多模态大语言模型长文本标注伦理审查申请书

申请日期： 2025 年 3 月 21 日

课题名称： Multi-Turn Interleaved Preference Alignment with Human Feedback	
单位： Institute for Artificial Intelligence, Peking University	
项目介绍	为了提升多模态大语言模型的通用性，我们使用多模态大语言模型（开源 Qwen2.5-VL 系列和闭源模型）构造了用于多轮多模态对齐的人类偏好数据集 InterMT Dataset，并开发了适用于人类价值对齐的 moderation。 To enhance the generalizability of multimodal large language models, we constructed the InterMT Dataset, a human preference dataset for multimodal alignment across multiple rounds, using both open-source (Qwen2.5-VL series) and closed-source multimodal large language models. Additionally, we developed a moderation approach for aligning with human values.
审查事项	1. **知情同意**：标注人员必须完全了解标注任务的性质、研究的目的、标注的数据将如何使用以及可能涉及的任何风险。 **Informed Consent:** Annotators must fully understand the nature of the annotation task, the purpose of the research, how the annotated data will be used, and any potential risks involved. 2. **隐私和保密**：标注过程可能涉及处理敏感或可识别个人身份的信息。确保参与者数据被匿名化并安全存储。 **Privacy and Confidentiality:** The annotation process may involve handling sensitive or personally identifiable information. Ensure participant data is anonymized and securely stored. 3. **补偿和剥削**：需要考虑人工标注者的补偿公平性，以避免剥削。 **Compensation and Exploitation:** Consider the fairness of compensation for human annotators to avoid exploitation. 4. **心理影响**：大语言模型安全标注任务可能涉及暴露于潜在有害或令人不安的内容，特别是涉及攻击性或暴力语言的任务。该项目需要评估并减轻参与者的任何心理风险。 **Psychological Impact:** The safe annotation tasks for large language models may involve exposure to potentially harmful or disturbing content, especially tasks involving offensive or violent language. The project needs to assess and mitigate any psychological risks to participants.

单位 盖章	申请课题组：杨耀东课题组； 申请缘由：投稿工作使用人类标注工参与模型偏好标注，用以支撑论文投稿至 NeurIPS 2025； 盖章： 申请时间：2025 年 5 月 21 日