



毕业论文（设计）开题报告

题 目： 基于大语言模型的项目申报书形式审查系统

姓 名： 吕付昊 学号： 19221128

学 院： 人工智能学院

专 业： 计算机科学与技术

指导教师： 丁钰 职称： 讲师

论文地点： 南京农业大学滨江校区

2024 年 11 月 22 日

一、结合毕业论文（设计）课题任务情况，根据所查阅的文献资料，撰写选题背景和意义及国内外文献综述：

大型语言模型（LLMs）是人工智能领域的一项重要技术突破，专门设计用于理解和生成人类语言[1]。作为参数规模庞大的深度学习模型，LLMs通过在海量文本数据上进行无监督或自监督预训练，能够捕捉并模拟语言的复杂性和多样性。它们通常基于Transformer架构，这种架构依赖于自注意力机制[2]，允许模型在处理一个单词时考虑到整个文本序列中的其他单词，从而更好地理解上下文和语义信息。

这些模型不仅具备广泛的语言知识，能够在不同领域间有效迁移，还能在少样本或零样本的情况下快速适应新任务[3]，展现出强大的上下文学习和推理能力。大语言模型的生成能力尤为突出，可以生成流畅、自然且多样化的文本，满足个性化需求，并将多种自然语言处理任务统一为生成式框架，减少了级联损失。此外，LLMs能够通过指令遵循与用户交互[4]，理解复杂意图，并通过外部工具集成增强功能，如调用搜索引擎或计算器来获取实时信息。这种多模态融合和动态知识更新的能力，使得大语言模型不仅能处理纯文本任务，还能应对图文结合的内容[5]，进一步扩展了其应用范围。

尽管存在鲁棒性和伦理安全性的挑战，但通过技术如知识蒸馏[6]、模型压缩[7]以及基于人类反馈的强化学习，大语言模型正在不断提升性能并优化输出质量，推动自然语言处理领域的持续创新和发展。相对于传统的机器学习方法，LLMs的主要优势在于它们能够处理更复杂的语言结构和语义，提供更自然、更准确的语言理解和生成能力，这使得它们在机器翻译、文本摘要、问答系统等多个领域中发挥着越来越关键的作用。同时，LLMs的灵活性和泛化能力也使得它们能够适应不断变化的语言使用模式和新兴的应用场景。大模型在自然语言处理领域展现出了前所未有的理解与生成能力。这些大模型，如GPT系列、BERT等[8]，通过海量的数据训练与复杂的网络结构，已经能够非常出色地理解输入的语句。它们不仅能够捕捉到语句的字面意义，还能理解上下文语境、隐含情感以及言外之意，从而在对话系统、问答系统、文本生成等多个场景中表现出色[9]。这种理解能力使得大模型能够更好地服务于用户，提供更精准、更人性化的交互体验。

项目申报书作为科研或政府资助申请的核心文件，其复杂性和繁琐程度往往令人望而生畏。它不仅需要详尽阐述项目的背景、目标、方法、预期成果等关键信息，还必须严格遵守特定的格式和排版要求，以确保评审人员能够清晰、快速地获取所需内容。这意味着，在撰写过程中，申报者不仅要投入大量精力进行内容的构思和撰写，还需时刻留意字体、字号、行距、页边距、标题层级等细节，以避免因格式问题而影响评审结果。格式检查作为这一过程中的重要环节，不仅耗时费力，而且极易出错，稍有疏忽便可能导致整个申报书被退回重改，从而延误宝贵的申请时机。近几年来,随着国家对基础科研经费投入的加大,国家自然科学基金的项目申请数每年都在快速增长。从2008-2011年,国家自然科学基金委员会(以下简称“自然科学基金委”)接收的项目申请数量分别为79 860、97 755、115 179和147 703项[10]。面对项目申请量的迅速增长,实现格式内容自动化检查可以提高申请申报质量、减少科研压力。

因此，高效、准确地完成格式检查，成为了可以减轻众多申报者的格式压力的有效办法。结合大语言模型在自然语言处理中的优势，可以推断出其在自动化检查项目申报书内容方面具有巨大潜力。大语言模型能够快速识别并纠正语法错误、逻辑混乱等问题，提高申报书的质量和准确性。同时，它还能根据特定的格式和要求，对申报书格式审查进行低代码量个性化定制，确保符合相关标准和规范。因此，确保申报书的准确性、规范性和逻辑性有重要意义。

**参考文献：**

[1] Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.

[[2]VASWANI A,SHAZEER N,PARMAR N,et al.Attention is All You Need[EB/OL](2017-06-12)[2024-05-10].https://arxiv.org/abs/1706.03762.](javascript:void(0))

[3]丁培耕,刘智威,王金桥.多模态人工智能大模型综述及展望[J/OL].无线电工程,1-20[2024-12-01].http://kns.cnki.net/kcms/detail/13.1097.tn.20241121.1402.002.html.

[[4]YE J J,CHEN X T,XU N,et al. A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models[EB/OL].(2023-03-18)[2024-07-11].https://arxiv.org/abs/2303.10420.](javascript:void(0))

[5]车万翔,窦志成,冯岩松,等.大模型时代的自然语言处理:挑战、机遇与发展[J].中国科学:信息科学,2023,53(09):1645-1687.

[6]邵仁荣,刘宇昂,张伟,等.深度学习中知识蒸馏研究综述[J].计算机学报,2022,45(08):1638-1673.

[7]高晗,田育龙,许封元,等.深度学习模型压缩与加速综述[J].软件学报,2021,32(01):68-92.DOI:10.13328/j.cnki.jos.006096.

[8]岳增营,叶霞,刘睿珩.基于语言模型的预训练技术研究综述[J].中文信息学报,2021,35(09):15-29.

[9]程齐凯,李信,陆伟.领域无关学术文献词汇功能标准化数据集构建及分析[J].情报科学, 2019, 37(7):41-47.

[10]王文棋,张建波,李国栋,等.国家自然科学基金申请项目人员查重方法应用和比较[J].中国科学基金,2012,26(05):288-290.DOI:10.16262/j.cnki.1000-8217.2012.05.010.

二、毕业论文（设计）任务要研究或解决的问题和拟采用的研究方法、实验方案及可行性分析：

（1）要研究或解决的问题

当前项目申报书的撰写与审查过程中，格式检查与内容审核极为复杂且耗时。传统的人工审查方式不仅效率低下，而且容易出错，难以满足日益增长的申报需求。本研究旨在开发一个智能化的项目申报书审查系统，通过解析Docx文档内容与格式信息，利用规则引擎与自然语言理解技术，实现申报书的自动格式审查与内容审核，从而提高审查效率与准确性。

（2）拟采用的研究方法

1. Docx文档解析：利用Python-docx库读取Docx文档的内容与格式信息，对文档进行分段处理，提取每段的文本内容、字体、字号、对齐方式等格式信息。这些信息将作为输入提交给大模型进行后续处理。
2. 自然语言理解模块：经过大模型检测输入的格式要求DOCX文档，输出应该处理的格式要求。然后作为输入送给规则模块，使其可以按照文档要求的格式进行查验。
3. 规则引擎模块：通过Word文档读取项目申报书形式审查规则库，涵盖各类申报指南、模板要求等。利用自然语言理解模块提取的信息与规则库进行匹配，实现规则校验。
4. 多Agent处理模块：设计格式审查Agent2和要求理解Agent1，分别负责文档的格式审查和格式要求理解。通过Agent1根据格式内容文档的要求，设计格式审查Agent2和建议修改Agent3的prompt，实现智能审查功能。
5. 反馈生成模块：根据规则校验结果，生成针对性的反馈意见DOCX，包括错误提示、建议修改等。支持多种反馈形式，如文本、表格、高亮显示等，方便用户查看和修改。
6. 人机交互模块：研究并设计友好的用户界面，支持用户对审查结果进行查看和对话，提高系统的易用性和用户体验。

（2）实验方案

数据准备：收集大量项目申报书样本及对应的格式要求、内容要求等规则库。

模型训练：利用收集的数据训练自然语言理解模型，提高其对申报书内容的深度语义理解能力。

系统集成：将Docx文档解析、规则引擎、自然语言理解、多Agent处理、反馈生成及人机交互等模块集成在一起，形成完整的智能化项目申报书审查系统。

测试与优化：对系统进行测试，评估其审查效率与准确性。根据测试结果进行优化对应的模块和内容，提高系统的性能和稳定性。

（3）可行性分析

技术可行性：Python-docx库、自然语言理解技术、规则引擎及多Agent系统等技术已相对成熟，且已有类似应用案例，因此技术上是可行的。

数据可行性：项目申报书样本及规则库可通过多种途径获取，如公开数据库、合作单位等，因此数据上是可行的。

三、研究计划及预期成果

|  |  |
| --- | --- |
| 时间段 | 主要工作内容 |
| 阶段一：需求分析与系统设计（2024.11.10-2025.01.20） | 详细调研项目申报书形式审查需求，设计系统架构，确定技术路线。 |
| 阶段二：数据准备与模型训练挑选（2025.01.20-2025.02.01） | 收集大量项目申报书数据，构建训练数据集，训练大语言模型。 |
| 阶段三：建立合适的规则库（2025.02.01-2025.02.05） | 建立完善的项目申报书形式审查规则库。 |
| 阶段四：系统开发与集成（2025.02.05-2025.03.07） | 开发系统前后端模块，进行系统集成测试。 |
| 阶段五：系统修改与验收（2025.03.07-2025.04.07） | 完成系统集成测试。部署系统，进行测试，完善系统。 |
| 阶段六：论文写作和修改（2025.04.07-2025.05.01） | 论文写作和修改 |

四、特色或创新之处

1. 智能化审查流程：项目通过集成Docx文档解析、规则引擎、自然语言理解等多个模块，实现了申报书的自动格式审查与内容审核。这一智能化审查流程极大地提高了审查效率与准确性，减少了人工审查的繁琐与错误。
2. 多Agent协同处理：项目设计了格式审查Agent，理解Agent以及建议Agent分别负责文档的格式检查等这种多Agent协同处理的方式，使得系统能够更加精细地处理申报书的不同方面，提高了审查的全面性和准确性。
3. 灵活的反馈机制：项目根据规则校验结果，生成针对性的反馈意见，并支持多种反馈形式，如文本、表格、高亮显示等。这种灵活的反馈机制，使得用户能够更加方便地查看和修改申报书，提高了系统的易用性和用户体验。
4. 高度可扩展性：项目采用模块化设计，各个模块之间相对独立，易于扩展和维护。这意味着随着申报书审查规则的更新或新需求的出现，系统可以方便地添加或修改相应的模块，以适应新的审查要求。

**指导教师意见（对课题的深度、广度及工作量的意见和对毕业论文（设计）结果的预测）：**

指导教师签字：

年  月  日