基于 BERT 和 GenAI 的研究生学位论文质量监测系统

A postgraduate degree thesis quality monitoring system based on BERT and GenAI

兰丽娜[1*]，郭子豪[1]，刘冬[2]，徐天岭[2]，赵佳瑶[1]
[1] 北京邮电大学人文学院
[2] 北京邮电大学研究生院
[*] lanlina@bupt.edu.cn

【摘要】 针对传统研究生学位论文质量监测依赖人工评阅、问题挖掘不充分的痛点，构建基于 BERT 模型和生成式人工智能（GenAI）的论文质量监测系统。首先，提出基于 BERT 模型的学位论文评阅典型问题挖掘方法，借助其深层语义理解能力，精准识别论文在选题价值、逻辑架构、学术规范等维度的共性与个性化问题；其次，依托 GenAI 技术自动生成多维度质量分析报告，包含论文质量量化评估结果，针对性输出面向学生的论文修改优化建议和面向导师的指导重点改进方案。实验结果表明，该系统可显著提升论文问题挖掘的准确性与效率，为研究生培养质量提升提供智能化支撑。

【关键词】 BERT 模型；生成式人工智能；研究生学位论文；质量监测系统；问题挖掘

Abstract: *Aiming at the pain points of traditional postgraduate thesis quality monitoring, such as over-reliance on manual analysis, low efficiency, and inadequate excavation of existing problems, this study constructs a postgraduate thesis quality monitoring system based on the BERT model and Generative Artificial Intelligence (GenAI). First, a method for mining typical problems in thesis review based on the BERT model is proposed, which takes advantage of its powerful deep semantic understanding capability to accurately identify both common and individual defects of theses in dimensions including research topic value, logical structure, and academic norms. Second, relying on GenAI technology, a multi-dimensional quality analysis report is automatically generated, which not only presents the quantitative evaluation results of thesis quality, but also provides targeted optimization suggestions for students and improvement plans for supervisors' guidance priorities. Experimental results show that the proposed system can significantly improve the accuracy and efficiency of thesis problem mining, providing intelligent support for the improvement of postgraduate training quality.*

Keywords: BERT Model, Generative AI, Postgraduate Thesis, Quality Monitoring System, Problem Mining

1. 引言

研究生学位论文是衡量研究生培养质量与学术能力的核心载体。当前研究生招生规模扩大，传统以人工为主的论文评阅模式存在效率低、典型问题挖掘不深入、缺乏针对性改进指导方案等突出痛点，难以适配规模化、精准化的质量管控需求。随着自然语言处理技术和人工智能技术的快速发展，BERT 模型具有卓越的深层语义理解能力，生成式人工智能（GenAI）则为自动化报告生成提供了可能。本研究结合两者优势，构建基于 BERT 模型和 GenAI 的研究生学位论文质量监测系统，旨在突破人工评阅局限，实现论文核心缺陷问题的精准识别并生成量化评估与双向改进建议，提升论文质量监测的智能化水平。

2. 研究综述

当前研究生学位论文质量监测的人工评审机制面临标准模糊和主观偏差等问题（董云川，2025），而现有辅助技术多局限于查重和浅层文本分类，难以精准挖掘逻辑严谨性等隐性缺陷（王益彬，2025）。在文本分析领域，BERT 及其衍生模型凭借双向注意力机制在深层语义理解上表现优异，垂直微调的 BERT 模型在特定诊断任务中比通用大语言模型更具稳定性和可解释性（Elmassry A M，2025）。另外，GenAI 能有效细化指导并提升论文修改效率，但直接应用于学术评价时易产生“幻觉”，其推理的准确率瓶颈亟需引入前置诊断机制来克服（Liang，2025）。现有研究尚未构建“问题挖掘-质量评估-改进指导”一体化的智能监测系统。因此，本文提出构建基于 BERT 模型与 GenAI 的研究生学位论文质量监测系统，实现论文典型问题的精准挖掘和针对性改进建议的生成。

3. 监测系统设计

3.1. 系统结构

基于 BERT 和 GenAI 的研究生学位论文质量监测系统结构如图 1 所示。

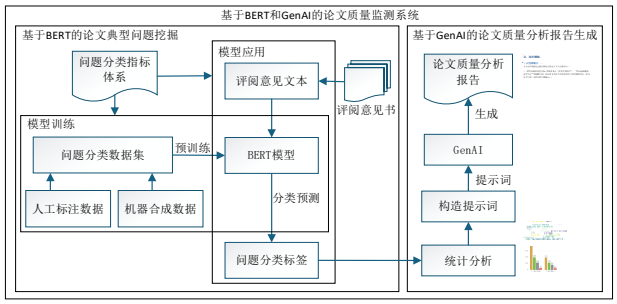

图 1 *基于 BERT 和 GenAI 的论文质量监测系统*

系统主要包括 2 大功能模块。①基于 BERT 的典型问题挖掘模块：构建问题分类指标体系与高质量数据集，通过预训练 BERT 模型实现对大规模评阅意见文本的精准分类。②基于 GenAI 的质量分析报告生成模块：对问题分类结果进行多角度统计，构造提示词并调用 GenAI 接口，自动生成包括对师生双方建议的完整质量分析报告。

3.2. 基于 BERT 模型的学位论文典型问题挖掘方法

3.2.1. 基于 BERT 模型的论文问题挖掘方法流程

基于 BERT 模型的论文问题挖掘方法流程如图 2 所示。分为三个阶段。①数据增强与构建阶段，主要构建高质量监督训练数据集；②模型微调阶段，主要进行预训练，实现参数优化；③结构化输出阶段，对非结构化的文本进行分类预测得到结构化的问题分类标签数据，用于质量分析和诊断。

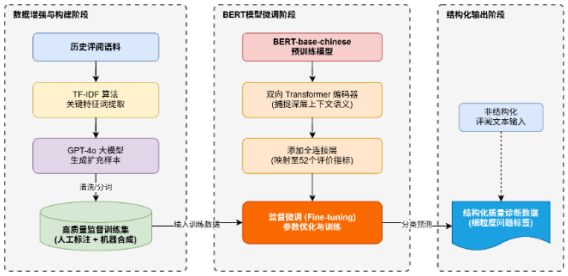

图 2 *基于 BERT 模型的论文典型问题挖掘方法流程*

3.2.2. 细粒度问题分类指标集构建

基于理工科学位论文的研究范式与评阅重点，结合已有评价标准，通过文献梳理、专家访谈与评阅意见扎根分析，构建了一个层次分明、覆盖全面的问题分类指标体系。该体系包含 6 个一级指标和 55 个二级指标，如表 1 所示。指标体系旨在从不同维度细化论文意见的问题描述，为后续的量化分析提供标准化的标签基础。

表 1 *问题分类指标体系（部分）*

一级分类指标	二级分类指标
1 论文选题价值	1-1 研究意义不明确 1-3 研究内容陈旧 1-5 选题前沿性不足 1-6 其他选题问题
2 文献综述与研究现状	2-1 综述不全面 2-3 相关性不强 2-4 问题导出描述弱
3 研究内容	3-2 系统性/逻辑性不足 3-5 分析深度不足 3-6 理论依据不足
4 研究方法和实验	4-1 实验不充分 4-3 实验描述不完整 4-4 结果分析不充分
5 成果创新性和应用价值	5-1 成果创新性不足 5-5 成果应用价值不足
6 论文结构与写作规范性	6-1 结构不清晰 6-2 章节标题不恰当 6-7 图表不规范 6-11 参考文献标注问题

3.2.3. 问题分类数据集构建

基于某高校近 5 年博士论文评阅数据，采用专家人工标注（1096 条）与机器合成（2797 条）结合的方式，构建了 3893 条问题分类数据集，示例如表 2，分布见图 3。机器合成阶段利用 TF-IDF 提取特征词并结合 GPT-4o 进行数据增强，有效提升了模型泛化能力。

表 2 *问题分类数据集示例*

预训练数据	评阅意见文本 text	标签 label
人工标注数据（1096 条）	研究问题不明确，没能很好地体现问题的重要性	1-2 研究问题不明确
	在第二章现有工作分析中，可适当补充近几年参考文献	2-1 综述不全面
机器合成数据（2797 条）	综述部分对研究现状的分析不够全面，缺乏对研究不足的系统总结	2-1 综述不全面

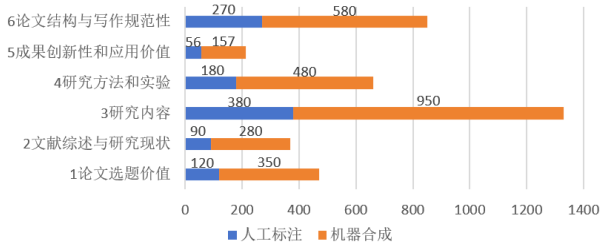

图 3 *问题分类数据集分布*

3.2.4. BERT 模型预训练与应用

加载 bert-base-chinese 模型并以 9:1 划分训练与测试集进行预训练，代码如图 4。

```python
from transformers import BertTokenizer, BertForSequenceClassification, Trainer, TrainingArguments
from datasets import load_dataset
# 1. 加载预训练模型和分词器
model_name = "bert-base-chinese"
tokenizer = BertTokenizer.from_pretrained(model_name)
model = BertForSequenceClassification.from_pretrained(model_name, num_labels=52)
# 2. 数据预处理
def tokenize_function(examples):
    return tokenizer(examples["text"], padding="max_length", truncation=True)
# 3. 设置训练参数与初始化
training_args = TrainingArguments(
    output_dir="./results",
    evaluation_strategy="epoch",
    learning_rate=2e-5,
    num_train_epochs=5,
)
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_datasets["train"],
    eval_dataset=tokenized_datasets["test"],
)
trainer.train()
```

图 4 *BERT 模型预训练代码*

```python
def generate_report(problem_stats):
    prompt = f"""
    基于BERT分析，主要共性问题如下：
    1. {problem_stats['top1']}（频率：{problem_stats['freq1']}）

    请生成质量分析报告：
    (a) 共性薄弱点深度剖析；
    (b) 给研究生的具体操作建议；
    (c) 给导师的管理与把关建议。
    """
    # 调用 OpenAI API（伪代码）
    return openai.ChatCompletion.create(model="gpt-4", messages=[...])
```

图 5 *GenAI 生成论文改进建议的接口代码*

表 3 *不同分类模型的特点与性能对比*

模型	特征提取方式	上下文理解	优点	不足	准确率

朴素贝叶斯	TF-IDF (词频统计)	否	计算速度快	无法捕捉语义关联，对否定句和长难句识别较差	60.2%
BERT 模型	Transformer(自注意力机制)	是	精准捕捉深层语义，鲁棒性强	训练算力要求较高	87.1%

在同一数据集上进行了基准对比实验（表 3）。实验结果表明，经过微调的 BERT 模型分类准确率达 87.1%，比朴素贝叶斯算法高约 17%。该模型已应用于 4002 份博士学位论文评阅书，切分为约 2 万条短句，实现了问题分类的自动化预测与画像分析。

3.3. 基于 GenAI 的学位论文质量分析报告生成机制

利用大语言模型自动生成改进建议，构造提示词及生成论文改进建议的代码如图 5 所示。AI 生成的质量分析报告部分截图如图 6 所示：

(a)共性薄弱点　(b)对学生的改进建议　(c)对导师的改进建议

图 6 *质量分析报告中 AI 生成的部分*

3.4. 可视化分析

系统可视化分析界面如图 7 所示。展示了问题分布，包括高频问题排位、总体分布及问题词云图。分析可见，目前最突出的三大缺陷集中于“分析深度不足”、“图表不规范”及“研究内容系统性不足”，为提升论文严谨性与规范性提供了明确的专项优化方向。

(a) 问题前 10　(b) 问题总体分布　(c) 问题词云图

图 7 *可视化分析界面*

4. 小结

本文构建了基于 BERT 与 GenAI 的学位论文质量监测系统，实现了精准问题挖掘与针对性建议生成的深度集成。该系统有效突破了传统监测工具识别表层、反馈形式化和针对性不足问题，推动质量监测从“人工主导”向“智能赋能”的转型。未来将聚焦多学科样本扩充与多模型融合，提升系统的普适性与智能化水平，为研究生培养质量提升提供更全面支撑。

致谢：本文受北京邮电大学研究生教育教学改革研究重点项目“面向高质量发展的研究生学位论文质量监测与评价体系研究”（2024Y027）资助。

参考文献

董云川,刘静.(2025).“盲审”还需“明辨”补：博士学位论文评审机制反思.学位与研究生教育,04,78-86.

王益彬,石艳.(2025).学位论文 AIGC 检测的现实逻辑、教育效应与技术想象重构.中国高教研究,12,25-32.

Elmassry A M, Zaki N, Alsheikh N,et al. (2025). A Systematic Review of Pretrained Models in Automated Essay Scoring, *IEEE Access, 13*, 121902-121917.

Liang, Z., et al. (2025). Assessing Large Language Models for Automated Feedback Generation in Learning Problem Solving. arXiv preprint arXiv:2503.14630.