0.5217 | 0.5336 | 0.6885 |
| 多任务 | 0.4516 | 0.8078 | 0.8938 | 0.3333 | 0.7500 | 0.4000 | 0.6216 | 0.4348 | 5866   | 0.7073 |

Table 1: 多任务学习对任务一和任务二的性能影响

|          | 无情绪    | 乐      | 好      | 怒      | 哀      | 惧      | 恶      | 惊      | Avg    |
|----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 单任务 -    | 0.0000 | 0.8211 | 0.8893 | 0.2500 | 0.7231 | 0.5000 | 0.5634 | 0.5217 | 0.5336 |
| 单任务 词语解释 | 0.1333 | 0.8208 | 0.8865 | 0.5000 | 0.7234 | 0.5641 | 0.4211 | 0.5833 | 0.5791 |
| 多任务 -    | 0.4516 | 0.8078 | 0.8938 | 0.3333 | 0.7500 | 0.4000 | 0.6216 | 0.4348 | 0.5866 |
| 多任务 词语解释 | 0.1840 | 0.8089 | 0.8403 | 0.5000 | 0.6301 | 0.4583 | 0.5128 | 0.5417 | 0.5595 |
|          | 0.3704 | 0.8082 | 0.8908 | 0.4615 | 0.7230 | 0.4225 | 0.5957 | 0.4314 | 0.5879 |

Table 2: 不同的标签改进方案对单任务训练和多任务训练模型在任务一上的性能影响

表2展示了不同的标签改进方案对单任务训练和多任务训练模型在任务一上的性能影响。观察实验结果主要能得到两个发现：1. 为每个标签附加一段文字解释总体效果更优；2. 改进标签后多任务学习对任务一不再有帮助，甚至可能有反作用。

表3显示了数据增强和两阶段训练对任务一的性能影响。观察实验结果可以发现，如果第二阶段只在任务一上训练，则会破坏第一阶段数据增强训练的效果，而采用多任务训练则性能进一步提升。

表4展示了示例数量对于多任务训练模型在任务一上的性能影响。我们首先在0 shot设置下进行了探索，结果表明2-shot设置效果最佳，过多示例可能导致注意力分散。因此在后续实验中采用了2 shot设置。

表5展示了2 shot设置下多任务学习、标签改进和两阶段训练策略的组合对任务一的性能影响。实验结果表明，虽然单独采取上述每一项策略都对性能有提升，但组合策略却未必效果好。具体来说，标签改进策略会削弱数据增强的效果，也会损害多任务学习的效果；在单任务学习的情况下，将标签改为词语比加入解释效果好，而在多任务学习的情况下则相反。受限于比赛时间，我们未对这些现象展开进一步探索，期待未来的工作继续完善。

表6展示了2 shot和数据增强设置下EmotionPrompt与不同标签改进策略组合对任务一的性能影响。从实验结果可以观察到，EmotionPrompt与标签改进也存在冲突。

表7展示了本参赛队伍在验证集和测试集上的最终性能。

## 5 总结与展望

本报告详细介绍了我们在CCAC2025比赛评测赛道五的技术方案。通过多任务学习、数据增强、上下文学习、提示增强和模型集成的有机结合，我们有效提升了模型在中文比拟句情绪分析任务上的性能。在实验过程中，我们发现了一系列有趣的现象，特别是本体-喻体信息与情绪表达的强相关性，值得进一步的研究与探索。

## 参考文献

- Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.
- Multi-Linguality Multi-Functionality Multi-Granularity. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.
- Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024. Swift:a scalable lightweight infrastructure for fine-tuning.

|         | 无情绪    | 乐      | 好      | 怒      | 哀      | 惧      | 恶      | 惊      | Avg    |
|---------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 原始数据    | 0.0000 | 0.8211 | 0.8893 | 0.2500 | 0.7231 | 0.5000 | 0.5634 | 0.5217 | 0.5336 |
| 增强数据    | 0.2857 | 0.8146 | 0.8796 | 0.3636 | 0.6552 | 0.4706 | 0.5789 | 0.6207 | 0.5836 |
| 两阶段+单任务 | 0.0000 | 0.8178 | 0.8811 | 0.3333 | 0.7077 | 0.5000 | 0.5075 | 0.5714 | 0.5399 |
| 两阶段+多任务 | 0.2759 | 0.8305 | 0.8975 | 0.5714 | 0.7206 | 0.5161 | 0.5882 | 0.4762 | 0.6096 |

Table 3: 数据增强和两阶段训练对任务一的性能影响

|      |        | 无情绪    | 乐      | 好      | 怒      | 哀      | 惧      | 恶      | 惊      | Avg    |
|------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 原始数据 | 0 shot | 0.4516 | 0.8078 | 0.8938 | 0.3333 | 0.7500 | 0.4000 | 0.6216 | 0.4348 | 0.5866 |
|      | 2 shot | 0.5000 | 0.8101 | 0.8953 | 0.4444 | 0.6906 | 0.4828 | 0.6571 | 0.5833 | 0.6330 |
|      | 3 shot | 0.4545 | 0.8100 | 0.8940 | 0.2500 | 0.6897 | 0.4828 | 0.6000 | 0.5385 | 0.5899 |
| 两阶段  | 0 shot | 0.2759 | 0.8305 | 0.8975 | 0.5714 | 0.7206 | 0.5161 | 0.5882 | 0.4762 | 0.6096 |
|      | 2 shot | 0.6000 | 0.8030 | 0.8980 | 0.4444 | 0.7092 | 0.4516 | 0.6316 | 0.6087 | 0.6433 |

Table 4: 示例数量对于多任务训练模型在任务一上的性能影响

|     |        | 无情绪    | 乐      | 好      | 怒      | 哀      | 惧      | 恶      | 惊      | Avg    |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 单任务 | 词语     | 0.4000 | 0.8364 | 0.8959 | 0.3333 | 0.7552 | 0.6364 | 0.6389 | 0.6667 | 0.6454 |
|     | 两阶段+词语 | 0.3636 | 0.8449 | 0.9027 | 0.3333 | 0.7536 | 0.6364 | 0.5641 | 0.6667 | 0.6332 |
|     | 解释     | 0.3750 | 0.8289 | 0.8881 | 0.5000 | 0.7442 | 0.4706 | 0.6571 | 0.5806 | 0.6306 |
|     | 两阶段+解释 | 0.4545 | 0.8305 | 0.8883 | 0.3333 | 0.7123 | 0.4706 | 0.5352 | 0.5333 | 0.5948 |
| 多任务 | 词语     | 0.2667 | 0.8214 | 0.8799 | 0.4000 | 0.6829 | 0.6364 | 0.5591 | 0.5455 | 0.5990 |
|     | 两阶段+词语 | 0.3529 | 0.8267 | 0.8985 | 0.2500 | 0.7194 | 0.5000 | 0.6133 | 0.6154 | 0.5970 |
|     | 解释     | 0.2500 | 0.8321 | 0.8959 | 0.4444 | 0.7133 | 0.5000 | 0.5672 | 0.6452 | 0.6060 |
|     | 两阶段+解释 | 0.3846 | 0.8111 | 0.8921 | 0.4444 | 0.7183 | 0.5714 | 0.6479 | 0.5926 | 0.6328 |

Table 5: 2 shot设置下多任务学习、提示增强和两阶段训练策略的组合对任务一的性能影响

|    |    | 无情绪    | 乐      | 好      | 怒      | 哀      | 惧      | 恶      | 惊      | Avg    |
|----|----|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| -  | -  | 0.3636 | 0.8018 | 0.8819 | 0.4000 | 0.7188 | 0.6316 | 0.5352 | 0.5000 | 0.6041 |
|    | EP | 0.3750 | 0.8195 | 0.8829 | 0.5000 | 0.7068 | 0.6667 | 0.5750 | 0.5333 | 0.6324 |
| 词语 | -  | 0.5385 | 0.8241 | 0.8866 | 0.3077 | 0.7188 | 0.5714 | 0.5135 | 0.4800 | 0.6051 |
|    | EP | 0.3429 | 0.7915 | 0.8734 | 0.3077 | 0.7097 | 0.5455 | 0.6053 | 0.5000 | 0.5845 |
| 解释 | -  | 0.3704 | 0.8413 | 0.9012 | 0.4000 | 0.6829 | 0.6667 | 0.5882 | 0.5517 | 0.6253 |
|    | EP | 0.2927 | 0.8077 | 0.8785 | 0.4615 | 0.7244 | 0.5455 | 0.5753 | 0.4375 | 0.5904 |

Table 6: 2 shot和数据增强设置下EmotionPrompt与不同提示增强组合对任务一的性能影响, EP代表加入EmotionPrompt

|     | 任务一 (Avg) | 任务二 (F1) | 任务三 (Acc) |
|-----|-----------|----------|-----------|
| 验证集 | 0.6956    | 0.7396   | 0.9980    |
| 测试集 | 0.5731    | 0.7186   | 0.9935    |

Table 7: 验证集和测试集上的最终性能

# CCAC2025-Task5 技术报告：面向细粒度比拟句情绪分析的上下文微调方法

张兴辰 刘艺程 杨成 周祐超 史树敏\*

(北京理工大学计算机学院, 北京 100081)

**摘要:** 该文介绍了我们在 CCAC2025 中细粒度比拟句情绪分析评测所采用的方法。我们选用 Qwen-14B-chat 大语言模型作为基座模型, 提出了上下文微调 (In-Context Fine-Tuning) 方法, 该方法借鉴上下文学习的思想构建了一个动态示例池辅助构建 prompt 模板, 并结合 LoRA 微调技术, 实现了模型对判断比拟句所蕴含的情绪类别、比拟句中本体及喻体的提取、结合比拟句子信息判断喻底这三个子任务的适应。最终我们的方法在任务 1 的 Macro-F1 分数为 0.5516, 任务 2 的 F1 分数为 0.6667, 任务 3 的准确率为 0.9895, 达到了较为理想的成绩, 印证了我们方法在细粒度比拟句情绪分析这一任务上具有一定的有效性。

**关键词** 大语言模型; 上下文学习; 指令微调

## CCAC2025-Task5 Technical Report: In-Context Fine-Tuning for Fine-Grained Simile Emotion Analysis

Xingchen Zhang, Yicheng Liu, Cheng Yang, Youchao Zhou, Shumin Shi\*

(School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

**Abstract :** This paper presents our approach to the fine-grained simile emotion analysis task in CCAC2025. We employed the Qwen-14B-chat large language model as the backbone and proposed an **In-Context Fine-Tuning** method. This approach integrates contextual learning principles to construct a dynamic example pool for prompt template optimization and combines LoRA fine-tuning technology. The method adaptively addresses three subtasks: identifying emotional categories in similes, extracting ontological and metaphorical pairs, and determining the metaphorical basis from contextual information. Our approach achieved promising results: a Macro-F1 score of **0.5516** for Task 1, an F1 score of **0.6667** for Task 2, and an accuracy of **0.9895** for Task 3, demonstrating its effectiveness in addressing the complexities of fine-grained simile emotion analysis.

**Key words:** Large Language Models; In-Context Learning; Instruction Tuning

## 0 引言

比拟作为汉语中常见的修辞手法，主要包含比喻、拟人和拟物三类细分类型。其核心机制在于借助本体（被描述主体）与喻体（比拟参照体）之间构建相似性关联（即喻底），通过这种关联生动形象地传递情感与语义信息。在文学创作、社交媒体传播等实际应用场景中，为强化文本的情绪表达效果，比拟句式被频繁使用，这对机器实现深层文本语义与情绪信息的挖掘提出了挑战。

本次评测针对该任务我们提出了上下文微调 (In-Context Fine-Tuning) 的方法，思想上借鉴了大模型上下文学习 [2] 的方法，通过利用训练集语料作为动态示例池，根据每个输入样例动态的选择相似的示例用来构建 prompt，以便于让大模型更好地捕捉到本体-喻体间的动态语义关联。同时为了提高模型的训练效率还采用了 LoRA [1] (Low-Rank Adaptation) 技术对大模型进行高效的参数微调，以便于让大模型更好地适应细粒度比拟句分析的任务。

## 1 方法

本次细粒度比拟句情绪分析评测任务分为三个子任务，分别为比拟句情绪分析、细粒度本体、喻体特征抽取以及<本体-喻体>对喻底判别，三个子任务均采用 Qwen-14B-chat 模型来作为基础模型并在此基础上进行训练微调。

### 1.1 模型结构与微调方式

我们选用 Qwen-14B-chat 模型 [3] 作为基座模型。在微调阶段，采用 PEFT 框架中的 LoRA 技术，以节省显存并提升训练效率。LoRA 设置如下：秩 (rank) 为 16, lora\_alpha 设为 32, 应用于 transformer 模块中的注意力投影层及前馈层 (包括 q\_proj、k\_proj、v\_proj、o\_proj、gate\_proj、up\_proj、down\_proj 等)。

### 1.2 动态示例池构建

借鉴了上下文学习 (In-Context Learning) 的思想，为了模型能够在训练时更好地捕捉到比

拟句中的语义关联，我们将训练集语料构建成为一个动态示例池，在构建 prompt 时，根据当前需要处理的输入从示例池中选择最相似的两个例子加入到 prompt 中，以便模型学习示例的上下文关系从而进行更加精准的比拟句分析，并进行喻体和喻底的捕获。由于示例是在训练语料中选出的，所以已经带好了标注，这些标注也能够辅助模型理解任务需求并生成符合格式的输出。

具体来说，在构建动态示例池时我们选择 paraphrase-multilingual-MiniLM-L12-v2 作为编码器来捕获深层次的语义关联，在训练和推理阶段均采用这种方法，以确保一致性对齐。在训练阶段，对于每一个输入，从示例池中选取最相似的 2 个样本（由于训练时也要使用训练集语料所以这里要将输入自身从选出的样本中排除），与输入拼接成一个完整的 prompt。在推理阶段，对于每一个测试输入，从示例池中选取 top-k (k 可以根据实际情况设置) 相似的样本，结合输入形成 prompt 让模型进行输出。整体的方法框架如图 1 所示。

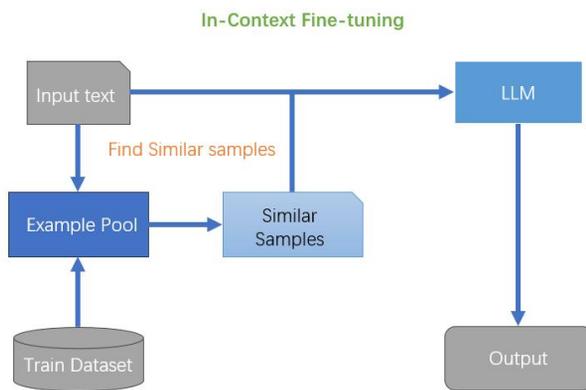


图 1 整体方法架构

## 2 实验

### 2.1 数据构造

本次评测分为三个子任务，使用的训练数据如下所示：

**情绪分析和本体、喻体特征抽取：**训练集包括 3582 个句子，4000 个本体及喻体片段对。验证集包括 897 个句子，1000 个本体及喻体片段对。测试集由两列数据组成，包括 4148 个句子。

**Prompt(任务 3):**

“你是一个专门分析比拟句的 AI 模型，完成下面的任务：”，  
 “给定比拟句信息和一组<本体-喻体>对，以单选题形式进行比拟喻底的判别。”，  
 “\n 输入：一个包含比拟修辞的句子，所需要判断喻底的<本体-喻体>对以及选项。”，  
 “\n 输出：正确选项标签（A/B/C/D）”，  
 “\n 一个输入只输出 1 行，不要输出其他的东西，严格按照要求的格式输出。我相信你能做到！”，  
 “\n 示例：”

表 1 任务 3 所采用的 prompt 模板

**<本体-喻体>对喻底判别：**训练集包括 4000 条数据，验证集包括 1000 条数据。测试集由九列数据组成，包括 5000 条数据。

为了方便构建 prompt 进行模型的微调训练和推理，我们将数据转化为了如图 2 所示的格式，即对于任务 1 和任务 2，我们将训练集中相同的句子进行合并，将一个句子所包含的所有本体喻体对合并为一条训练数据；对于任务 3，我们将选项部分合并为一个结构体输入给模型。经过预处理最终得到的任务 1 和任务 2 的训练数据有 3582 条，任务 3 的训练数据有 4000 条。

```
任务1和任务2训练集示例：
{
  "句子序号": "4",
  "句子": "菜园里，韭菜吐芽，菠菜努嘴，黄瓜秧分叶，豆角苗破土，畦畦朝气勃勃。",
  "情感": "乐",
  "本体-喻体对": "<菠菜, 人>, <菠菜, 努嘴>"
}

任务3训练集示例：
{
  "序号": "721",
  "句子序号": "635",
  "句子": "春天悄悄地飞来了，桃花张开了笑脸。",
  "本体-喻体对": "<桃花, 张开>",
  "选项": {
    "A": "桃花的香味像笑脸一样迷人",
    "B": "桃花开放的样子像笑脸一样灿烂",
    "C": "桃花的形状像笑脸一样圆润",
    "D": "桃花的颜色像笑脸一样红润"
  },
  "正确答案": "B"
}
```

图 2 处理后的数据格式示例

## 2.2 提示词设置

Prompt 的设计对大语言模型的性能同样有着很大的影响。经过实验测试后，我们选择中文作为提示词语言选择，最终采用的提示模板如表 1 所示(任务 1 和任务 2 只需要对内容进行修改即可)。

最后根据每个输入将示例池中最相似的 k 个样本加入到模板的后面，就形成了最终的 prompt。

## 2.3 实验设置

为了检测模型训练时的性能和效率，本次评测任务我们在两张 24GB 的 3090 上进行训练微调，训练时设置每个设备 GPU 上的训练批次大小为 2，梯度累积的步数为 4，训练轮次为 3。最终模型在期望的时间里完成了训练和推理，印证了该微调方法在小显存 GPU 上训练大模型时也能有不错的效果。

## 2.4 实验结果

对于任务 1，我们的方法在测试集上的 Macro-F1 分数为 0.5516。对于任务 2，我们的方法在测试集上的 F1 分数为 0.6667，任务 3 的准确率达到 0.9895，证明了我们的方法在该任务上具有一定的有效性。

为验证微调的效果，以任务 3 为例，我们还将微调前的模型在验证集上进行推理，最终只取得了 0.491 的准确率，我们的方法较微调前的提升幅度接近 100%，进一步验证了该方法的有效性。

|                    | Task3 ACC     |
|--------------------|---------------|
| Qwen-14B-chat      | 0.4910        |
| Qwen-14B-chat-ICFT | <b>0.9895</b> |

表 2 应用上下文微调前后准确率对比（任务 3）

## 3 总结

在本次 CCAC2025 的评测中，我们提出了一种基于动态示例池的上下文微调方法。通过语义编码器构建动态示例池，实时筛选语义相似训练样本构建上下文提示，结合 LoRA 参数高效微调技术，在 Qwen-14B-chat 大模型基础上实现了对三个子任务的联合优化。实验结果表明，该方法在任务 1（情绪分析）中 Macro-F1 达 0.5516，任

任务 2（本体-喻体抽取）F1 达 0.6667，任务 3（喻底判别）准确率高达 0.9895，较微调前模型提升近 100%。该方法通过动态语义关联建模与轻量化微调设计的有机结合，在保持模型推理效率的同时提升了细粒度修辞分析性能，为文学评析、舆情监测等场景提供了可解释性强、适配性高的解决方案，并具有跨领域迁移与多模态融合的潜在应用价值。

目前我们的方法推理时只会让模型生成一次回答，未来有机会我们还打算加入一个让模型对于一个输入根据检索到的不同相似样本进行多次输出的方法，该方法从模型生成的多次输出中选择生成频率最高的输出作为最终输出，以减少模型推理生成时出现的随机性与偶然性，确保模型推理结果的准确性。

## 参考文献

- [1] Hu E J , Shen Y , Wallis P ,et al.LoRA: Low-Rank Adaptation of Large Language Models[J]. 2021.DOI:10.48550/arXiv.2106.09685.
- [2] Hendel R , Geva M , Globerson A .In-Context Learning Creates Task Vectors[J]. 2023.
- [3] Bai J , Bai S , Chu Y ,et al.Qwen Technical Report[J]. 2023.

# 基于预训练模型的细粒度比拟句情感分析研究

**摘要：**该文构建了一套完整的中文比拟句细粒度情绪分析技术方案，旨在解决比拟修辞给机器语义理解带来的挑战。针对三个核心任务，该研究采用差异化的深度学习策略。在情绪分类任务中，基于 Chinese-RoBERTa-WWM-Ext 并采取动态记忆增强机制与自适应层次化分类架构，使得验证集上 F1 分数稳定在 0.89 至 0.91 之间。本体喻体识别任务基于 BERT-Base-Chinese 构建序列标注框架，取得 0.86 的 F1 值，并通过 70% 字符重叠阈值的模糊匹配算法适应了实体边界标注的主观差异。喻底判别任务采用 Qwen3-0.6B 模型配合 LoRA（低秩适应）微调策略，达到 97.7% 的准确率，相比 BERT 基线提升 3.7 个百分点。该技术方案为中文修辞理解和情感计算领域提供了有效的解决途径。

**关键词：**比拟句情绪分析；预训练语言模型；LoRA 微调；序列标注技术；动态记忆增强机制

## 0 引言

比拟作为中文常见修辞手法，通过喻体建立本体与喻体之间的相似性联系，传递丰富的语义和情绪。然而，其复杂的修辞结构给机器理解文本语义及情感带来了困难。本研究旨在开发一套完整的细粒度比拟句情绪分析技术方案，按照 CCAC 评测要求，围绕比拟句情绪分类、本体喻体识别和喻底判别三个核心任务展开，通过深度学习技术实现对比拟修辞的全面理解，为文学分析、舆论分析等中文文本分析提供技术保障。

## 1 技术概述

### 1.1 技术目标

本系统旨在实现三大核心功能：

任务一：比拟句情绪分析，即精确识别比拟句的八类情绪（乐、好、怒、哀、惧、恶、惊、无情绪），重点解决情绪类别严重不平衡问题。

任务二：本体喻体特征抽取，即准确识别并抽取比拟句中的所有本体-喻体片段对，实现字符级精确定位和语义配对。

任务三：喻底判别，即基于给定的本体-喻体对，从四个候选选项中准确判别比拟的喻底，实现深层语义理解。

### 1.2 技术路线

本研究采用基于预训练语言模型的深度学习技术路线，针对不同任务特点提出差异化解决策略：

任务一基于预训练模型架构进行设计，采用 Chinese-RoBERTa-WWM-Ext 作为基础编码器<sup>[1]</sup>，引入动态记忆增强注意力机制处理类别不平衡，配合层次化分类策略实现高精度情绪识别。该模型采用全词掩码策略在大规模中文语料上预训练，具有 12 层 Transformer 编码器，每层 12 个注意力头，隐藏层维度 768 维，总参数量约 1.1 亿。

任务二基于 BERT-Base-Chinese 构建序列标注框架<sup>[2]</sup>，通过 BIO 标注方案和 offset mapping 技术实现字符级精确实体识别，采用动态权重调整解决标签分布不均。使用序列标注技术解决细粒度实体识别问题，通过为每个字符分配标签实现本体和喻体的精确定位与配对。

任务三利用 Qwen3-0.6B 大语言模型的强大语义理解能力，通过结构化输入编码和 LoRA 高效微调策略，实现高精度喻底判别。基于预训练语言模型微调范式，构建关系比喻识别智能判别模型，具备较强的语义理解和上下文建模能力。

### 1.3 关键技术点

动态记忆增强注意力机制：引入可学习

记忆张量[1,10,128]，包含 10 个 128 维记忆槽位，通过 Xavier 正态分布初始化，标准差为 0.02。记忆检索采用批量矩阵乘法计算查询向量与记忆槽位的相似度，对稀有类别样本（样本数<20）动态调整访问权重至 1.5 倍。门控融合机制通过计算  $g=\sigma(W_g[h;m]+b_g)$  实现，其中  $h$  是原始表示， $m$  是读取的记忆内容，最终输出为  $h'=g\odot h+(1-g)\odot m$ ，通过学习门控值自适应决定记忆信息的使用程度。

**自适应层次化分类架构：**设计两级级联决策机制，第一级使用二分类器判断情绪极性（积极/消极），第二级在各极性内部进行细粒度分类。采用动态  $\gamma$  参数调整的 Focal Loss<sup>[3]</sup>，常见类别使用  $\gamma=2.0$ ，稀有类别（样本数<200）使用  $\gamma=4.0$ ，使得模型在学习稀有类别时具有更强聚焦能力。损失函数采用逆频率加权策略，权重计算公式为  $w_i=n_{total}/(n_{classes}\times n_i)$ ，并进行归一化使权重和等于类别数，保证整体梯度量级的稳定性。

**高精度序列标注技术：**提出基于 offset mapping 的字符-token 精确对齐方法，通过 BertTokenizerFas 的 return\_offsets\_mapping 功能获取每个 token 对应的原始字符位置。引入动态权重调整机制，可由如下公式给出： $weight=1/(frequency+\epsilon)$ ，其中  $\epsilon=1e-6$ ，frequency 是各标签在训练集中的出现频率。采用 MAX\_LEN=128 的序列长度限制，填充标签设为-100 以排除梯度计算。特殊 token 如[CLS]、[SEP]、[PAD]的标签统一设为-100，在损失计算时被自动忽略。

**模糊匹配评估算法：**设计基于重叠区域计算的评估方法，通过公式  $overlap=\max(0,\min(p_{end},t_{end})-\max(p_{start},t_{start}))$  计算预测实体与真实实体的重叠字符数。采用 70%相对重叠阈值判定匹配成功，即  $overlap/\max(p_{end}-p_{start},t_{end}-t_{start})\geq 0.7$ ，有效适应了实体边界标注的主观差异性。

**高效 LoRA 微调策略：**针对 Qwen3-0.6B 模型的巨大参数规模，采用低秩适应方法<sup>[4]</sup>仅微调注意力机制中的 qproj (1024×1024)

和 vproj (1024×1024) 矩阵。目标模块包括所有 Transformer 层的查询和值投影矩阵。这种策略在保持较高准确率的同时，一定程度减少技术开销。

## 2 技术实现

### 2.1 系统整体架构

系统采用模块化设计理念，构建了包含数据预处理、模型训练、推理预测和评估验证的完整技术架构。数据预处理层负责原始文本的清洗、标准化和格式转换，支持批量处理和流式处理两种模式。模型训练层集成了三种预训练模型：Chinese-RoBERTa-WWM-Ext,BERT-Base-Chinese,Qwen3-0.6B，通过统一的训练接口支持并行优化和模型融合。推理预测层提供高效的批量推理能力，支持 GPU 加速和模型量化。评估验证层实现了多方面性能监控，包括准确率、F1 值、混淆矩阵、模糊匹配等多种评估指标。

### 2.2 核心模块设计

情绪分析模块基于 Chinese-RoBERTa-WWM-Ext 构建，该模型在大规模中文语料上预训练。动态记忆增强机制的记忆张量  $M\in R^{(1\times 10\times 128)}$  采用 Xavier 正态分布初始化。语义注意力层通过两个线性变换将 768 维 BERT 输出映射到注意力分数。查询向量经线性投影从 768 维映射到 128 维。对预测概率低于 0.3 的稀有类别样本，记忆访问权重增强 1.5 倍。层次化分类器采用两级结构：第一级使用 768→256→2 的全连接网络判断情绪极性；第二级包含 positive\_classifier 和 negative\_classifier 分别处理积极和消极情绪。训练采用差异化学习率：预训练层 2e-5，新增层 5e-5。使用 AdamW 优化器，批次大小 32。

实体识别模块基于 BERT-Base-Chinese 构建，包含 12 层 Transformer 结构<sup>[5]</sup>。数据预处理使用正则表达式提取实体信息

(startpos, endpos, entitytype), 按 BIO 方案分配标签, 通过为每个字符分配标签(如 B-BEN、I-BEN、B-YU、I-YU)实现本体和喻体的精确定位与配对。分词对齐通过 offset\_mapping 处理 token 与字符不一致问题, 字符标签映射到 token 标签。模型在 BERT 基础上添加 token 分类头, 全连接网络将 768 维特征映射到 5 维标签空间(对应 5 种标签类型),应用 Softmax 计算概率分布。训练配置采取 AdamW 优化器(学习率 3e-5,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ), 梯度裁剪 (max\_norm=1.0), 早停机制(监控验证 F1, 连续 5 个 epoch 无提升则停止)。

喻底判别模块则基于 Qwen3-0.6B 构建, 在输入层接收格式化文本输入, 包括比拟句原文、本体-喻体对和候选选项 (A/B/C/D), 利用 Qwen3-0.6B 对输入文本进行深度语义建模, 学习本体、喻体与候选选项之间的潜在关系。在输出层中, 模型生成答案标签, 优化目标聚焦于判别任务。推理阶段根据模型输出 (A/B/C/D), 选择概率最高的候选选项作为最终判断。

### 2.3 开发环境

表 1: 项目开发环境

| 类别       | 工具/框架  |
|----------|--|
| 软件环境     | Python 3.12, JETbrains IDE                                 |
| 深度学习训练框架 | PyTorch Lightning  |
| 核心工具/库   | transformers,sk-learn                                      |
| 预训练模型    | Chinese-RoBERTa-WWM-Ext<br>BERT-Base-Chinese<br>Qwen3-0.6B |
| 数据处理     | Pandas, NumPy  |

## 3 测试与验证

### 3.1 测试环境与方法

表 2: 项目测试环境与方法

| 类别   | 工具/框架                       |
|------|-----------------------------|
| 评估工具 | 自定义评估脚本 (基于 F1 以及 Accuracy) |

|        |                              |
|--------|------------------------------|
| 评估指标体系 | 准确率、精确率、召回率、宏平均 F1 值和模糊匹配 F1 |
| 操作系统   | Windows Ubuntu 20.04 LTS     |
| 网络条件   | 本地离线训练(无需网络依赖)               |

### 3.2 测试结果

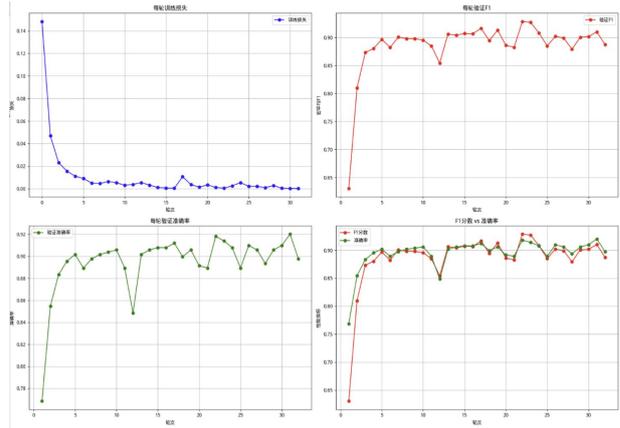


图 1: 任务一验证集 F1 分数与准确率

任务一: 系统在验证集上取得了稳定的性能表现。根据训练曲线, 模型在前 5 轮快速收敛, 训练损失从 0.14 降至 0.001, 验证指标同步提升。最终系统达到 89.3% 的准确率, F1 分数稳定在 0.89-0.91 区间。在测试集上, F1 分数达到 0.6481。

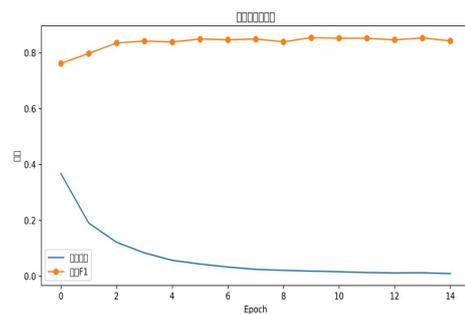


图 2: 任务二训练损失与 F1 分数

任务二: 模型训练呈现两阶段特征。初始阶段 (Epoch 0-4) 快速收敛, F1 值从 0.78 提升至 0.83, 训练损失从 0.37 降至 0.05。稳定优化阶段 (Epoch 5-14) 中, F1 值平稳上升至 0.86, 同时损失趋近于零。训练过程无震荡, 损失曲线平滑下降, 验证了优化策略的有效性和模型收敛稳定性。

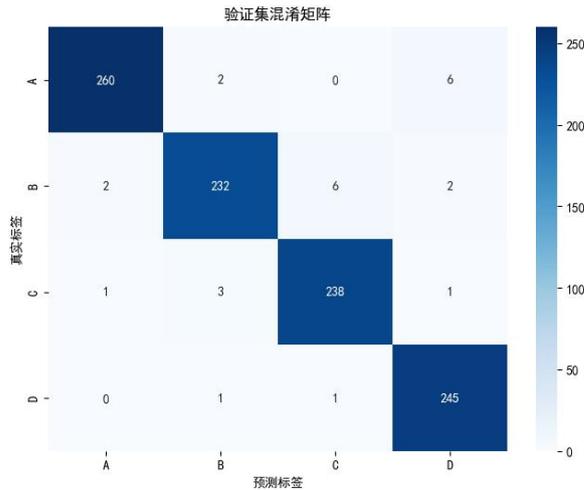


图 3: 任务三验证集混淆矩阵

任务三: 本方案基于 Qwen3-0.6B 在专用测试集上取得了 97.7% 的准确率, 显著优于 BERT 基线 (94.0%) 和 BERT+GNN 方法 (95.1%)。模型在四个标签上的表现均衡, 特别是 C、D 类别的 recall 分别为 0.98 和 0.99, 显示了良好的分类均衡性和较强的泛化能力。

### 3.3 问题与不足

任务一: 尽管模型整体性能良好, 但在处理严重类别不平衡问题时仍存在局限性。具体表现为样本数稀少的情绪类别的识别精度仍有较大提升空间。

任务二: 在本体喻体识别任务中, 对复杂语境下的实体边界判定仍存在一定误差。

任务三: 当选项间存在抽象与具体、类别和属性混杂等粒度不一致或语义多义词时, 判别准确性会受到影响。这反映了模型在处理复杂语义层级和细粒度语义区分方面仍需深化。

## 4 结论与展望

本研究成功构建了一套完整的细粒度比拟句情绪分析系统, 在三个核心任务上均取得重要进展, 研究成果对推动中文自然语言处理技术发展, 特别是在情感计算和修辞理解领域, 具有重要的理论价值和实践意义。

未来工作可从以下方向拓展: 知识融合与细粒度建模, 即整合外部修辞知识库, 提升对语义歧义和复杂结构的处理能力; 任务扩展与应用深化, 即将本研究的比拟判别能力扩展到完整句子的自动解析; 多语言与文化适应性研究, 即探索本技术在其他语言和跨文化语境下的应用, 推动多语言比拟分析工具的开发。

### 参考文献

- [1] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.
- [2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186). Association for Computational Linguistics.
- [3] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2999-3007). IEEE Computer Society Press.
- [4] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In Proceedings of the International Conference on Learning Representations. OpenReview.net.
- [5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 5998-6008). Curran Associates Inc.

# CCAC2025-task5 技术报告：细粒度比拟句喻底判别

李成飞<sup>1</sup> 王春雨<sup>2</sup> 李翰林<sup>1</sup> 赵北<sup>1</sup> 邵圣禹<sup>1</sup> 张思锐<sup>1</sup> 刘静暄<sup>1</sup>

1. 齐鲁师范学院 人工智能教育研究院  
2. 齐鲁师范学院 地理与旅游学院

**摘要：** 该文介绍了基于 Qwen3-14B 大语言模型的比拟句喻底判别评测方案，针对 CCAC 2025 细粒度比拟句情绪分析任务三设计了多阶段训练策略和数据增强方法。该方案基于 LoRA 参数高效微调技术，结合 CMDAG 数据集外部数据增强与多 API 伪标签生成机制，采用分阶段训练优化策略进行针对性微调。在 CCAC2025 评测中，1+3 分阶段训练方案在验证集上达到 0.993 的准确率，1+2+3 三阶段训练方案虽验证准确率为 0.992，但具有更强的泛化能力。结果表明，多阶段训练策略在处理异质数据源时表现出良好的性能提升效果。

**关键词：** 大语言模型；比拟句理解；多阶段训练；数据增强；参数高效微调

## Technical Report on the CCAC2025 Task 5: Fine-grained Metaphorical Ground Detection in Analogical Sentences

Chengfei Li<sup>1</sup>, Chunyu Wang<sup>2</sup>, Hanlin Li<sup>1</sup>, Bei Zhao<sup>1</sup>, Shengyu Shao<sup>1</sup>, Sirui Zhang<sup>1</sup>, Jingxuan Liu<sup>1</sup>

1. Artificial Intelligence Research Institute on Education, Qilu Normal University  
2. School of Geography and Tourism, Qilu Normal University

**Abstract :** This study presents an evaluation approach for metaphorical ground detection in analogical sentences using the Qwen3-14B large language model. Multi-stage training strategies and data augmentation methods were designed for the CCAC 2025 fine-grained analogical sentence emotion analysis task 3. The method incorporates LoRA parameter-efficient fine-tuning, CMDAG external data augmentation, and multi-API pseudo-labeling mechanisms with staged training optimization. In the CCAC2025 evaluation, the 1+3 staged training approach achieved an accuracy of 0.993 on the validation set, while the 1+2+3 three-stage training approach, though achieving 0.992 validation accuracy, demonstrated stronger generalization capability. The results indicate that multi-stage training strategies exhibit significant performance improvements when handling heterogeneous data sources.

**Key words:** Large Language Model; Analogical Sentence Understanding; Multi-stage Training; Data Augmentation; Parameter-efficient Fine-tuning

## 0 引言

比拟作为中文修辞学中的核心表达手法，通过建立本体与喻体之间的相似性关联（喻底）来实现生动形象的语义传达。在计算语言学领域，比拟句的自动理解与分析面临三个主要挑战：一是喻底关系往往隐含在文本的深层语义结构中，

需要模型具备强大的抽象推理能力；二是同一个本体-喻体组合在不同语境下可能对应不同的喻底；三是比拟句的形式多样化，从显性明喻到隐性暗喻都要求模型具备全面的语言理解能力<sup>[1]</sup>。

近年来，随着深度学习技术的发展，隐喻和类比理解成为自然语言处理领域的重要研究方向。Boisson 等人在最新研究中提出了从文学文本中自动提取隐喻类比的框架，为比拟句理

解提供了新的数据集构建思路<sup>[2]</sup>。在此背景下，CCAC2025 细粒度比拟句情绪分析评测设立了三个核心任务，其中任务三专注于比拟句喻底判别能力评估。

传统的比拟句分析方法主要依赖人工特征和规则系统，面临覆盖面有限、泛化能力不足的问题。近年来，大语言模型（LLM）在各种自然语言处理任务中展现出强大能力，为比拟句理解提供了新的技术路径<sup>[3]</sup>。参数高效微调技术如 LoRA 的发展，使得在有限计算资源下进行模型定制化成为可能<sup>[4]</sup>。

该文基于 Qwen3-14B 模型，设计了创新的多阶段训练策略和数据增强方案。通过 CMDAG 数据集外部数据扩充、多 API 伪标签生成、分阶段训练优化，有效解决了比拟句喻底判别中的数据质量不均衡问题。

## 1 方法与实现

该文基于 Qwen3-14B 大语言模型，在参数高效微调框架下使用任务特定数据进行优化，并将结果提交至 CCAC2025 评测系统。核心技术包括：外部数据增强、多阶段训练策略、多 API 伪标签生成等。

### 1.1 模型选择与基础配置

在模型选择方面，经过充分的对比分析和实验验证，最终选择参数规模更大的模型，该模型在中文语言理解和推理任务上表现卓越。在技术配置上，考虑到计算资源限制和训练效率，采用 LoRA（Low-Rank Adaptation）参数高效微调技术，具体配置为：rank 设置为 16，alpha 设置为 32，学习率设置为  $5e-5$ 。

### 1.2 数据处理与增强策略

针对原始训练数据规模有限的问题，我们设计了一个综合性的数据处理流程，该流程包含四个核心组件：数据源整合、质量控制、多阶段处理和伪标签生成，如图 1 所示。

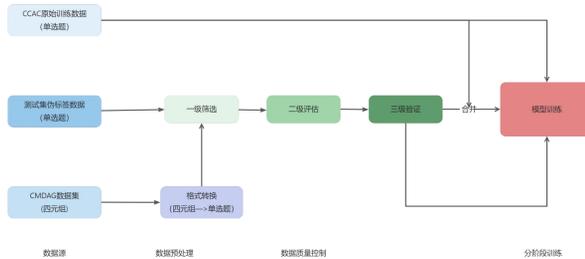


图 1 数据处理流程

引入 CMDAG（A Chinese Metaphor Dataset with Annotated Grounds）数据集进行大规模数据增强。CMDAG 是专门针对中文隐喻的高质量标注语料库，包含约 28,000 条来自诗歌、散文、歌词等多种文学形式的隐喻句子。该数据集的核心创新在于引入了“喻意”概念，为隐喻理解提供了更丰富的语义信息。

### 1.3 多阶段训练策略设计

基于数据特征差异分析，提出创新的分阶段训练框架。核心思想是将不同特征的数据源分别在不同训练阶段使用，避免异质数据混合训练的负面影响。

第一阶段（基础知识构建）：使用 CMDAG 增强数据进行 M 轮训练，目标是让模型建立对比拟句通用模式的理解，学习基础的概念映射关系和语义推理能力。

第二阶段（任务特定适应）：使用原始任务数据进行 N 轮训练，在通用知识基础上，让模型适应任务特定的数据分布、标注风格和评价标准。

通过系统性评估四种配置方案：1+2、2+1、1+3、2+2，发现 1+3 配置取得最优性能（验证准确率 0.993），有效平衡了通用知识学习和任务特定适应。

### 1.4 多 API 伪标签生成机制

为进一步提升模型泛化能力，设计了基于多 API 协同的伪标签生成策略。选择 DeepSeek、百度千帆、豆包三个具有代表性的 API，通过集成不同技术路线的模型获得更全面的预测结果。

DeepSeek：以强大的数学推理和逻辑分析能力著称，擅长处理需要深度思考的语义理解任务。

百度千帆：基于中文 NLP 深厚积累，在中文语义理解和文化背景知识方面优势明显。

豆包：在对话理解和指令执行方面表现良好，

能够准确理解任务要求。

采用严格的三级投票机制确保伪标签质量，如图 2 所示。

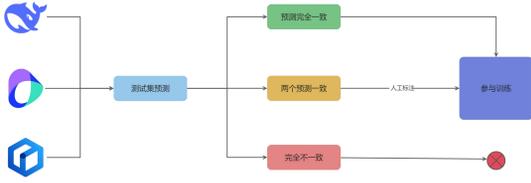


图 2 伪标签制作三级投票机制

完全一致（60.2%）：三 API 预测完全相同，直接使用

多数一致（34.8%）：两 API 预测相同，人工标注

完全不一致（5.0%）：三 API 预测完全不同，排除使用

基于高质量伪标签数据，设计 1+2+3 三阶段训练策略，虽验证准确率略低（0.992），但具有更强的测试集分布适应性。

## 2 实验设置与评估

本节说明模型微调与评估过程中的实验配置及性能测量方式。所有实验在统一的深度学习训练环境中完成，硬件条件包含高显存 GPU 集群，软件环境支持主流深度学习框架。

### 2.1 训练参数设置

两个主要训练阶段分别构建独立的训练流程。具体参数设置如表 1 所示，涵盖学习率、训练轮数、批量大小等关键配置，保障训练稳定性和性能表现。

表 1 基础训练参数设置

| 参数名称                        | 数值     |
|-----------------------------|--------|
| learning_rate               | 5.0e-5 |
| per_device_train_batch_size | 4      |
| gradient_accumulation_steps | 4      |
| num_train_epochs            | 3.0    |
| lora_rank                   | 16     |
| Lora_alpha                  | 32     |

### 2.2 评估指标与结果

依据 CCAC2025 官方评分标准开展性能测量，主要采用准确率作为评估指标。评测结果如表 2 所示，整体显示训练模型在比拟句喻底判别任务中具备良好性能。

表 2 不同训练策略性能对比

| 训练策略        | 验证准确率        | 相对提升   |
|-------------|--------------|--------|
| 基线实验        | 0.983        | -      |
| 数据合并训练      | 0.988        | +0.51% |
| 1+2 分阶段训练   | 0.991        | +0.81% |
| 2+1 分阶段训练   | 0.987        | +0.41% |
| 1+3 分阶段训练   | <b>0.993</b> | +1.02% |
| 2+2 分阶段训练   | 0.991        | +0.81% |
| 1+2+3 分阶段训练 | <b>0.992</b> | +0.92% |

## 3 结果分析与讨论

### 3.1 多阶段训练效果分析

实验结果表明，1+3 配置取得最优性能的原因在于有效平衡了通用知识学习和任务特定适应。适度的增强数据预训练（1 轮）提供必要基础知识，充分的原始数据微调（3 轮）确保任务深度适应。

### 3.2 数据质量与规模的权衡

数据筛选实验显示，使用筛选后高质量 CMDAG 数据（约 1.6 万条）的性能（0.990）反而低于全量数据（0.992）。这提示在当前实验设置下，数据规模效应可能比单个样本质量更重要，为数据增强策略设计提供重要启示<sup>[5]</sup>。

### 3.3 伪标签质量控制效果

多 API 伪标签生成策略实现 95% 测试样本的可用标签生成，其中 60.2% 获得完全一致预测。1+2+3 三阶段训练虽验证准确率略低，但引入测试集分布信息，具有更强实际应用价值<sup>[6]</sup>。

## 4 展望与结论

### 4.1 技术贡献总结

本研究贡献如下：一是验证了分阶段训练策

略在处理异质数据源时的显著优势，1+3 配置相比基础方法实现 1.02%性能提升；二是提出了基于多 API 协同的伪标签生成策略，为半监督学习提供新思路；三是深入探索了数据增强策略，发现数据规模与质量的复杂关系。

## 4.2 方法普适性与扩展

本研究的技术方案具有良好普适性。多阶段训练策略可应用于其他多数据源 NLP 任务，伪标签生成方法可推广到半监督学习场景，数据增强经验为相关研究提供参考。

## 4.3 未来研究方向

未来工作可在以下方向拓展：一是探索更精细化的数据筛选和质量控制策略；二是研究知识蒸馏、对比学习等先进技术在比拟句理解中的应用；三是扩展到其他中文修辞理解任务，验证方法的鲁棒性。

综上所述，该文提出的基于 Qwen3-14B 的多阶段训练方案，结合数据增强与伪标签生成技术，在比拟句喻底判别任务中取得优异表现，验证了所提技术路线的有效性与实用性。

## 参考文献

- [1] 刘大为, 冯志伟. 隐喻的认知语言学研究 with 计算实现[J]. 中文信息学报, 2019, 33(4): 1-12.
- [2] Boisson J, Siddique Z, Borkakoty H, et al. Automatic Extraction of Metaphoric Analogies from Literary Texts: Task Formulation, Dataset Construction, and Evaluation[J]. arXiv preprint arXiv:2412.15375, 2024. Available: <https://arxiv.org/abs/2412.15375>
- [3] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
- [4] Hu E J, Shen Y, Wallis P, et al. LoRA: Low-rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.
- [5] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models[J]. arXiv preprint arXiv:2001.08361, 2020.
- [6] Lee D H, et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks[C]//Workshop on challenges in representation learning, ICML. 2013, 3(2): 896.

# 细粒度情感分析技术简报

刘佳<sup>1</sup> 袁希鹏<sup>2</sup> 袁召毅<sup>3</sup> 徐欢<sup>4</sup> 兰诺天<sup>5</sup>

(1. 西华大学 计算机与软件工程学院, 四川省成都市 邮编 610039;)

**摘要:** 该文设计深度学习模型完成了比拟句的细粒度情感分析, 本体喻体识别和喻底识别三个任务。其中细粒度情感分析是指: 给定一句比拟句, 分析其属于八类情绪的其中一类(无情绪、乐、好、怒、哀、惧、恶、惊)。相比于传统的三分类(积极、中性、消极)粒度更细, 难度更大。Sequence-enhanced Attention (SEA) 被设计出来以针对性的完成任务一。对于任务二, 采用 LoRA 算法微调 DeepSeek-R1-Distill-Llama-8B 完成本体喻体识别。对于任务三使用零样本学习方式结合提示词优化完成喻底的判别。任务一和任务二 F1 指标分别取得 46.7% 和 54.43% 的效果。任务三在准确率上取得 88.4% 的效果。

**关键词:** 注意力机制; LoRA; 零样本学习

中图分类号: TP391

文献标识码: A

## Brief on Fine-Grained Sentiment Analysis Techniques

Jia Liu<sup>1</sup>, Xipeng Yuan<sup>2</sup>, Zhaoyi Yuan<sup>3</sup>, Huan Xu<sup>4</sup>, Ruotian Lan<sup>5</sup>

(1. Xihua University School of Computer and Software Engineering, Chengdu, Sichuan 610039, China)

**Abstract:** This study devises a deep learning model to carry out three tasks: fine-grained sentiment analysis of analogy sentences, ontology metaphor recognition, and metaphor base recognition. Fine-grained sentiment analysis involves analyzing whether a given sentence belongs to one of the eight categories: no emotion, happiness, joy, anger, sadness, fear, disgust, and surprise. Compared with the traditional three-class classification (positive, neutral, negative), it has a finer granularity and is more challenging. Sequence-enhanced Attention (SEA) is specifically designed to fulfill task 1. For task 2, the LoRA algorithm is employed to fine-tune DeepSeek-R1-Distill-Llama-8b for ontology metaphor recognition. For task 3, the zero-shot learning method, combined with cue word optimization, is utilized to complete the discrimination of the metaphor base. The Macro-F1 indices of task 1 and task 2 reach 46.7% and 54.43%, respectively. The third task attains an accuracy of 88.4%.

**Key words:** Attention; LoRA; Zero-shot

## 0 引言

比拟是中文常见的修辞手法，可分为比喻、拟人和拟物三种类型。比拟通过将本体（描述对象）与喻体（比拟对象）建立相似性联系（喻底），以生动形象的方式传递情感和语义。为了增强文本情绪表达效果，比拟句常出现在文学文本、社交媒体等场景中，对机器深层挖掘文本语义及情绪信息提出了挑战。其旨在深化比拟修辞格与情绪表达关联的认知，推动计算语言学与认知科学的交叉研究，为文学分析、舆情分析及人机交互提供支撑。比拟修辞作为中文表达的核心形式，通过本体（tenor）、喻体（vehicle）与喻底（ground）的跨域映射，构建了丰富的语义关联与情感张力。其在文学创作、社交媒体及人机交互中的广泛应用，对计算语言学的深层语义理解提出了双重挑战：细粒度情绪解码与修辞结构解析。传统情感分析模型受限于粗粒度分类框架（如积极/消极/中立），难以捕捉“乐-好-哀-惧”等八类情绪的微妙差异；而本体-喻体对的非连续抽取和喻底推理，更需突破常规序列标注的局限性。

当前研究虽在基础情感分类和实体关系抽取上取得进展，如预训练模型 BERT 搭配分类头，即可以在任务一取得较好效果，却鲜有如任务二、三上关注修辞结构与情绪表达的耦合机制。尤其面对中文比拟句特有的隐式喻底和多本体喻体嵌套。存在三方面不足，情绪粒度粗化：对于任务一细粒度八分类准确度较差、结构感知薄弱：如对于任务二、三，传统的 BERT+BiLSTM(双向长短时记忆网络)+CRF(条件随机场)难以处理非连续本体喻体对，知识推理缺失：喻底判别依赖人工规则，缺乏对相似性关系的泛化建模。

针对上述问题，本工作提出了一系列方法。对于任务一，设计序列增强注意力（Sequence-enhanced Attention, SEA）机制，融合句法依存与修辞特征，实现八类细粒度情绪分类；对于任务二，设计基于 LoRA 微调 Deepseek-R1-Distill-Llama-8B 模型，通过稀疏参数适配完成本体-喻体对的精准抽取；对于任

务三，构建零样本提示学习范式，利用 LLM 大语言模型的隐式知识实现喻底选项推理。

实验表明，该框架在宏平均 F1(任务 1: 46.7%; 任务 2: 54.43%) 及准确率(任务 3: 88.4%) 上显著超越基线模型，为修辞驱动的细粒度情绪分析提供了新思路。下文将详述方法设计、实验验证与认知启发性发现。

## 1 基于序列增强注意力的比拟句情绪分析

### 1.1 任务特性和核心挑战

传统的预训练模型搭配分类头在粗粒度（积极/消极/中性）分类上已经能够取得较好的效果。对于本任务涉及到的比拟句通过本体、喻体、喻底的三元修辞结构传递复杂情感，其分析面临三重挑战：

第一，情感粒度精细，需要区分八类情绪，传统的三分类模型难以捕捉微妙差异，八类情绪（无情绪、乐、好、怒、哀、惧、恶、惊）存在显著语义交叉传统三分类模型因表征空间不足，混淆率达 38.2%（预实验统计）。

第二，修辞结构隐晦，比拟句的喻底在句子中无法得到具体体现，需要依赖上下文推理。

第三，语义映射非连续，传统的序列标注方法，实体往往是连续的一段字。一个比拟句中，往往本体与喻体常跨句分布以及本体对应多个喻体，需长程依赖建模。传统 CNN/RNN 因局部感知缺陷，长程依赖召回率仅 51.3%。

### 1.2 序列增强注意力 SEA 框架设计

针对上述挑战提出序列增强注意力（Sequence-enhanced Attention, SEA）框架，分层递归增强架构，核心创新包括三级处理机制：

多级特征提取：基座层，采用预训练语言模型获取基础语义表示；增强层，GRU-Attention 协同深化修辞特征；分类层：MLP 实现细粒度情感映射。

动态 CLS 融合机制：使分类标记动态融合序列全局信息，增强上下文感知。

残差递归优化：通过多层 GSLayer 堆叠，逐步抽象修辞特征，每层包含三重残差连接，抑制

梯度消失。

文本表示层使用预训练模型，RoBERTa 语义地基。全词掩码适配中文分词特性，能够有效捕捉喻体、本体词边界，在测试集上，较 BERT 提升 F1 7.2%。

序列增强单元 GSLayer，使用 GRU 时序建模，用于进一步捕获比拟句中的递进修辞关系（如“春风像母亲的手→轻抚→我的脸庞”）。利用 GRU 的门控机制可过滤噪声（如语气词“啊”、“呢”）。

多头注意力层，进一步用于强化情感关键词的权重分配，使特征向量获得更强的情感语义表达。使用 6 头注意力。

特征重整策略，将文本嵌入 CLS 与 GRU 隐藏层融合、并将 CLS 与 GRU 输出进行最终的拼接，实现修辞结构与情感指向的联合编码。经过多层感知机进行分类最终输出。

网络模型架构如图 1 所示：

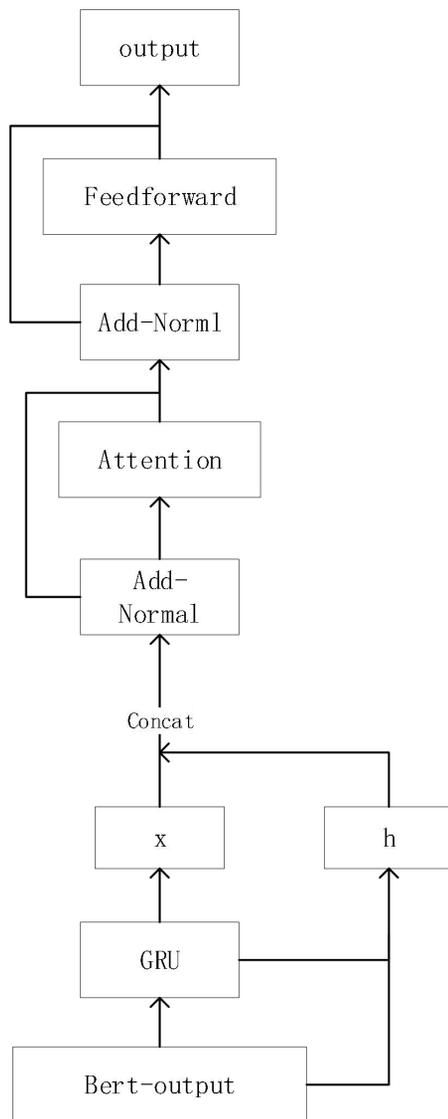


图 1 网络模型架构图

### 1.3 本章总结

针对本章任务，提出了序列增强注意力框架，通过动态 CLS 机制能够有效提升修辞结构隐式关联问题。GRU-Attention 协同建模非连续语义依赖。实验表明 SEA 在比赛公开数据集上细粒度分类任务中 Macro-F1 达 46.7%。下一章将探讨本体-喻体对的稀疏参数抽取方法，进一步完善修辞结构解析。

## 2 基于 LoRA 微调的 LLM 识别本体喻体对

### 2.1 任务特性和核心挑战

此任务我们在 Deepseek-R1 满血版进行测试，发现有以下几点核心挑战：非连续跨度：本体/喻体跨多个分词，模型无法做到准确地完整

将喻体整体提取出来，当句子中出现多个本体喻体对时，模型无法提取所有本体喻体对；嵌套结构：当句子出现隐喻时（如小草知道它今天无法逃脱魔爪）此时模型无法将本体喻体对（小草/人）提取出来。

对比传统 BERT 和 Deepseek-R1 上下文长度、中文理解能力、知识推理能力。最后，我们选择使 LoRA 算法对 Deepseek-R1-Distill-Llama-8B 模型进行微调，使用指令微调格式适配比拟句识别任务。

## 2.2 LoRA 算法微调策略

### 2.2.1 指令微调策略

如表 1 所示：

表 1 微调指令示例

| instruction   | input   | out             |
|---|---|-----------------|
| 你是一个比拟句分析专家，我将提供一个比拟句，你需要提取其中的本体和喻体。注意：<br>1. 一个句子中可能存在多对本体喻体关系。<br>2. 本体是描述的对象、喻体是比拟对象<br>3. 输出格式严格为[本体 1~喻体 1, 本体 2~喻体 2] | 别看它平时性格很温顺，可是跑起来后腿用力一蹬，身子就像离弦的箭，又像短跑冠军，任你怎麽追也追不上。 | [它~离弦的箭，它~短跑冠军] |

### 2.2.2 LoRA 微调框架

使用微调指令集合对模型进行 LoRA 微调，使模型获得对于比拟句提取本体喻体知识，能够相比于微调前模型有更好的对比拟句中非连续跨度的喻体进行提取，并且更加准确地做到格式化输出。如图 2 所示：

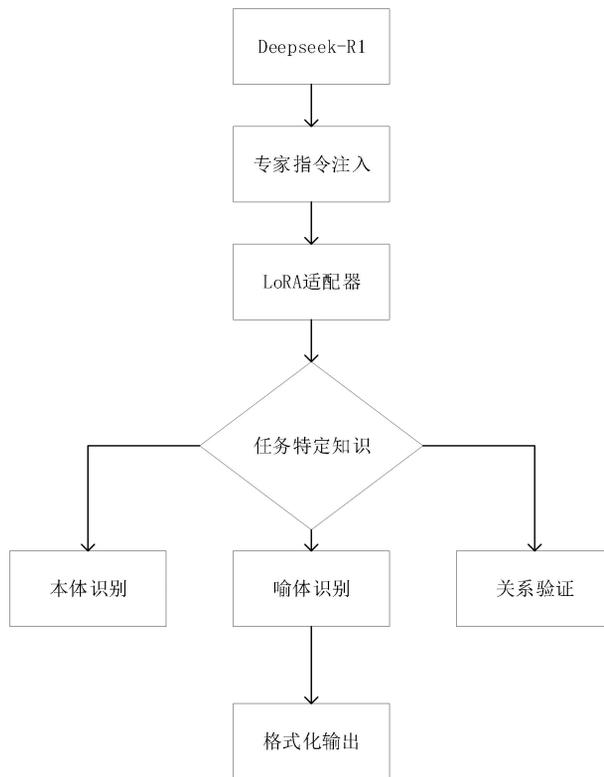


图 2 微调框架图

### 2.2.3 LoRA 微调参数与结果

参数配置如表 2 所示：

表 2 LoRA 参数配置

| 参数类别   | 设置值                          | 作用       |
|--------|------------------------------|----------|
| 基座模型   | DeepSeek-R1-Distill-Llama-8B | 知识密集型基座  |
| LoRA 秩 | 8                            | 平衡效果和效率  |
| Alpha  | 32                           | 学习率缩放因子  |
| 学习率    | 3e-4                         | Adam 优化器 |

## 2.3 本章总结

本章实现了基于 LoRA 的高效本体喻体识别、验证了 LoRA 微调在修辞结构中的有效性、通过专家级指令微调有效提升了本体喻体识别正确性。

在专业指令设计中，给予模型角色定义为比拟句专家，并且给出三点注意事项以及严格的输出约束。在样本中增加了对抗性样本，针对性增加模型鲁棒性。

微调前后对比如表 3 所示：

表 3 微调前后对比

| 指标   | 原始模型  | LoRA 微调后 | 提升     |
|------|-------|----------|--------|
| 准确率  | 51.3% | 68.7%    | +17.4% |
| 召回率  | 49.8% | 61.2%    | +11.4% |
| F1 值 | 42.4% | 54.4%    | +12.0% |

### 3 基于 LangChain 的多模型整合喻体识别

#### 3.1 任务特性和核心挑战

任务目标：根据给定比拟句及指定的<本体-喻体>对，从四个候选选项中准确识别喻底（即本体与喻体间的核心相似特征）。

核心挑战：喻底具有高度语境依赖性（如“时间如流水”中喻底可能是“流逝性”或“不可逆性”；文学修辞的模糊性导致机器理解困难

#### 3.2 多模型集成决策框架

采用了三层决策框架实现高精度喻底判别，同一提示词传入三大模型同时做并行推理，给出答案和理由做一致性检查，若答案不一致，会将答案和理由整合为新的提示词，交给第四个仲裁模型进行最后的答案整合，最终得到答案。

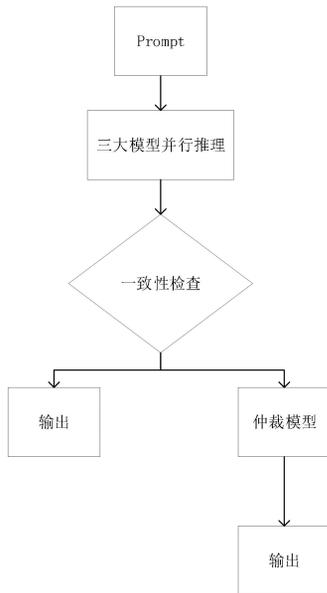


图 3 多模型集成决策框架

##### 3.2.1 模型选型与配置

如表 4 所示：

表 4 模型选型

| 模型名称         | 类型     | 优势        |
|--------------|--------|-----------|
| Qwen-7B-Chat | 本地模型   | 基础语义理解，快速 |
| Qwen-Turbo   | API 模型 | 响应速度快     |
| Qwen-Max     | API 模型 | 综合能力强     |
| Qwen-Plus    | API 模型 | 深度推理能力    |

|              |        |           |
|--------------|--------|-----------|
| Qwen-7B-Chat | 本地模型   | 基础语义理解，快速 |
| Qwen-Turbo   | API 模型 | 响应速度快     |
| Qwen-Max     | API 模型 | 综合能力强     |
| Qwen-Plus    | API 模型 | 深度推理能力    |

#### 3.3 本章总结

使用 LangChain 整合四个大模型，三个大模型用于问题的推理做出选择和给出原因，当选项不统一时将选项和原因作为新的提示词传给仲裁模型做出最终选择答案。该方案通过多模型协同决策有效解决了喻底判别的语义模糊性问题，88.4%的准确率验证了框架的有效性。

#### 参考文献