# 基于多模型协同的儿童互联网新闻风险管理与价值观引导框架

**梁宇蓝[1], 王悦[1], 于东[1,*], 刘鹏远[1,2],康晨[1]**
1.北京语言大学信息科学学院，北京，100083
2.国家语言资源监测与研究平面媒体中心，北京，100083
{202321198421,202321198096}@stu.blcu.edu.cn
yudong@blcu.edu.cn, liupengyuan@blcu.edu.cn, kangchen@blcu.edu.cn

## 摘要

随着互联网在儿童群体中的广泛普及，新闻内容的“毒性遗留”与价值观缺失已成为亟待解决的安全挑战。本文提出了一种多模型协同的儿童新闻改写框架（CRV-LLM），旨在从词汇、事件、标题和价值观四个维度，对原始新闻文本进行深度风险识别与精准改写。CRV-LLM集成了四个轻量化风险检测模型和DeepSeek-R1-Distill-Qwen-32B改写模型，通过模型间的协同与反馈，能够在保证儿童可读性的前提下，有效剔除潜在有害信息并植入积极价值引导。实验结果表明，CRV-LLM框架在安全性、教育性等核心指标上优于主流模型，且推理效率提升62%，为儿童互联网内容安全管理提供了一种高效、可扩展的技术方案。

**关键词：** 文本生成；大型语言模型；儿童新闻；风险控制；价值观对齐

# A Multi-Model Collaborative Framework for Child Internet News Risk Management and Value Guidance

**Yulan Liang[1], Yue Wang[1], Dong Yu[1,*], Pengyuan Liu[1,2], Chen Kang[1]**
1.Faculty of Computer Science,Beijing Language and Culture University,Beijing,100083
2.National Language Resources Monitoring and Research Center for Print Media,Beijing,100083
{202321198421,202321198096}@stu.blcu.edu.cn
yudong@blcu.edu.cn, liupengyuan@blcu.edu.cn, kangchen@blcu.edu.cn

## Abstract

With the widespread proliferation of the internet among children, residual toxic content and the absence of value-oriented guidance in online news have emerged as pressing safety challenges. This paper proposes a multi-model collaborative framework for children's news rewriting — CRV-LLM (Children's Risk-control and Value-guidance Large Language Model) — designed to conduct in-depth risk identification and precise rewriting across four key dimensions: vocabulary, events, headlines, and values. CRV-LLM integrates four lightweight risk detection models with a DeepSeek-R1-Distill-Qwen-32B rewriting model, achieving effective removal of potentially harmful information and embedding of positive value guidance, all while ensuring readability for young audiences. Experimental results demonstrate that CRV-LLM outperforms mainstream models on core indicators such as safety and educational value, with a 62% improvement in inference efficiency. This work offers an efficient, scalable technical solution for the safe management of children's online news content.

*为通讯作者
**基金项目：**教育部人文社科一般项目(23YJAZH184)，教育部语合中心国际中文教育研究课题(23YH38B)，北京语言大学梧桐创新平台(中央高校基本科研业务费，21PT04)
©2025 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

**Keywords:** Text Generation , Large Language Models , Children's News , Risk Control , Value Alignment

## 1 引言

新闻阅读对儿童的成长具有重要意义 (金丽莉., 2025)，它不仅能拓展他们的知识视野，还能培养批判性思维、信息素养和社会责任感。然而，许多传统新闻的语言、内容和表达方式并不适合儿童阅读，可能涉及复杂的社会议题、暴力事件或负面情绪，甚至可能对儿童的认知产生误导。因此，如何改写新闻，使其更加易懂、安全，并富有教育意义，成为儿童新闻发展的核心目标。

目前，针对儿童的文本风险管理主要依赖于关键词过滤和机器学习模型。关键词过滤 (Aho et al., 2024; 崔洪振et al., 2024)通过设定特定词汇或短语，识别并屏蔽潜在风险内容，而机器学习模型 (Rosenblatt,F., 1958)则利用算法对文本进行分类，以识别其中的风险。这些方法在一定程度上提升了内容安全性，但仍存在明显局限。例如，关键词过滤难以精准识别隐晦或语境相关的风险内容，可能导致误杀或漏检。机器学习模型则可能因训练数据的限制，无法有效识别儿童文本中的隐性风险，如潜在的心理影响或价值观偏差，甚至可能引入算法偏见，导致错误分类。

相比之下，大型语言模型（LLM）通过在海量文本数据上的预训练，展现出卓越的语义理解和生成能力，使其在文本改写和内容创作方面具有明显优势 (Zhao et al., 2023)。相较于传统的小型模型，LLM无需特定任务训练，即可有效改善文本的连贯性、可读性和内容丰富度。在安全控制方面，目前的人工智能安全研究仍主要聚焦于成人用户、国家安全及一般性的错误信息识别，一些研究 (Kurian,N, 2024; Prasanjit Rath et al., 2025; Junfeng Jiao et al., 2025)开始关注儿童与大模型交互的潜在风险，并尝试在生成过程中加入安全约束，以引导模型生成符合道德标准的儿童内容，但针对儿童内容安全的研究仍然相对较少，系统性研究更为稀缺。在儿童新闻改写方面，已有研究探索利用大模型辅助新闻生成，以优化文本质量 (Du et al., 2024)。然而，尽管可以通过提示词引导大模型基于成人新闻内容生成适合儿童阅读的版本，但生成结果仍可能保留部分不适宜儿童阅读的风险内容，并且缺乏有效的价值观引导（见图1）。这一现象表明，在儿童内容改写领域，仅依赖现有的技术手段仍存在明显不足，需要进一步探索更精细化的风险管理方法，以确保生成内容的安全性、教育性和价值观一致性。且现有大型语言模型虽具备强大的生成能力，但其全流程处理耗时较长，难以满足儿童新闻实时改写和大规模部署的需求。

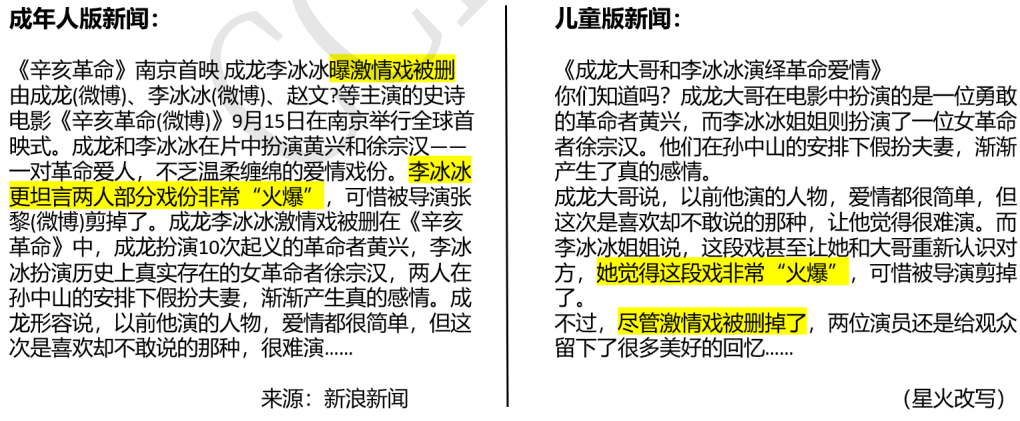



Figure 1: 同一主题新闻的成年人版和对应儿童版示例

针对上述问题，我们提出了一种基于四个轻量化风险检测模型与DeepSeek-R1-Distill-Qwen-32B（以下简称DeepSeek-32B）协同工作的创新框架，以实现更高效、更精准的儿童新闻改写。在该框架中，四个轻量化风险检测模型分别聚焦于词汇、事件、标题和价值观四个核心维度，对新闻文本进行全面的风险识别和控制，并针对不同风险点提供具体的改写建议。



(c) 2025 中国中文信息学会计算语言学专业委员会

这种模块化检测策略确保了风险筛查的精准性，同时避免了传统方法中过度过滤或误杀的弊端。DeepSeek-32B在该框架中承担智能风险沟通的角色，不仅能够综合采纳四个检测模型的改写建议，还能在语境理解和推理能力的加持下，实现更加自然、连贯且符合儿童认知特点的新闻改写。本框架通过解耦风险检测与改写优化，实现高效的并行处理，大幅降低了时间开销。实验表明，我们的框架大幅提升了改写的质量与可控性，显著提升了儿童新闻改写的实时性与可扩展性。基于该框架，我们构建了儿童新闻风险数据集，用于进一步优化风险检测能力，为未来儿童友好型新闻内容生成提供高质量的数据支持。该方法不仅提高了风险检测和改写的效率，还突破了当前儿童新闻改写中价值观引导不足的瓶颈。

本文的贡献：第一，提供了一种高效、可扩展的儿童互联网内容风险管理方案。该方法突破了现有儿童新闻改写中风险控制不足、价值观引导缺乏和实时生成困难的瓶颈，为儿童友好型新闻生成提供了一种可推广的技术路径。第二，构建了高效的风险检测机制。通过多维度风险识别，避免传统方法中过度过滤或误杀的问题，实现对儿童不适宜内容的精准控制。第三，构建儿童新闻风险数据集。通过统一的构造流程，我们建立了儿童新闻风险数据集，为未来的风险检测优化和安全内容生成提供高质量的数据支持。

## 2 相关工作

大模型的安全性研究指防止模型生成有害、不当或恶意内容，并提高模型对有害输入的鲁棒性。安全性研究通常涵盖对模型输出的实时监控、风险检测和行为调控，确保模型在实际应用环境中具备稳定性和可靠性 (V. Rathod et al., 2025)。对于有害输入的对抗性，Zhao (2025)通过数据增强和负偏好优化（NPO），提高模型在分布外（OOD）场景中的泛化能力。通过奖励机制强调关键拒绝token，进一步提高模型对越狱攻击的抵抗力。在风险检测领域，安全护栏模型结合了数据驱动的分类学习和基于概率图模型（PGM）的逻辑推理，能够有效识别并拦截不安全内容。Li et al. (2024)通过生成结构化的“风险-收益树”，提供了一种既可解释又可调节的内容审核系统。Han et al. (2024)结合了综合安全检测、上下文增强以及错误修复机制，从多个维度提升LLM 的安全性和响应质量。同时，Bibhu Dsah et al. (2024)提出一种以用户为中心的方法，该方法基于零样本学习，从输入和输出两端筛查并过滤恶意文本。设计一个结合LLMs、SGD和优化控制的框架，用于在存在潜在风险的情况下高效学习和执行复杂任务。在实时监控领域，Zhang et al. (2024)已经延伸至LLM 在交互环境中的行为安全性，Xie et al. (2024)探索了在线安全分析方法，即在LLM 生成输出的过程中实时监测和评估系统行为，以便及时发现并阻止潜在的安全风险。Kurian (2024)重点提出AI 系统在设计时需特别关注儿童用户的特殊需求与脆弱性，确保其交互体验既安全又适宜。

大模型对齐主要指确保大模型的行为与预期目标、价值观和规范一致的过程。现有的对齐技术包括监督微调（SFT）、基于人类反馈的强化学习（RLHF）和原则驱动的集成方法，使模型在面对各种输入时能够产生符合用户需求和伦理标准的输出 (Wang et al., 2024)。然而，这些方法存在局限性，如手动编写规则的不精确性和不全面性，以及模型缺乏安全训练导致的风险感知不足。Yao et al. (2023)提出了一种基于人类基本价值观的新范式，并以Schwartz的基本价值观理论为基础，构建了一个名为FULCRA的数据集，用于研究LLMs的行为与基本价值观之间的关系。国内相关工作也构建了中文核心价值-行为体系，标注了包含主体行为的文本句，构建了细粒度中文价值-行为知识库,并提出了价值观类别判别、方向判别及联合判别任务 (刘鹏远et al., 2024)。Raoul (2024)设计一个能够自动从自然语言文本中识别和提取上下文特定个人价值观的工具，以支持人道主义组织的决策过程，利用LLMs和基于字典的方法分别提取文本中的个人价值观，并结合这两种方法以提高提取的准确性和召回率。Luo et al. (2024)提出Guide-Align方法，通过构建一个全面且详细的指南库和相应的检索模型，确保LLMs的输出与人类价值观一致。Wang et al. (2024)通过在不同的2D网格世界环境中测试，验证LLM能够作为直接的奖励信号，指导RL代理避免负面副作用并安全探索。在新闻领域，现有的新闻价值提取方法主要依赖于人工标注，这限制了研究的规模和应用范围。Alicja Piotrkowicz et al. (2024)开发一种自动化的新闻价值提取方法可以支持更大规模的新闻学研究。提出一种能够从新闻标题中自动提取新闻价值的方法。Liu et al. (2024)将隐式价值函数应用于标记级别采样，显式价值函数应用于块级别束搜索，可以充分利用它们各自的优点，提高模型的整体对齐性能。然而，现实应用中需要对这些对齐的LLMs进行进一步微调以适应特定领域，这可能导致模型的安全性下降。此外，现有研究对LLMs内部参数如何维持安全性尚不明确，且冻结部



(c) 2025 中国中文信息学会计算语言学专业委员会

分神经元的方法未能有效防止安全性退化。Li et al. (2025)提取每层隐藏层的输出向量，计算正常-正常查询对和正常-恶意查询对的余弦相似性，观察不同层的分布差异，确认安全层的存在，在微调过程中冻结安全层的参数，仅更新安全层之外的参数。

## 3 数据集构建

本研究的数据主要来源于THUCNews新闻文本分类数据集 (孙茂松et al., 2016)。THUCNews由清华大学NLP研究组基于新浪新闻RSS订阅频道2005年至2011年间的历史数据，经过筛选与过滤后整理而成，包含74万篇新闻文档。在此基础上，本文选取了其中的体育、时政和娱乐三个新闻类别，以保证数据的多样性和代表性。为确保实验数据的均衡性与随机性，本文在上述三个类别中均匀随机抽取了一万篇新闻，构建基础数据集，以支持后续的风险检测与儿童友好性改写研究。为了科学界定新闻文本中的潜在风险，本文参考了《未成年人保护法》与《未成年人网络保护条例》中关于儿童网络不良信息的规范，从词汇、事件、标题和价值观四个维度制定了详细的风险标准，具体见附录B。这些标准涵盖了可能对儿童心理或价值观产生不良影响的内容，如带有暴力、恐怖、低俗、误导性或不适宜儿童理解的表达。为了进一步评估大模型在不同类型新闻文本上的风险管理能力，本文对新闻文本的情感倾向进行了分类。基于负面新闻的定义，我们首先使用GPT4模型进行新闻的情感分类，结合人工审核方式，对新闻文本进行情感属性标注，最终构建了一个包含1000篇新闻的测试数据集，其中包括：300篇正面新闻（内容积极、鼓舞人心、无明显负面情绪）、300篇中性新闻（信息客观中立，未包含强烈情感倾向）、400篇负面新闻（涉及社会冲突、暴力、不安情绪等内容）。这一分类有助于分析不同类型新闻在风险识别与改写过程中的表现差异，并进一步优化多模型协同架构，以实现更加精准的儿童新闻改写。

我们在正面、中性和负面新闻集上，分别采用讯飞星火-v3.5、ERNIE-3.5和GLM-4模型进行了儿童版本改写实验，使用统一的提示模板，以评估模型在改写过程中对内容风险的抵御能力。结果显示，模型在负面新闻集上的表现问题最为突出，频繁出现风险词汇、风险事件、风险标题以及价值观缺失等现象；相比之下，在正面和中性新闻集上的改写质量较高，风险表现相对可控。进一步分析表明，负面新闻本身包含更多潜在风险内容，其中娱乐领域的风险内容和价值观缺失的现象最为突出（见表1），原因是负面新闻中不良信息的出现概率显著高于正面与中性新闻，使得模型在价值观挖掘与风险过滤方面面临更大挑战，表现出明显的处理困难，因此后续实验主要使用负面新闻数据集。

| | 风险内容 | | | 风险价值观 | | |
|---|---|---|---|---|---|---|
| **新闻类别** | **讯飞星火-v3.5** | **ERNIE-3.5** | **GLM-4** | **讯飞星火-v3.5** | **ERNIE-3.5** | **GLM-4** |
| 政治 | 13 | 10 | 4 | 7 | 4 | 3 |
| 娱乐 | 34 | 58 | 42 | 23 | 34 | 20 |
| 体育 | 27 | 18 | 17 | 14 | 20 | 17 |
| **总和** | 74 | 86 | 63 | 44 | 58 | 40 |

Table 1: 负面新闻在不同类别下的风险内容和风险价值观分布

## 4 方法

### 4.1 总览

儿童互联网新闻风险管理与价值观引导框架（CRV-LLM）如图2所示。本文中，我们提出的基于多模型协同的儿童互联网新闻风险管理与价值观引导框架，主要分为三个阶段。首先，微调的轻量化模型对输入的新闻文本进行多维度的风险监测，分析新闻是否存在不适合儿童阅读的的风险点，并基于新闻的内容和风险点给出相应的改写建议或替代方案；同样的输入给到价值观引导模型，模型通过分析新闻的价值观的完整性，检测新闻是否缺失价值观，基于新闻提供具有教育意义的正面价值观。协同模型得到原新闻和来自不同维度的风险建议与价值观强化段落，最终输出生成符合要求的儿童版新闻。本框架采用模块化设计，将风险检测与改写任务解耦。四个轻量化检测模型基于DeepSeek-R1-Distill-Qwen-7B微调（下文简称DeepSeek-7B），参数量仅为基线模型的21.9%，通过并行推理实现高效风险筛查；改写模型则专注于语义重构，避免冗余计算。此设计大幅降低了端到端处理的时延与资源消耗。



(c) 2025 中国中文信息学会计算语言学专业委员会

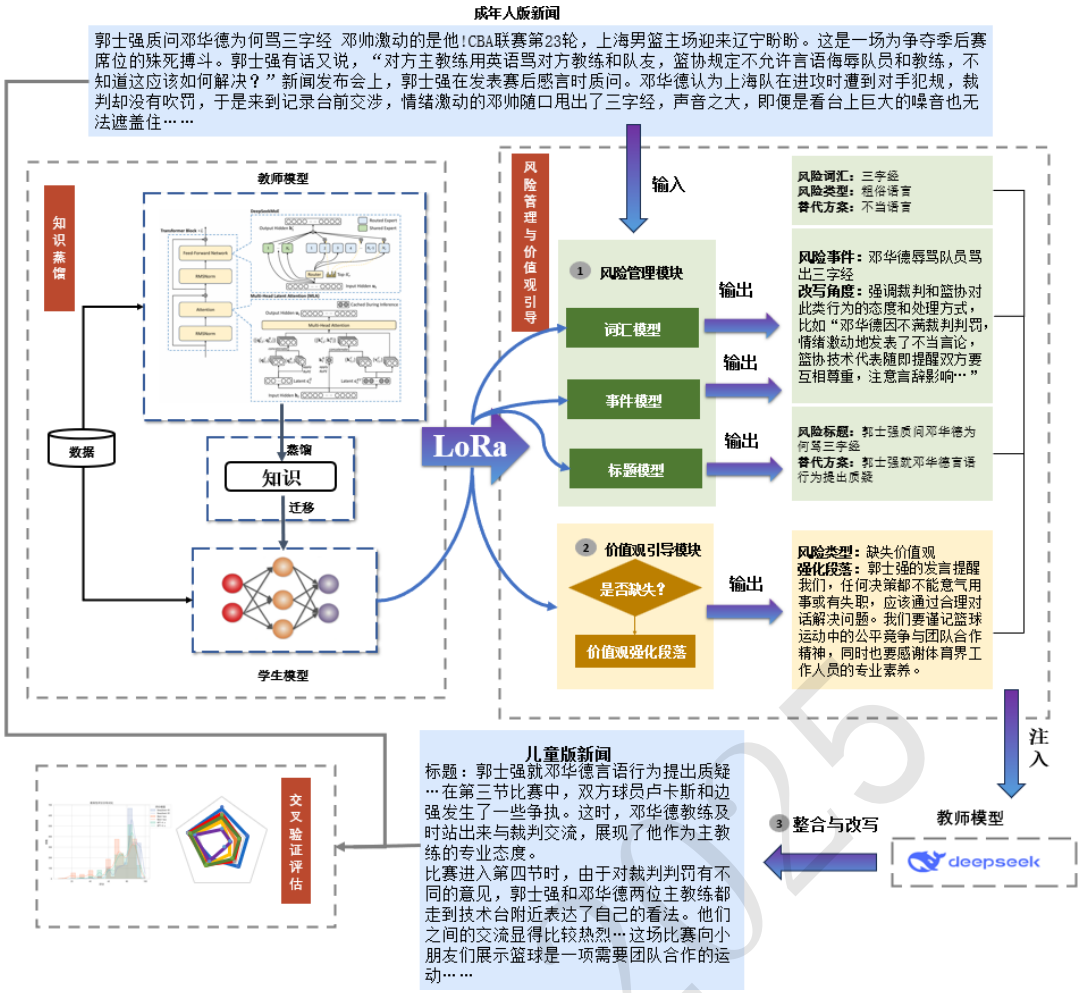



Figure 2: 框架流程图

本节先介绍指令数据集的构建过程，见第4.2节，接着描述风险识别大语言模型的微调过程，见第4.3节。

## 4.2 指令数据集构造

我们已有基础的新闻数据集，在此工作上，本文进一步针对不同维度的风险检测模块，包括词汇、事件、标题和价值观定制了四套个性化问卷框架，以便在数据挖掘过程中精准提取高质量的风险数据，并为大模型的优化提供有力支持。该问卷框架不仅考虑到风险内容的显性特征，如特定敏感词汇或极端表达，还涵盖了隐性风险，如信息误导、潜在负面暗示等，确保风险识别的全面性和准确性，具体问卷框架见附件C。问卷框架遵循思维链的思考路径，逐步激活思维链，将整个流程分为三个阶段：风险识别、深度分析和解决方案。

1. 风险识别：逐句扫描，根据不同维度的风险定义，识别出存在的所有可能的风险点。
2. 深度分析：对风险点进行分类、评分，二次确定风险点。
3. 解决方案：生成可落地的改造建议，对所有风险点提供替代方案和适宜性分析。

我们从THUCNews 数据集中额外抽取了1万篇新闻，通过本地部署的DeepSeek-32B教师模型，对这批文本分别进行了四轮风险挖掘。经自动清洗与人工审核后，初步形成了针对各维度的基础指令数据集。在此基础上，为了保证小模性之间的关联性以及让小模型充分学习并复现教师模型的思考路径，我们提出了一种“统一构造、多维关联”的数据集生成流程,如图3所示：

1. 串行化构造：所有新闻样本先后依次通过“风险识别—深度分析—解决方案”三个环节，每一步均注入思维链提示（Chain-of-Thought），保持样本标注流程的连贯性与一致性；



(c) 2025 中国中文信息学会计算语言学专业委员会

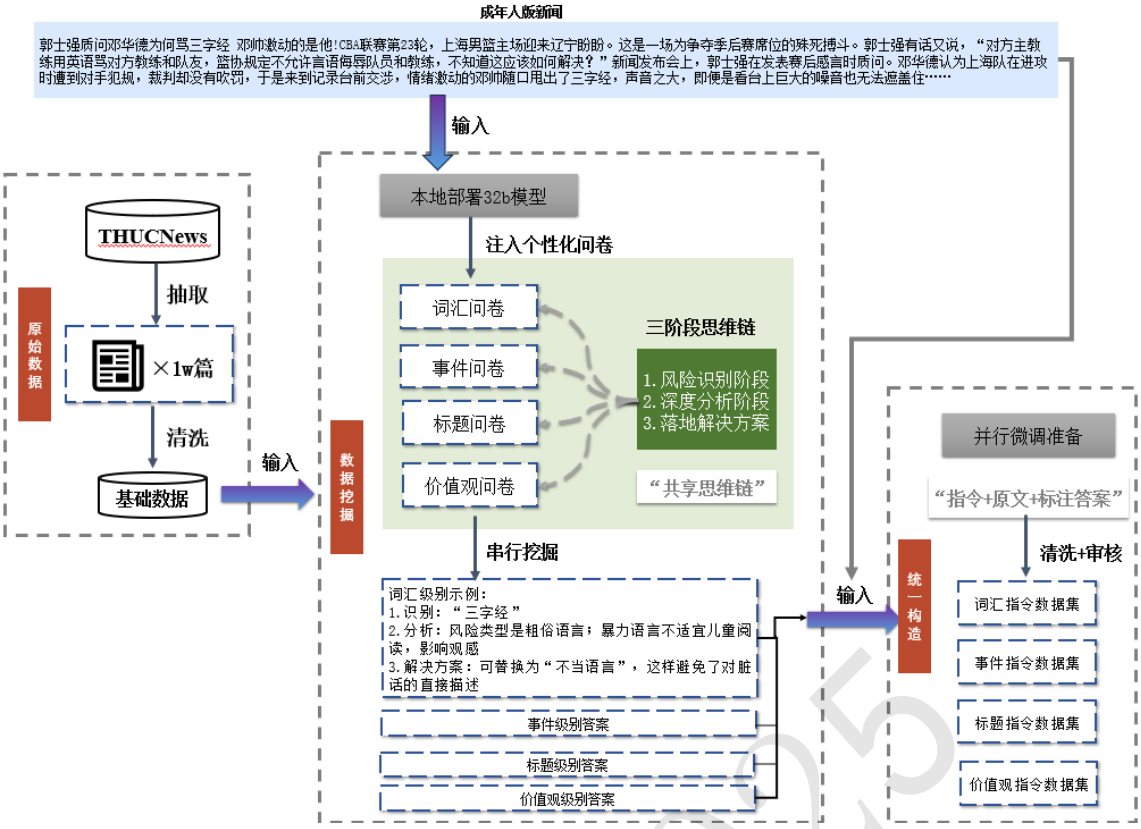



Figure 3: 指令数据集构造流程图

2. 跨维度关联：通过统一的样本标识（UUID）机制，确保同一篇新闻在不同风险维度下的“原文—问卷对”天然关联，无需为各模块重复构造独立数据集，从根本上杜绝了数据孤岛与标签冲突；

3. 并行微调：在串行化构造完成后，根据关联标签一次性切分出四个子数据集，用于并行训练四个小模型。此设计大幅提升了数据准备与模型训练的整体效率，同时保证了各模型微调数据源的一致性与可比性。

凭借这一统一而高效的构建方法，我们不仅保留了各风险检测模块间的内在关联，更显著降低了数据预处理和调优的成本，为多模型协同优化提供了坚实的数据支撑。

### 4.3 指令微调

DeepSeek-7B被用来作为微调四个风险模型的基座，使用Alpha格式的指令数据集，结合LoRA (Hu et al., 2021)技术对DeepSeek-7B模型进行微调。LoRA的原理是在模型的原始权重上添加了低秩调整，我们将LoRA应用于自注意力模块中的所有查询/键/值/输出投影矩阵。在本实验中LORA具体设置rank=8，初始学习率为5.0e-05，使用cosine学习率调度器和adamwtorch优化器，每个模型训练5 轮。

## 5 实验与结果

### 5.1 实验设计

为了全面评估所提出框架的有效性与适用性，本文选取了四种当前主流的大型语言模型（LLM）作为实验模型，分别为：讯飞星火-v3.5、ERNIE-3.5、hunyuan-turbo-latest 以及Baichuan3-Turbo。通过多模型对比分析，旨在检验该框架在多种生成体系下的稳定性与泛化能力。所有模型在实验中统一设定温度参数为0.5，以确保生成结果的一致性与可控性。此外，本文同步引入DeepSeek 系列模型进行对比测试，涵盖32B 模型与7B 模型，其中32B 模型被设定为框架的基线模型，以衡量整体性能提升。



(c) 2025 中国中文信息学会计算语言学专业委员会

实验设计方面，本文共构建三个子实验，分别从改写质量效果、风险管理模块效果、系统推理效率三个关键维度，对框架进行系统评估：

改写效果评估：采用交叉验证机制，从内容安全性、教育引导性、内容适宜性、信息完整性与语言连贯性五个维度对改写结果进行专家评审，全面评估框架生成文本的质量 (Chang et al., 2024)。该实验选取了三种当前领先的大型语言模型（LLM）作为评价模型。这些模型包括：GPT-4.o、Deeepseek-R1、Qwen-max。使用的数据集包括两个部分：D包含400篇儿童新闻文本，以及D中与儿童新闻文本主题内容相对应的400篇成年人新闻文本。考虑LLM生成文本的随机性，我们将文本传输3次，并计算最终的平均结果。

风险管理模块评估：为定量评估风险检测模块的性能，本文引入了风险回避率（Risk Avoidance Rate, RAR）和改写质量率（Rewriting Rate, RR）两个指标，分别衡量模型对潜在有害内容的检测和规避能力与检测到高风险内容后的改写表现。N为文本总数，对于第i篇文本，$R_i$为真实存在的风险点数量，$D_i$为模型检测出的风险点数量，$F_i$为检测出的风险点中，未被有效改写的数量。

$$\mathrm{RAR} = \frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{R_i} \tag{1}$$

$$\mathrm{RR} = \frac{1}{N}\sum_{i=1}^{N}\frac{F_i}{D_i} \tag{2}$$

效率评估：通过对比框架与全线使用DeepSeek-32B分别在推理阶段的总耗时（IL）与显存使用率指标，分析其在实际部署中的资源优化能力与执行效率。

通过以上多维度综合实验，本文验证了所提框架在儿童互联网新闻改写任务中的可控性、实用性与拓展性，为后续模型安全性与价值导向优化提供理论与实践依据。

### 5.2 实验结果与分析

**改写质量评估结果**:

| 模型名称 | 安全性 | 教育引导性 | 适宜性 | 信息完整性 | 连贯性 |
|---|---|---|---|---|---|
| CRV | **91.0** | **86.0** | 86.3 | 76.5 | **91.6** |
| 讯飞星火-v3.5 | 84.2 | 75.3 | 78.6 | 76.3 | 88.1 |
| Baichuan3-turbo | 88.2 | 70.5 | 82.7 | 79.8 | 91.5 |
| hunyuan-turbos-latest | 84.0 | 73.3 | 78.0 | **80.1** | 88.8 |
| ERNIE-3.5 | 89.7 | 74.7 | 72.3 | 76.3 | 86.2 |
| DeepSeek-R1-Distill-Qwen-32B | 86.6 | 80.2 | 81.0 | **80.1** | 90.0 |

Table 2: 三个评价模型五个维度上对生成文本的综合评估结果

通过对比表2与图4的实验结果可以发现，CRV-LLM框架在儿童新闻改写任务中展现了显著的多维度优势。首先，其安全性得分高达91.0，领先其他模型3–7分，表明CRV在过滤涉险词汇、弱化负面描写方面最为精准。其次，CRV的教育引导性以86.0分位居首位，远超其他模型的73–81分区间，说明其价值观引导模块能够有意识地植入励志、包容等正面内容。CRV在适宜性维度上也优于同类模型，体现了其语言风格和叙事节奏对儿童认知特点的高度契合。虽然CRV在信息完整性上略逊于某些模型，但这种因安全与教育考量而有意淡化细节的权衡，正是其确保阅读安全与价值导向的核心所在。最后，CRV以91.6分的连贯性成绩，证明其在删减与重构过程中依旧能够保持文本的自然流畅与逻辑通顺。综上所述，CRV-LLM通过风险管理与价值观引导的有机协同，不仅实现了文本安全性的最高保障，更在正向教育效果与阅读体验上取得了全面领先。

为了验证三个评价评价模型具有统计一致性，我们采取了随机抽样的方法，从每个测试模型中随机抽取了50篇文本，总计300篇，使用ICC对每个维度的评分进行一致性评估，辅以相关分析与非参数方差检验来验证评分稳定性，实验结果在附录F。根据实验结果，五个维度的一



(c) 2025 中国中文信息学会计算语言学专业委员会

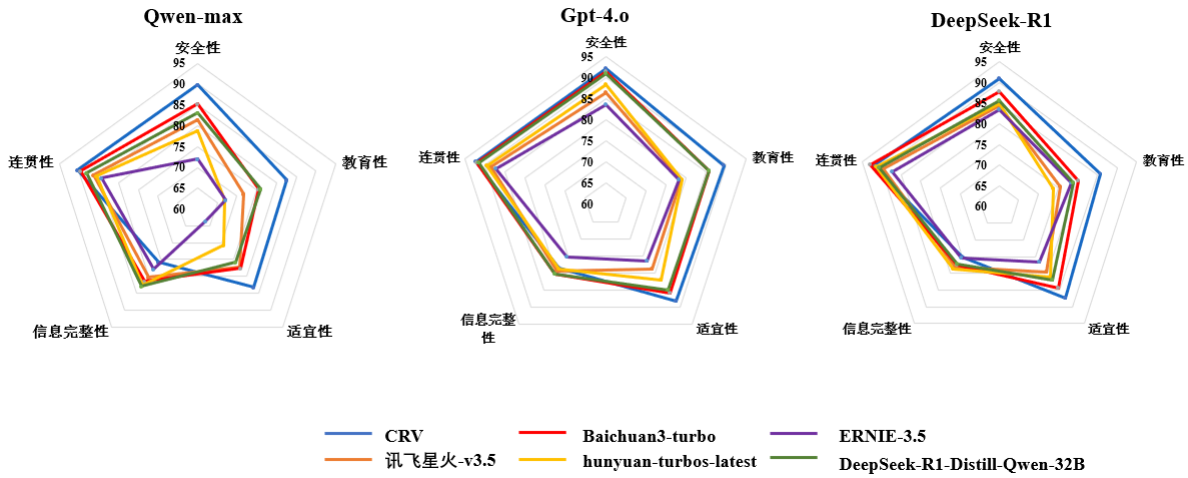



Figure 4: 多维度实验评估结果雷达图

致性相关系数均大于0.6，Spearman等级相关性在所有维度上均显示出显著的正相关性，且各模型在所有维度上的评分分布趋势显示出较高一致性。Friedman检验结果显示，在某些维度比如连贯性上模型的分布存在显著差异，表明各评价模型在设计时对不同文本的特点的关注重点不同，比如，DeepSeek-R1在教育性和适宜性方面表现均衡，而GPT-4.o在连贯性和适宜性方面更为敏感。但整体而言，这些差异并不影响模型评分的一致性，各模型在所有维度上的评分分布趋势显示出较高的一致性，进一步证明了模型在评价任务上的可靠性。

综合来看，CRV-LLM之所以在儿童新闻改写评估中表现卓越，首先得益于其“双模块”架构：风险管理模块依托，能够对原文中的潜在负面或煽动性词句、事件进行精准识别与过滤；价值观引导模块则通过嵌入“励志”“包容”“团队合作”等正向提示，确保内容既安全又富有教育意义。其次，CRV采取多模型融合策略，将擅长安全审查的子模型与精于生成润色的子模型有机串联，形成“筛选—建议—优化”闭环；与此同时，通过关键词、句式和段落三级细粒度控制，CRV能够针对不同上下文灵活调整过滤与替换强度。再者，框架在语言风格上面向儿童认知特点进行专门调优，严格把控句长、词汇难度与叙事节奏，以提升趣味性和可读性，使得模型更好地理解和适应儿童新闻文本生成任务与安全控制任务的需求，从而实现了安全性、教育性与连贯性的全面最优。

**风险管理模块评估结果:**

| 模型名称 | RAR | | | | RR | |
|---|---|---|---|---|---|---|
| | 词汇 | 事件 | 标题 | 价值观 | 标题 | 价值观 |
| CRV | 0.2626 | 0.1900 | **0.0197** | **0** | **0.0201** | 0.0754 |
| DeepSeek-r1-32b | **0.2835** | 0.1322 | 0.1578 | 0.0471 | 0.0703 | **0.0462** |
| DeepSeek-r1-7b | 0.5223 | 0.2396 | 0.1513 | **0.1069** | **0.2558** | **0.4718** |
| 讯飞星火-v3.5 | 0.3134 | **0.4942** | **0.1776** | 0.0597 | 0.0880 | 0.0635 |
| ERNIE-3.5 | 0.2388 | 0.1157 | 0.0328 | 0.0503 | 0.2721 | 0.1059 |
| hunyuan-turbos-latest | **0.0149** | **0.0991** | 0.0592 | 0.0094 | 0.1328 | 0.0506 |
| Baichuan3-Turbo | 0.0298 | 0.1042 | 0.0592 | 0.0440 | 0.0629 | 0.0526 |

Table 3: 生成文本在RR和RAR指标下的评估结果

在本研究中，我们分别针对词汇、事件、标题和价值观四个模块计算了风险回避率（RAR），并重点考察了标题与价值观两个维度的改写质量率（RR）。如图表3所示，框架在RAR指标中，标题与价值观两个维度的得分最低，表明系统对这两个维度的风险检测能力最为敏锐。特别是价值观维度，识别率达到百分百，充分展示了框架在识别潜在价值观偏误方面的出色性能。在词汇与事件维度的RAR评估中，混元大模型表现最优，分别达到0.0149与0.0991。与之相比，本框架在这两个维度上的RAR值虽略高于混元模型，但相较基线模型与7B模型均有明显提升，且与讯飞星火和ERNIE-3.5等通用模型的表现相近。这表明，本框架的整体风险识别能力并不逊色于主流通用模型。在RR指标方面，框架在标题改写维度



(c) 2025 中国中文信息学会计算语言学专业委员会

表现最为出色，RR值为0.0201，在所有对比模型中得分最低，表明其不仅成功识别高风险标题，还能提供准确、稳妥的改写建议。同时，框架在价值观改写维度的RR得分也与表现最优的32B模型非常接近，进一步印证其在确保高风险识别率的同时，具备较高的改写质量保障能力。

综上所述，本框架在多维度的风险识别与干预中均展现出稳健的性能，尤其在标题与价值观两个关键领域，兼具高识别率与高改写质量，体现出其在儿童新闻风险控制场景中的实际应用价值与优势潜力。

**效率评估结果**: 本实验中，所有模型均在NVIDIA NVIDIA L40S（46GB）GPU 上测试，使用半精度（FP16）推理。此外，我们未使用FlashAttention 等特定加速技术，而是采用常规库提供的高效实现。实验结果如图表4所示，在推理效率方面，本框架在处理一篇新闻时，从风险识别、建议生成到改写完成儿童版文本的平均用时为15.6秒，显著优于采用32B模型进行全流程处理所需的41.12秒。在显存占用方面，框架的平均占用为312.3 GB，约为DeepSeek-32B模型的708.5 GB的百分之44，明显低于后者。

| 模型名称 | IL(s) | GPU(G) |
|---|---|---|
| CRV | 15.6 | 391.5 |
| DeepSeek-R1-32B | 41.12 | 708.5 |

Table 4: 模型推理效率对比

该结果表明，在保证文本安全性与改写质量的同时，本框架在推理速度与资源效率上均具备明显优势，展现出其在实际部署与大规模应用场景中的良好适应性。CRV-LLM的加速优势源于两方面：其一，轻量化检测模型通过低秩适配（LoRA）减少参数规模，单模型推理速度提升3.2倍；其二，模块化设计支持风险检测与改写任务并行执行，较串行流程缩短60%时延。

### 5.3 消融研究

在探索不同模块对儿童新闻文本生成安全性的影响时，我们进行了消融实验，评估了风险词汇、风险事件、风险标题和价值观引导模块在独立及组合使用时的效果，具体结果见附录D.1。实验结果显示，单模块在特定维度上较Baseline 有提升：词汇、事件和标题模块在安全性和适宜性上表现较优，而价值观模块在教育性上表现突出。组合模块方面，双模块和三模块组合在多数维度上带来了显著的性能提升，不同组合在各维度上各有侧重，如词汇+事件组合显著增强了安全性，词汇+价值观组合则在教育性上表现最佳。三模块协同实现最优均衡，随着模块数量的增加，模型在各维度上的均衡性和整体性能均显著增强。CRV 框架在五个评估维度上展现了最为均衡且卓越的性能，证明了完整模块整合对于最大化模型综合性能的关键作用。

附录D.2的分组柱状图进一步凸显了价值观模块在提升文本教育性方面的显著优势。无论是在单模块、双模块还是三模块组合中，包含价值观模块的组合在教育性评分上均优于不含价值观模块的组合，且随着模块数量的增加，文本的教育性评分也呈现出逐步提升的趋势，这再次证实了价值观模块在多模块协同工作中的核心地位。

综上所述，实验结果充分展示了多模块协同工作在优化文本生成中的关键作用。各模块在不同维度上发挥着独特且不可或缺的作用，而合理组合与协同这些模块则能显著放大其优势，为生成高质量的儿童新闻文本提供了坚实的架构基础。

### 5.4 案例分析

如图5所示，右侧为CRV框架改写的儿童版新闻，中间是讯飞星火-v3.5模型使用普通提示模板生成的版本。对比可见，CRV框架在儿童内容风险控制上优势显著。讯飞星火的文本保留了“想吐”“惊悚”等激烈词汇，如“队友表示想吐！”直接传递负面情绪，可能影响儿童或引发模仿。其标题《马里昂的“歪”手指，队友表示想吐！》更用引号强化话题感，有煽动猎奇之嫌，不仅缺乏教育意义，还可能引导儿童模仿类似夸张表达。相反，CRV框架有效规避了此类不宜表达。它将“队友觉得手指特殊”改为“引发大家的关注”，表述更客观中性,有效消除了原始文本中的负面情绪传导。其标题《马里昂手指轻度畸形引关注》准确、中立，使用“轻度”“引关



(c) 2025 中国中文信息学会计算语言学专业委员会

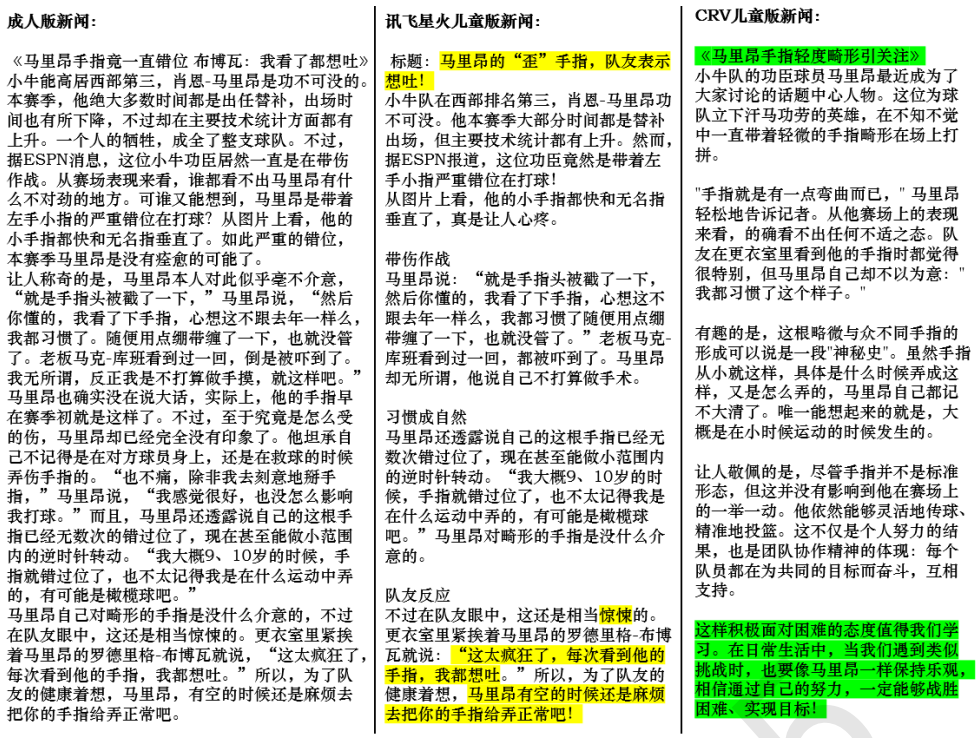



Figure 5: 成年人版新闻、讯飞星火-v3.5和CRV框架生成的儿童新闻示例

注”替代感官刺激词，使得标题在传达信息的同时，保持了中立理性的风格，更适配儿童接受度与心理发展阶段。原始新闻与讯飞星火改写中均详细描写了“队友的惊讶表情”“手指变形吓到队友”的具体细节，这种描写容易引发儿童的恐惧或嘲笑心理，甚至可能强化对身体差异的负面认知。CRV框架则有意弱化这些不利元素，不仅删去对马里昂手指状况的过度刻画，还通过补充背景信息让儿童理解“运动员可能遭遇伤病”，引导其用更理性、更包容的态度看待身体差异。此外，CRV文本特意加入如“这种积极应对困难的态度值得我们学习”“只要相信自己、努力，一定能克服困难，实现目标！”等正向总结段落。这不仅提升了新闻的教育意义，也帮助儿童在阅读中形成正确的价值判断。CRV框架通过“删减不良内容+重构正面引导+适龄化表达”的系统机制，显著提升了儿童新闻改写的价值观适配度和风险控制能力。相比之下，传统大模型直接改写仍容易保留原文中的风险词汇和表达习惯，缺乏针对儿童心理特征的精细打磨。

## 6 结论与未来工作

在本研究中，我们首次将大型语言模型系统性地应用于儿童新闻安全领域，提出了一种创新性的多模型协同框架——儿童互联网新闻风险管理与价值观引导（CRV-LLM）。该框架由两大核心模块构成：风险管理模块和价值观引导模块。经过实验我们发现风险管理模块能够高效过滤与管控潜在风险内容，显著提升生成文本的安全性，而价值观引导模块通过干预，能有效增强新闻文本的正面价值传递。实验结果表明，CRV-LLM依托细粒度控制与联合优化，不仅能精准理解并满足儿童新闻生成任务的需求，还在可读性、安全性和教育性等多项指标上实现了同步提升，在保证生成质量的前提下，推理效率较传统大模型提升62%，显存占用降低至44%。其模块化架构与轻量化设计，为儿童新闻的实时改写与大规模部署提供了高效、可扩展的解决方案。尽管框架已取得阶段性成果，但在复杂语境下的细粒度风险识别、隐含偏见检测及知识覆盖方面仍有不足；价值观引导策略亦需更好地适应多元文化与不同年龄段需求；此外，数据样本分布偏向负面可能在一定程度上影响情感泛化能力。未来工作将聚焦于知识的补充和整合以提升风险识别质量，以及构建更平衡的数据集以覆盖多种情感倾向新闻，同时打通权威知识库以强化内容准确性，研发个性化价值干预机制，并在更多真实场景中验证系统的通用性与稳定性。我们期望通过这些努力，能够进一步提升儿童新闻的教育价值，使其成为儿童学习和成长过程中的有益资源。



(c) 2025 中国中文信息学会计算语言学专业委员会

## 参考文献


Aho, Alfred V. and Corasick, Margaret J. 1975. *Efficient string matching: an aid to bibliographic search* Communications of the ACM.

Alicja Piotrkowicz., Vania Dimitrova., Katja Markert. 2024. *Automatic Extraction of News Values from Headline Text* Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 64–74, Valencia, Spain, April 3-7 2017. c2017 Association for Computational Linguistics.

Belmoukadam, O., Jonghe, J.D., Ajridi, S., Krifa, A., Damme, J.V., Mkadem, M., Latinne, P. 2024. *ADVERSLLM: A PRACTICAL GUIDE TO GOVERNANCE, MATURITY AND RISK ASSESSMENT FOR LLM-BASED APPLICATIONS* International Journal on Cybernetics and Informatics (2024): n. pag.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi,Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. *A survey on evaluation of large language models* ACM Trans. Intell. Syst. Technol., 15(3), mar.

Shanshan Han., Zijian Hu., Alay Dilipbhai Shah., Han Jin., Yuhang Yao., Dimitris Stripelis., Zhaozhuo Xu., Chaoyang He. 2024. *TorchOpera: A Compound AI System for LLM Safety* arXiv:2406.10847.

Nomisha Kurian. 2024. *'No, Alexa, no!': designing child-safe AI and protecting children from the risks of the 'empathy gap' in large language models* Learning, Media and Technology, 1–14.https://doi.org/10.1080/17439884.2024.2367052.

Shen Li., Liuyi Yao., Lan Zhang., Yaliang Li. 2025. *Safety Layers in Aligned Large Language Models: The Key to LLM Security* ICLR 2025 The Thirteenth International Conference on Learning Representations. https://openreview.net/forum?id=kUH1yPMAn7

Zhixuan Liu., Zhanhui Zhou., Yuanfu Wang., Chao Yang., Yu Qiao. 2024. *Inference-Time Language Model Alignment via Integrated Value Guidance* In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 4181–4195, Miami, Florida, USA. Association for Computational Linguistics.

Jing-Jing Li., Valentina Pyatkin., Max Kleiman-Weiner., Liwei Jiang., Nouha Dziri., Anne G. E. Collins., Jana Schaich Borg., Maarten Sap., Yejin Choi., Sydney Levine. 2024. *SafetyAnalyst: Interpretable, transparent, and steerable safety moderation for AI behavior* arXiv:2410.16665.

Prasanjit Rath, Hari Shrawgi, Parag Agrawal, and Sandipan Dandapat. 2025. *LLM Safety for Children* In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track), pages 809–821, Albuquerque, New Mexico. Association for Computational Linguistics.

Raoul BRIGOLA. 2024. *VIVE: AnLLM-based approach to identifying and extracting context-specific personal values from text* https://studenttheses.uu.nl/handle/20.500.12932/46894

Rosenblatt, F. 1958. *The perceptron: a probabilistic model for information storage and organization in the brain.* Psychological Review, 1958, 65:386-408.DOI:10.1037/h0042519.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. *Lora: Low-rank adaptation of large language models* arXiv preprint arXiv:2106.09685

Junfeng Jiao, Saleh Afroogh, Kevin Chen, Abhejay Murali, David Atkinson, Amit Dhurandhar. 2025. *LLMs and Childhood Safety: Identifying Risks and Proposing a Protection Framework for Safe Child-LLM Interaction* arXiv:2502.11242.

V. Rathod, S. Nabavirazavi, S. Zad and S. S. Iyengar. 2025. *Privacy and Security Challenges in Large Language Models* 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2025, pp. 00746-00752, doi: 10.1109/CCWC62904.2025.10903912.

Zhaoyue Wang. 2024. *Towards Socially and Morally Aware RL agent: Reward Design With LLM* arXiv:2401.12459.




(c) 2025 中国中文信息学会计算语言学专业委员会

Xinpeng Wang, Shitong Duan, Xiaoyuan Yi, Jing Yao, Shanlin Zhou, Zhihua Wei, Peng Zhang, Dongkuan Xu, Maosong Sun and Xing Xie. 2024. *On the Essence and Prospect: An Investigation of Alignment Approaches for Big Models* Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24).

Xuan Xie., Jiayang Song., Zhehua Zhou., Yuheng Huang., Da Song., Lei Ma. 2024. *Online Safety Analysis for LLMs: a Benchmark, an Assessment, and a Path Forward* arXiv:2404.08517.

Jing Yao., Xiaoyuan Yi., Xiting Wang., Yifan Gong. and Xing Xie. 2023. *Value FULCRA: Mapping Large Language Models to the Multidimensional Spectrum of Basic Human Values* arXiv:2311.10766v1

Yi Luo, Zhenghao Lin, YuHao Zhang, Jiashuo Sun, Chen Lin, Chengjin Xu, Xiangdong Su, Yelong Shen, Jian Guo, and Yeyun Gong. 2024. *Ensuring Safe and High-Quality Outputs: A Guideline Library Approach for Language Models* In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1152–1197, Mexico City, Mexico. Association for Computational Linguistics.

Zhao., Xuandong., Cai., Will., Shi., Tianneng., Huang., David., Lin., Licong., Mei., Song., Dawn. 2025. *Improving LLM Safety Alignment with Dual-Objective Optimization* arXiv:2503.03710.

Zhexin Zhang ,Shiyao Cui ,Yida Lu ,Jingzhuo Zhou ,Junxiao Yang ,Hongning Wang, Minlie Huang. 2024. *AGENT-SAFETYBENCH: Evaluating the Safety of LLM Agents* arXiv:2412.14470.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, Ji-Rong Wen. 2023. *A Survey of Large Language Models*

Du Xiaomeng, Yu Dong, and Liu Pengyuan. 2024. 文本样式和主题框架引导下的大模型辅助儿童新闻生成 In Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference), pages 150–170.

金丽莉2025. 少年儿童新闻传播的价值与策略 中国新闻出版广电报.2025.01.23.

崔洪振, 张龙豪, 彭云峰, 邬雯. 2024. 关键词提取算法研究综述 中文信息学报, 2024, 38 (02): 1-14+24.

刘鹏远, 张三乐, 于东, 薄琳. 2024. *CoreValue:* 面向价值观计算的中文核心价值-行为体系及知识库 中文信息学报. 2024, 38(11): 13-26

孙茂松，李景阳，郭志，赵宇，郑亚斌，司宪策，刘知远. 2016. *THUCTC*：一个高效的中文文本分类工具包

CCL 2025



(c) 2025 中国中文信息学会计算语言学专业委员会

# A 提示模板

## A.1 评估生成文本的提示模板

你是个儿童新闻文本编辑专家，请客观地从五个维度出发，参照成人版新闻，评估模型生成文本的质量。每个维度从1（最差）到100（最好）进行评分，其中每个维度的具体信息如下：

**内容安全性** 词汇风险规避：文本是否规避敏感词、暴力、恐怖、色情、歧视、污言秽语、迷信等词汇或表述。

事件内容安全性：文本描述事件是否规避儿童不宜接触的风险类型，例如暴力、色情、邪教、极端主义、分裂主义等。

标题表达合适性：文本标题是否避免使用耸动、惊悚、恐吓或夸张的表达方式或误导性、情绪煽动的内容。

**教育引导性** 正面价值观：文本是否传达正面情绪、社会责任感，如诚实、勇敢、同情心等。

行为引导：文本是否提供了实用的信息或建议，帮助儿童学习生活中的技能，鼓励儿童做出正确判断和行动。

**内容适宜性**

题材与语境适配：文本是否避免涉及复杂政治、伦理冲突、成人情节，符合儿童心理发展。

语言和知识难度匹配：文本是否使用常见词汇，语句简洁，是否避免过度抽象或术语密集。

**信息完整性**

事实保留：原文本的核心事件、人物、背景是否被准确重述。

结构还原：原文本的起因、经过、结果等因果关系是否保留。

**连贯性**

语法与用词正确：文本是否有语病、错别字、用词不当。

句式结构自然：文本是否有断裂、跳跃、重复句。

语气风格统一：文本是否符合儿童口吻，表达连贯、易读。

**原文本** = 成年人版新闻

**需要评估的文本** = 模型生成的文本

## A.2 国内模型改写新闻的提示模板

成人版新闻D：XXX

你是一位资深的儿童文学专家，现在有一个根据成年人版新闻D来生成一篇儿童版的新闻，请参考文本样式定义模块s。（注：确保内容真实准确，避免误导儿童，选择儿童感兴趣的话题和事件。）

文本样式模块s:

文本难度：适合小学的学生阅读

儿童新闻整体板块:

标题：简洁明了，控制在10-20 个字以内，避免使用复杂的句子和专业术语。

导语：建议导语长度控制在3-4 句话以内，凝练表达新闻的主要内容，使用生动形象的语言，激发孩子的阅读兴趣。

正文：可以分三个板块进行叙述，每个板块都有序号和小标题，每个板块都是完整的文本段落。

第二十四届中国计算语言学大会论文集，第402页-第421页，济南，中国，2025年8月11日至14日。

(c) 2025 中国中文信息学会计算语言学专业委员会

# B 风险标准

## B.1 相关政策法规

本文风险标准制定主要参考《未成年人保护法》与《未成年人网络保护条例》中关于儿童网络不良信息的规范。具体条例节选自《未成年网络保护条例》网络信息规范的第二十二条和第二十三条:

第二十二条：任何组织和个人不得制作、复制、发布、传播含有宣扬淫秽、色情、暴力、邪教、迷信、赌博、引诱自残自杀、恐怖主义、分裂主义、极端主义等危害未成年人身心健康内容的网络信息。

第二十三条：网络产品和服务中含有可能引发或者诱导未成年人模仿不安全行为、实施违反社会公德行为、产生极端情绪、养成不良嗜好等可能影响未成年人身心健康的信息的，制作、复制、发布、传播该信息的组织和个人应当在信息展示前予以显著提示。

## B.2 具体风险标准

**风险词汇定义:**
指危害未成年人身心健康的词语、短语或隐喻表达，这些词汇可能:

i 直接涉及淫秽、色情、暴力、邪教、迷信、赌博、引诱自残自杀、恐怖主义、分裂主义、极端主义等内容。

ii 诱导未成年模仿不安全、违反社会公德的行为、产生极端情绪以及养成不良嗜好等。

**风险事件定义:**
指危害未成年人身心健康的事件，这些事件可能:

i 直接涉及淫秽、色情、暴力、邪教、迷信、赌博、引诱自残自杀、恐怖主义、分裂主义、极端主义等内容。

ii 诱导未成年参与危险活动、实施违反社会公德行为、产生极端情绪、养成不良嗜好等。

**风险标题定义:**
指可能吸引未成年人点击并暴露其接触不良信息的标题，这些标题可能:

i 直接涉及淫秽、色情、暴力、邪教、迷信、赌博、引诱自残自杀、恐怖主义、分裂主义、极端主义等内容。

ii 具有猎奇性、耸动性或诱惑性，容易引起未成年人的好奇心、恐慌或误解。

**风险价值观定义:**
指可能扭曲未成年人正确价值观念的内容或行为导向，这些价值观可能:

i 导致未成年人形成错误的人生观、世界观和道德观，影响其健康成长。

ii 缺失教育性，淡化事情的严重性，强行美化错误行为。

CCL 2025

第二十四届中国计算语言学大会论文集，第402页-第421页，济南，中国，2025年8月11日至14日。

(c) 2025 中国中文信息学会计算语言学专业委员会

# C 问卷框架

| | | |
|---|---|---|
| **## 任务说明**<br>您需要以教育专家的身份，对原始新闻内容D进行词汇分析，并根据输出要求，输出结构化改造方案。必须严格遵循以下思考路径和改写原则：<br><br>【思维链激活】<br>1. 风险识别阶段：逐句扫描识别风险点<br>2. 深度分析阶段：对风险点进行多维评分<br>3. 解决方案阶段：生成可落地的改造建议<br>【改写原则】<br>1.替换词汇4不改变原始词汇含义<br>2.人名、电影名、专业术语等专有名词不在识别与改写范围<br><br>**## 处理流程**<br>**### 第一阶段：风险识别**<br><识别开始><br>[风险词汇定义]:<br>指危害未成年人身心健康的词语、短语或隐喻表达，这些词汇可能：<br>i. 直接涉及淫秽、色情、暴力、邪教、迷信、赌博、引诱自残自杀、恐怖主义、分裂主义、极端主义等内容<br>ii. 诱导未成年人模仿不安全行为、是是违反社会公德的行为、产生极端情绪以及养成不良嗜好<br>1. 风险定位：先逐句扫描，根据风险词汇定义，识别包含风险词汇的原文句子（字段：risk_sent) | 2. 输出逻辑:<br>IF 若所有词汇的使用语境适宜，则可跳过分析和建议阶段:<br>->输出类型：无风险模式输出<br>ELSE:<br>->完整执行原「深度分析+解决方案」流程<br></识别结束><br><br>**### 第二阶段：深度分析**<br><分析开始><br>1. 风险词汇分析<br>1.1 风险词汇（字段：risk_word)<br>1.2 判断风险类型（字段：risk_type）：<br>a.淫秽 b.色情 c.暴力 d.邪教 e.迷信 f.赌博 g.引诱自残自杀 h.恐怖主义 i.分裂主义 j.极端主义 k.其他（标注）<br>2. 风险等级判定（1分）（字段：risk_score)<br>风险词汇均判定为1分<br></分析结束><br><br>**### 第三阶段：解决方案**<br><建议开始><br>替代方案生成（需生成3个备选和适宜分析）<br>替代词1（字段：alt_word1）<br>适宜性分析(字段：reason)<br>替代词2（字段：alt_word2）<br>适宜性分析(字段：reason)<br></建议结束> | **## 输出要求**<br>思考后，必须严格遵循以下两种输出模式，保留所有字段：<br>【模式一：存在风险】<br>{{<br>"risk_status": "有风险",<br>"vocab_analysis": [<br>{{<br>"risk_sent": "",<br>"risk_word": "",<br>"risk_type": "",<br>"risk_score": 1,<br>"replacements": [<br>{{ "alt_word1": "", "reason": "" }},<br>{{ "alt_word2": "", "reason": "" }}<br>]<br>}}<br>],<br>// 更多风险点...<br>}}<br><br>【模式二：无风险】<br>{{<br>"risk_status": "无风险",<br>"check_result": "全文未发现不适宜儿童的风险词汇"<br>}} |

Figure 1: 词汇专项问卷框架

| | | |
|---|---|---|
| **## 任务说明**<br>您需要以教育专家的身份，对原始新闻内容进行事件分析，并根据输出要求，输出结构化改造方案。必须严格遵循以下思考路径和改写原则：<br><br>【思维链激活】<br>1. 风险识别阶段：逐句扫描识别风险点<br>2. 深度分析阶段：对风险点进行多维评分<br>3. 解决方案阶段：生成可落地的改造建议<br><br>【改写原则】<br>1.改写事件不改变原文含义<br>2.避免过于保护而丢失原文信息<br><br>**## 处理流程**<br>**### 第一阶段：风险识别**<br><识别开始><br>[风险事件定义]:<br>指危害未成年人身心健康的事件，这些事件可能：<br>i. 直接涉及淫秽、色情、暴力、邪教、迷信、赌博、引诱自残自杀、恐怖主义、分裂主义、极端主义等内容<br>ii. 诱导未成年人参与危险活动、实施违反社会公德行为、产生极端情绪、养成不良嗜好等 | 1. 风险定位：先逐句扫描，根据风险事件定义，识别包含风险事件的段落位置（字段：original_text_excerpt)<br>2. 输出逻辑:<br>IF 若文本不包含风险事件，则可跳过分析和建议阶段:<br>->输出类型：无风险模式输出<br>ELSE:<br>->完整执行原「深度分析+解决方案」流程<br></识别结束><br><br>**### 第二阶段：深度分析**<br><分析开始><br>1. 风险事件分析<br>1.1 事件解构，总结事件（字段：event_desc)<br>1.2 判断风险类型（字段：risk_type)<br>a.淫秽 b.色情 c.暴力 d.邪教 e.迷信 f.赌博 g.引诱自残自杀 h.恐怖主义 i.分裂主义 j.极端主义 k.其他（标注）<br>1.3 事件风险原因（字段：reason）<br>2. 风险等级评分：（1-3分，分数越高，风险越高）（字段：risk_score)<br>1分：可能引发误解但无直接危害<br>2分：隐含危险倾向或价值观误导<br>3分：直接违反法律或严重危害儿童身心<br></分析结束> | **## 输出要求**<br>思考后，必须严格遵循以下两种输出模式，保留所有字段：<br>【模式一：存在风险】<br>{{<br>"risk_status": "有风险",<br>"event_analysis": [{{<br>"original_text_excerpt": "",<br>"event_desc":"",<br>"risk_type":"",<br>"risk_score": ,<br>"reason": "" ,<br>"suspect_adjust":"",<br>"narrative_advice": "",<br>}}]<br>//更多风险点...<br>}}<br><br>【模式二：无风险】<br>{{<br>"risk_status": "无风险",<br>"check_result": "全文未发现不适宜儿童的风险内容"<br>}} |

Figure 2: 事件专项问卷框架

第二十四届中国计算语言学大会论文集，第402页-第421页，济南，中国，2025年8月11日至14日。

(c) 2025 中国中文信息学会计算语言学专业委员会

| | | |
|---|---|---|
| **## 任务说明**<br>您需要以教育专家的身份，对原始新闻内容D进行标题分析，并根据输出要求，输出结构化改造方案。必须严格遵循以下思考路径和改写原则：<br><br>【思维链激活】<br>1. 风险识别阶段：逐句扫描识别风险点<br>2. 深度分析阶段：对风险点进行多维评分<br>3. 解决方案阶段：生成可落地的改造建议<br><br>【改写原则】<br>1.改写标题不改变原标题含义<br>2.原标题含有的人名、电影名、地名等专有名词不在识别和改写范围内<br><br>**## 处理流程**<br>**### 第一阶段：风险识别**<br><识别开始><br>[风险标题定义]:<br>指可能吸引未成年人点击并暴露其接触不良信息的标题，这些标题可能：<br>i. 直接涉及淫秽、色情、暴力、邪教、迷信、赌博、引诱自残自杀、恐怖主义、分裂主义、极端主义等内容<br>ii. 具有猎奇性、耸动性或诱惑性，容易引起未成年人的好奇心、恐慌或误解<br>1. 风险定位：扫描标题，根据风险标题定义，识别风险标题（risk_title） | 2. 输出逻辑:<br>IF 若标题没有风险，则可跳过分析和建议阶段:<br>->输出类型：无风险模式输出<br>ELSE:<br>->完整执行原「深度分析+解决方案」流程<br></识别结束><br><br>**### 第二阶段：深度分析**<br><分析开始><br>1. 现状分析<br>1.1 原始标题风险点标注（字段：title_risk_points）<br>1.2 原始标题风险类型（risk_type）<br>a.淫秽 b.色情 c.暴力 d.邪教 e.迷信 f.赌博 g.引诱自残自杀 h.恐怖主义 i.分裂主义 j.极端主义 k.其他（标注）<br>2. 标题风险评分（risk_score）（1-2分）<br>1分：含有暗示性或引发误解的内容<br>2分：包含风险词汇<br></分析结束><br><br>**### 第三阶段：解决方案**<br><建议开始><br>优化方案（生成3个备选标题和适宜性分析）<br>替代方案生成（字段：alt_title）<br>适宜性分析（字段：reason）<br></建议结束> | **## 输出要求**<br>思考后，必须严格遵循以下两种输出模式，保留所有字段：<br>【模式一：存在风险】<br>{{<br>"risk_status": "有风险",<br>"title_optimization": {{<br>"risk_title":"",<br>"title_risk_points":"",<br>"risk_score":,<br>"risk_type":"",<br>"replacements": [<br>{{ "alt_title1": "", "reason": "" }},<br>{{ "alt_title2": "", "reason": "" }}<br>]<br>}}<br>}}<br><br>【模式二：无风险】<br>{{<br>"risk_status": "无风险",<br>"check_result": "标题未发现不适宜儿童的风险内容"<br>}} |

Figure 3: 标题专项问卷框架

| | | |
|---|---|---|
| **## 任务说明**<br>您需要以教育专家的身份，对原始新闻内容D进行价值观分析，并根据输出要求，输出结构化改造方案。必须严格遵循以下思考路径和改写原则：<br><br>【思维链激活】<br>1. 风险识别阶段：逐句扫描识别风险点<br>2. 深度分析阶段：对风险点进行多维评分<br>3. 解决方案阶段：生成可落地的改造建议<br><br>【改写原则】<br>改写价值观与原文相符<br><br>**## 处理流程**<br>**### 第一阶段：风险识别**<br><识别开始><br>识别逻辑：先逐句扫描，<br>IF 新闻存在价值观传递：<br>[风险价值观定义]:<br>指可能扭曲未成年人正确价值观念的内容或行为导向。这些价值观可能<br>i. 导致未成年人形成错误的人生观、世界观和道德观，影响其健康成长<br>ii. 缺失教育性，淡化事情的严重性，强行美化错误行为<br>1. 风险定位：根据风险价值观定义，识别包含风险价值观的段落位置（risk_value） | 2. 完整执行原「深度分析+解决方案中的矫正」流程，输出类型：风险模式二输出<br>ElSE IF 新闻仅描述事实，没有明显的价值观传递：<br>->执行原「解决方案」中的强化流程<br>->输出类型：风险模式一输出<br></识别结束><br><br>### 第二阶段：深度分析<br>#### 价值观分析<br><分析开始><br>价值观质量分析<br>1. 判断偏移类型(risk_type)<br>a.不良价值观 b.教育性缺失<br>2. 价值观偏移检测（deviation）(1-3分):<br>1分：新闻体现了价值观，但和原文想传递的含义有差异<br>2分：新闻传递了部分正面价值观，但有偏向性，例如淡化事情的严重性或对犯罪嫌疑人的人格化描述，可能会误导儿童<br>3分：新闻中的价值观，偏向消极，影响儿童的身心成长<br></分析结束><br><br>**### 第三阶段：解决方案**<br><建议开始><br>1. 强化<br>增加3-4句话、与原文相应的价值观段落（字段：strengthen） | 2. 矫正<br>2.1 风险原因分析（reason）<br>2.2 参开以下角度，给出修正方案<br>i.不良价值观纠正，传递正面价值观（字段：correction）<br>ii.教育延伸，使原文价值观更严肃客观（字段：education）<br></建议结束><br><br>**## 输出要求**<br>思考后，将改造方案严格按以下JSON格式输出，保留所有字段：<br>【模式一：风险模式一】<br>{{<br>"risk_status": "风险一",<br>"value_exist": {{<br>"integrity": 1,<br>"risk_type":"缺失价值观",<br>"strengthen":""<br>}}<br>}}<br>【模式二：风险模式二】<br>{{<br>"value_deviation":{{<br>"risk_value":"",<br>"risk_type":"",<br>"reason":"",<br>"deviation":,<br>"correction":"",<br>"education":""<br>}}<br>//更多风险点...<br>}} |

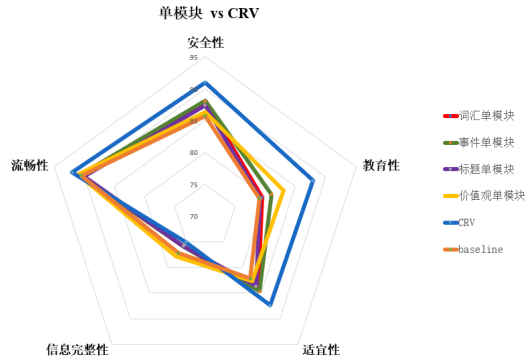
Figure 4: 价值观专项问卷框架



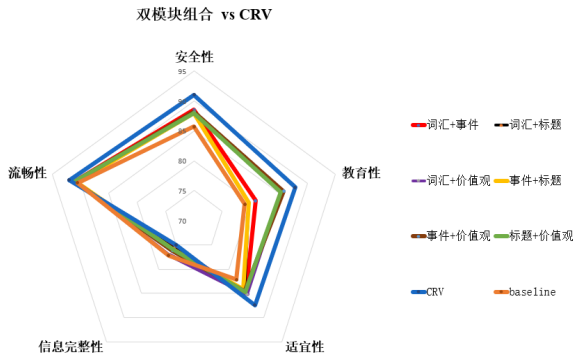
(c) 2025 中国中文信息学会计算语言学专业委员会

# D 消融实验的评估结果

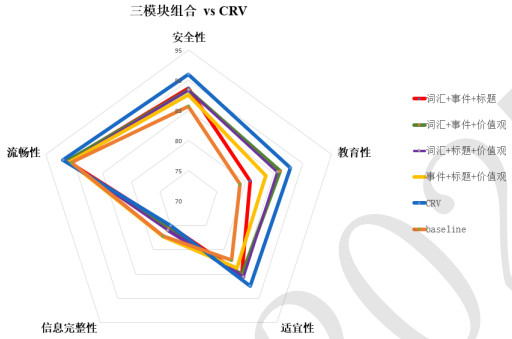
## D.1 各模块评估雷达图




(a) 单模块在五维度上的雷达图




(b) 双模块在五维度上的雷达图




(c) 三模块在五维度上的雷达图

Figure 5: 各模块组合在多维度上的评估结果雷达图

## D.2 价值观模块评估结果

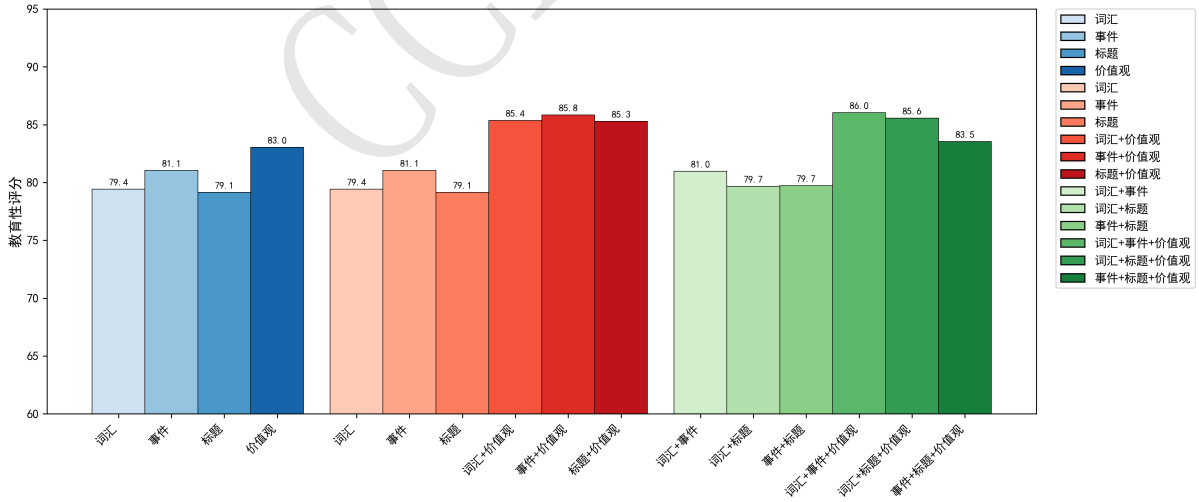



Figure 6: 有无价值观模块在教育性上的评估结果分组柱状图

CCL 2025



(c) 2025 中国中文信息学会计算语言学专业委员会

## D.3 各模块评估热力图

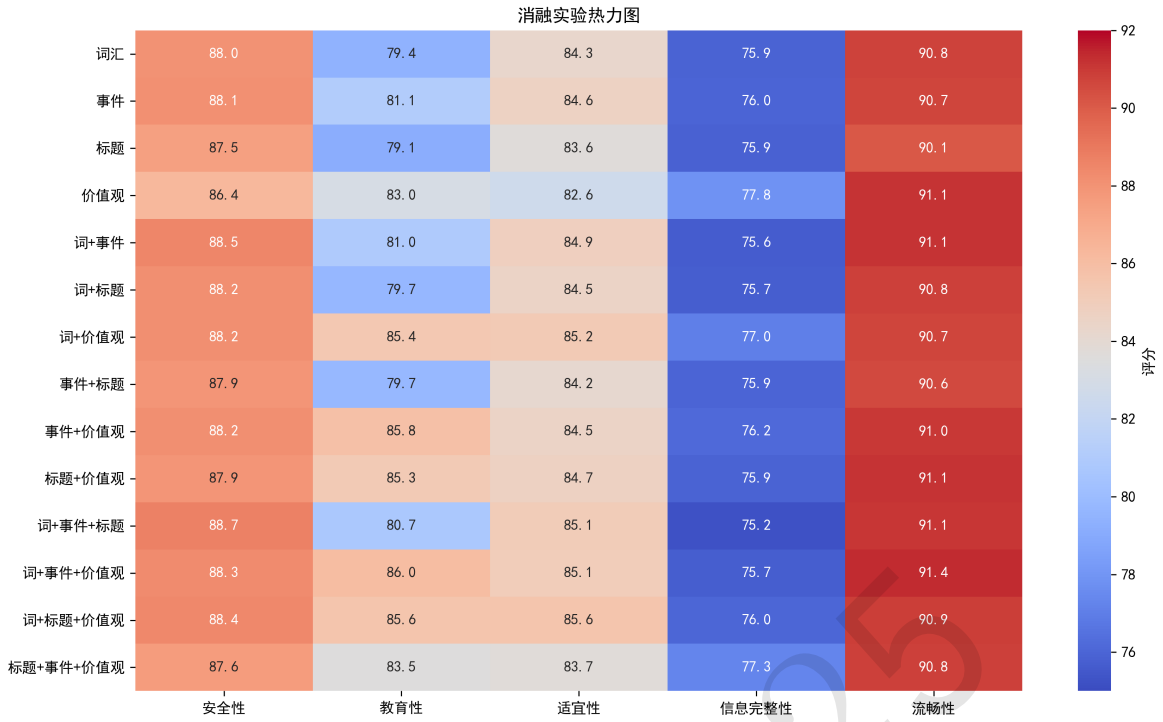



Figure 7: 各模块组合在多维度上的评估结果热力图

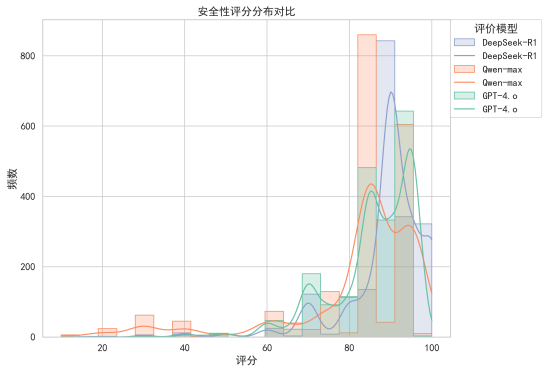
CCL 2025

第二十四届中国计算语言学大会论文集，第402页-第421页，济南，中国，2025年8月11日至14日。

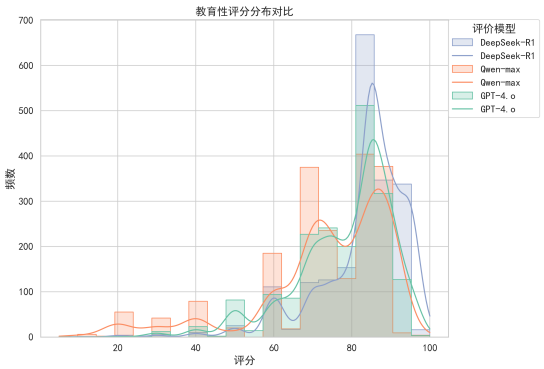
(c) 2025 中国中文信息学会计算语言学专业委员会

# E 一致性实验的评估结果

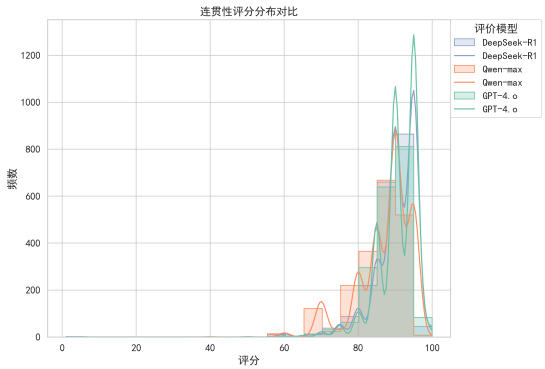
## E.1 各维度评分分布与一致性结果



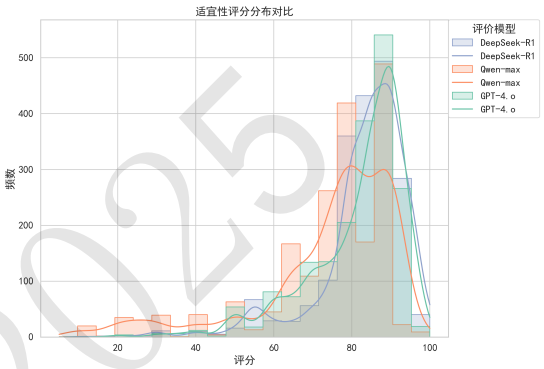

(a) 安全性



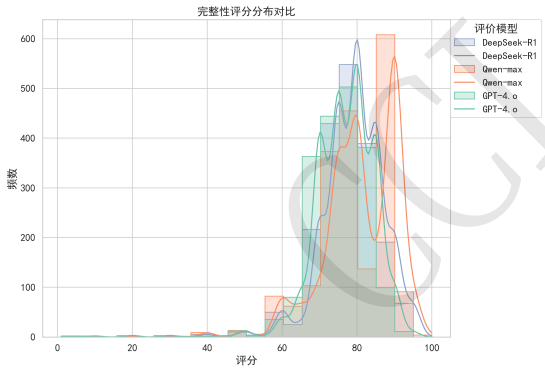

(b) 教育性



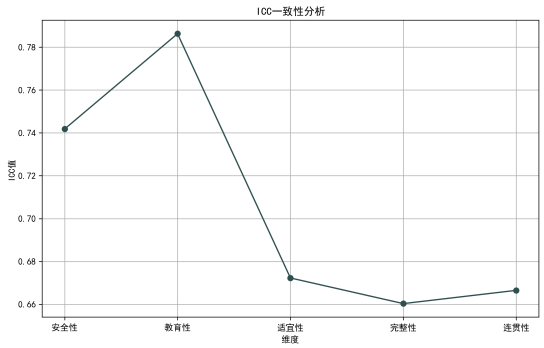

(c) 连贯性




(d) 适宜性




(e) 完整性




(f) 一致性

Figure 8: 各维度评分分布对比与ICC一致性

ACL 2025



(c) 2025 中国中文信息学会计算语言学专业委员会

### E.2 评价模型的相关性结果

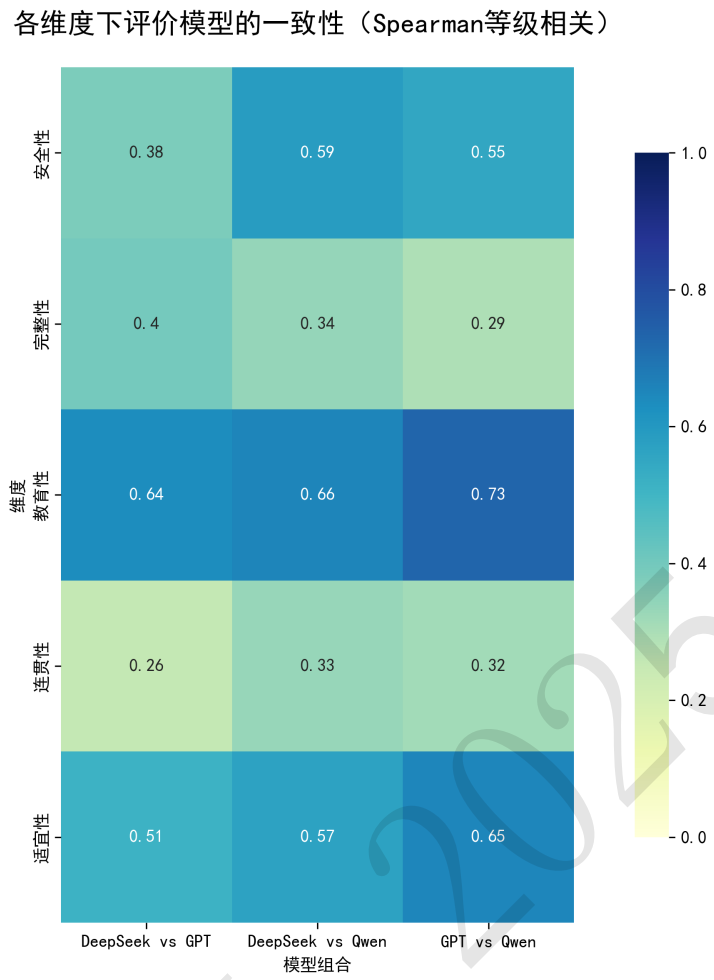



Figure 9: 三个评价模型在各维度上的相关性

## F 与人类评估的相关性补充实验

本文从每个领域随机选取30 篇共计90篇新闻，邀请两位语言学硕士在四个维度上进行风险点标注，每篇新闻的标注结果包括四个维度的所有风险点、风险原因及解决方案。使用CRV框架进行风险识别和意见生成后，计算不同维度间的风险识别结果与标注结果的ROUGE 和BLEU 指标。

| | ROUGE | | | | BLEU | | | |
|---|---|---|---|---|---|---|---|---|
| 模型名称 | 词汇 | 事件 | 标题 | 价值观 | 词汇 | 事件 | 标题 | 价值观 |
| CRV | **45.91** | **43.3867** | **60.686** | **68.8245** | **40.6486** | **44.9729** | **57.4998** | **56.4670** |
| ERNIE-3.5 | 27.1134 | 29.1043 | 40.6748 | 21.8103 | 9.6640 | 11.6528 | 14.4768 | 23.4050 |
| 讯飞星火 | 32.9053 | 33.8856 | 47.0113 | 27.2138 | 12.5410 | 14.1248 | 16.5117 | 26.4667 |

Table 1: 人类相关性对比

结果显示，在两个指标上，CRV 对比两个国内模型均高出较多，证明了CRV 在与人类的相关性上的显著优势。

第二十四届中国计算语言学大会论文集，第402页-第421页，济南，中国，2025年8月11日至14日。

(c) 2025 中国中文信息学会计算语言学专业委员会