方法报告

Track 1: 作文切题度自动评分

1 研究目标与方法概述

本研究旨在开发并评估一种自动化系统，用于对中小学生作文档的切题度进行多级别分类（涵盖“优秀”至“不合格”五个等级）。为实现此目标，我们构建了一种混合架构，该架构融合了基于大型语言模型（LLM）的多智能体协作系统与一个辅助性的 BERT（Bidirectional Encoder Representations from Transformers）分类模块。

2 LLM 多智能体分析框架

该框架的核心是一个通过精心设计的提示工程（Prompt Engineering）引导的、由多个功能专一的 LLM 智能体（基于 DeepSeek 模型）组成的级联处理流程。各智能体按序执行特定分析任务，为最终评分提供多维度信息：

1.题目要求解析智能体：负责精确解析作文题目的显性与隐性指令，建立规范化的切题度评估基准。

2.文章主旨提取智能体：负责从作文文本中提炼其核心论题或记叙主线，并评估主旨表达的明确性。

3.内容-任务符合度校验智能体：基于前两个智能体的输出，严谨校验作文内容与既定题目要求之间的符合度，识别关键的契合点与偏离点。

4.素材-主题契合度审核智能体：评估作文中所选用的素材（如事例、细节描写）支撑核心主旨及回应题目要求的有效性与相关性。

5.综合裁决智能体：整合上述所有分析报告，依据官方评分标准细则，生成初步的切题度分类建议。

6.分类结果提取智能体：从综合裁决报告中精确提取最终的分类标签。

3 BERT 辅助分类模块

为增强系统评估的鲁棒性，我们引入了一个基于 bert-base-chinese 预训练模型的增强型 BERT 架构（EnhancedBertEssayModel）。该模块以作文的题目要求、年级、标题及正文内容的拼接文本作为输入。通过无监督学习范式进行训练（具体采用了基于隐藏层状态表示的对比损失函数），旨在捕捉文本深层语义特征，并独立输出一个辅助性的切题度预测分类。训练数据源自测试集本身，以适应特定数据分布。

4 混合决策与冲突仲裁机制

系统的最终分类决策采用如下混合机制：

以 LLM 多智能体框架（特别是综合裁决智能体）输出的分类作为主要判定依据。

BERT 模块提供独立的辅助分类预测。

当 LLM 框架与 BERT 模块的评估结果出现不一致时，系统将激活一个专门设计的 LLM 冲突仲裁智能体 (Arbitration Agent)。该仲裁智能体接收作文原文、题目要求、LLM 与 BERT 的冲突分类结果以及 LLM 综合裁决报告的关键片段作为输入。其核心任务是依据官方评分标准，进行独立的最终裁决，以解决分类分歧。

若 LLM 与 BERT 评估一致，或 BERT 预测失败，则直接采纳 LLM 框架的裁决结果。在仲裁智能体调用失败的罕见情况下，亦保留 LLM 框架的原始裁决结果。

该混合方法通过结合 LLM 的深度语境理解、上下文分析能力与 BERT 的语义特征提取能力，并引入冲突仲裁层，旨在实现更准确、更可靠的作文切题度自动评估。

Track 2：相关性评论自动生成

1 研究目标与方法概述

本任务的核心目标是基于作文的切题度分析，自动生成一段仅聚焦于作文中心思想及其与题目要求相关性的评价性文本（评语）。为此，我们设计并实现了一个多阶段的 LLM 智能体流水线（Pipeline）。

2 系统输入与依赖

该系统的运行不仅依赖于输入的作文数据（题目要求、年级、标题、内容），亦关键性地利用了 Track 1 任务输出的最终切题度分类结果 (DUFL2025_track1.json) 作为指导信息。

3 LLM 评论生成流水线

该流水线包含以下按序执行的 LLM 智能体：

1.题目要求分析智能体：再次对题目要求进行分析，为后续评语生成提供精确的上下文参照。

2.作文主旨提取智能体：提取作文的核心观点或记叙主旨。

3.切题度评估智能体：基于前两步的分析结果，生成一段关于作文切题表现的描述性文本（此阶段非输出分类标签，而是文本化评估）。

4.初步评语草稿生成智能体：此智能体结合切题度描述性文本与 Track 1 的最终分类结果，生成初步评语。其提示被严格约束，确保：(a) 评语内容严格限定于切题性及中心思想表达；(b) 整体评价基调与 Track 1 分类结果保持一致；(c) 若切题度评估结果非最优（如非“优秀”），则提供仅针对提升切题性的具体改进建议。

5.合规性检查与精炼智能体：作为流水线的最后关键环节，此智能体对草稿评语执行多重验证与优化：

内容约束验证: 强制移除任何涉及语言风格、结构逻辑、素材细节、修辞手法、错别字、篇幅等非切题性/非中心思想相关的所有内容。

格式规范化: 清除所有 Markdown 标记、特殊符号、不必要的标点及格式，确保输出为单一、连续的纯文本段落。

长度约束执行: 严格确保最终评语的字符数精确落在 150 至 250 字符的目标区间内。

分类一致性复核: 再次确认评语的评价倾向与 Track 1 的输入分类标签完全吻合。

语言专业性润色: 在满足所有约束的前提下，提升语言表达的准确性、流畅度与专业性，使其符合中小学作文评语规范。

通过这一系列智能体的协同工作，特别是最终环节的严格过滤与精炼，本系统旨在生成高度聚焦于任务要求（切题度与中心思想）、满足所有格式与长度限制的高质量自动化评语。

Method Report

Track 1: Automated Essay Relevance Scoring

1 Objective and Methodological Overview

This study aims to develop and evaluate an automated system for the multi-level classification of essay relevance in primary and secondary school student writing, spanning five categories from "Excellent" to "Unqualified". To achieve this, we constructed a hybrid architecture integrating a multi-agent collaborative system based on Large Language Models (LLMs) with an auxiliary Bidirectional Encoder Representations from Transformers (BERT) classification module.

2 LLM Multi-Agent Analysis Framework

The core of this framework, guided by meticulous prompt engineering, comprises a cascaded processing pipeline of specialized LLM agents (based on the DeepSeek model). Each agent sequentially performs a specific analytical task, providing multi-dimensional information for the final score:

1.Prompt Requirements Parsing Agent: Precisely parses the explicit and implicit instructions of the essay prompt to establish standardized relevance assessment benchmarks.

2.Essay Theme Extraction Agent: Extracts the core thesis or narrative thread from the essay text and assesses the clarity of the main idea's expression.

3.Content-Task Alignment Verification Agent: Rigorously verifies the alignment between the essay content and the established prompt requirements (based on outputs from the preceding two agents), identifying key points of congruence and divergence.

4.Material-Theme Coherence Auditing Agent: Evaluates the effectiveness and relevance of supporting materials (e.g., examples, descriptive details) in bolstering the core theme and addressing prompt requirements.

5.Final Adjudication Agent: Integrates all preceding analysis reports and, based on official scoring rubric specifications, generates a preliminary relevance classification recommendation.

6.Classification Extraction Agent: Accurately extracts the final classification label from the Final Adjudication Agent's report.

3 BERT Auxiliary Classification Module

To enhance the robustness of the system's assessment, we incorporated an enhanced BERT architecture (EnhancedBertEssayModel) based on the bert-base-chinese pre-trained model. This module takes concatenated text—comprising the essay's prompt, grade level, title, and content—as input. It is trained using an unsupervised learning paradigm, specifically employing a contrastive loss function based on hidden state representations, aiming to capture deep semantic features of the text. It independently outputs an auxiliary relevance classification. Training data was sourced from the test set itself to adapt to the specific data distribution.

4 Hybrid Decision-Making and Conflict Arbitration Mechanism

The system's final classification decision employs the following hybrid mechanism:

The classification output by the LLM multi-agent framework (particularly the Final Adjudication Agent) serves as the primary determination.

The BERT module provides an independent, auxiliary classification prediction.

When a discrepancy arises between the assessments of the LLM framework and the BERT module, the system activates a specifically designed LLM Conflict Arbitration Agent. This arbitration agent receives the original essay, prompt requirements, the conflicting classification results from both LLM and BERT, and key excerpts from the LLM adjudication report as input. Its core task is to perform an independent final adjudication based on the official scoring rubrics to resolve the classification disagreement.

If the LLM and BERT assessments are consistent, or if the BERT prediction fails, the decision from the LLM framework is directly adopted. In the rare event of the arbitration agent failing, the original adjudication result from the LLM framework is retained.

This hybrid approach, combining the deep contextual understanding and analytical capabilities of LLMs with the semantic feature extraction proficiency of BERT, augmented by a conflict resolution layer, aims to achieve more accurate and reliable automated essay relevance scoring.

Track 2: Automated Relevance-Focused Comment Generation

1 Objective and Methodological Overview

The primary objective of this task is, based on the essay's relevance analysis, to automatically generate evaluative text (comments) strictly focused on the essay's central theme and its relevance to the prompt requirements. To this end, we designed and implemented a multi-stage LLM agent pipeline.

2 System Input and Dependencies

The operation of this system relies not only on the input essay data (prompt, grade, title, content) but also crucially utilizes the final relevance classification results from Track 1 (DUFL2025_track1.json) as guiding information.

3 LLM Comment Generation Pipeline

This pipeline involves the sequential execution of the following LLM agents:

1.Prompt Requirements Analysis Agent: Re-analyzes the prompt requirements to provide a precise contextual reference for subsequent comment generation.

2.Essay Theme Extraction Agent: Extracts the core viewpoint or narrative focus of the essay.

3.Relevance Assessment Agent: Generates descriptive text evaluating the essay's relevance

performance based on the analyses of the preceding two agents (Note: this stage outputs a textual assessment, not a classification label).

4.Draft Comment Generation Agent: This agent combines the descriptive relevance assessment text and the final classification result from Track 1 to generate an initial comment draft. Its prompts are strictly constrained to ensure: (a) comment content is rigorously limited to relevance and central theme expression; (b) the overall evaluative tone aligns with the Track 1 classification result; (c) if the relevance assessment indicates suboptimal performance (e.g., not "Excellent"), it provides specific improvement suggestions solely targeting relevance enhancement.

5.Compliance Checking and Refinement Agent: As the critical final stage, this agent performs multiple validation and optimization steps on the draft comment:

Content Constraint Validation: Mandates the removal of any content pertaining to language style, structural logic, material details, rhetorical devices, grammatical errors, length, or other aspects unrelated to relevance or the central theme.

Format Standardization: Eliminates all Markdown markup, special symbols, extraneous punctuation, and formatting, ensuring the output is a single, continuous plain text paragraph.

Length Constraint Enforcement: Strictly ensures the final comment's character count falls precisely within the 200 to 250 character target range.

Classification Consistency Review: Re-verifies that the comment's evaluative stance fully aligns with the input classification label from Track 1.

Professional Language Polishing: Enhances linguistic accuracy, fluency, and professionalism, adhering to the norms of secondary school essay feedback, while satisfying all preceding constraints.

Through the collaborative work of this agent pipeline, particularly the stringent filtering and refinement applied in the final stage, the system is designed to generate high-quality, automated comments that are highly focused on the specified task requirements (relevance and central theme) and meet all formatting and length specifications.