赛题 04

数据标注平台

目录

1. 项目目标

1.1.1.1.项目背景

1.1.1. 公司背景

恒生电子是一家金融软件和网络服务供应商，也是一家以“让金融变简单”为使命的金融科技公司，1995 年成立于杭州， 2003 年在上海证券交易所主板上市（代码 600570.SH）。恒生电子以技术为核心竞争力，聚焦金融行业，致力于为证券、期货、基金、信托、保险、银行、交易所、私募等机构提供整体解决方案和服务。恒生已连续 15 年入选 FinTech100 全球金融科技百强榜单，2022 年排名第 24 位。目前拥有超过 13300 名员工，其中产品技术人员占比约 65%。多年来，恒生以技术服务为核心，凭借多年金融 IT 建设经验，以及对金融业务的深刻洞察和理解，用优质的产品与服务，持续赋能金融机构创新发展。

同时，恒生电子还积极履行企业社会责任，在投资者教育、扶贫济困、关爱自闭症儿童等领域持续贡献力量，实现企业与社会共同可持续发展。

1.1.2. 行业背景

随着人工智能技术的迅速发展，越来越多的领域需要使用文本和图片数据进行训练和优化，例如自然语言处理、计算机视觉等领域。例如，在自然语言处理领域，机器翻译、情感分析、语音识别等任务都需要使用大量的自然语言数据进行训练；在计算机视觉领域，图像分类、目标检测、人脸识别等任务需要使用大量的图像和视频数据进行训练。而这些数据的标注工作是非常繁琐、耗时的，需要大量的人力和时间投入。

为了解决这个问题，数据标注平台应运而生。这些平台可以为研究者和企业提供一个方便、快捷、高效的标注工具，使得标注过程更加高效、准确和可靠。同时，平台也可以提供多种标注方式和标注规范，以确保标注数据的质量和一致性。此外，平台还可以提供数据管理、质量控制、安全保障等方面的支持，以确保标注过程的稳定性和可靠性。

1.1.3. 业务背景

恒生电子从 2018 年开始涉足人工智能相关领域，目前公司内已有自然语言处理、图像处理、知识图谱等相关人工智能产品。在这个过程中，积累了大量的金融词库、语料及图片标注等数据

，目前这些数据分别由不同的系统管理。随着数据量和业务规模的不断的增长，原有的数据标注方式和管理模式已现瓶颈，我们需要搭建统一的标注平台，制定标准的数据标注流程和规范，对标注数据进行统一的管控，并提供版本控制、用户隔离等机制，以提升相关研究人员的工作效率，并保障数据资产安全。

1.2. 项目目标

1.2.1. 项目说明

【问题说明】

针对目前数据标注存在的问题，企业需要搭建数据标注平台。构建一套企业的数据接入、数据标注、标注审核、数据发布的标准流程。对于标注数据平台需要提供标注数据版本管理机制，以达到数据与下游模型版本匹配的效果。同时，对于流程中不同的角色，平台需要提供差异化的功能。此外，平台还需要对标注数据提供数据隔离机制，不同角色、不同级别的人所能看到数据不同，以保障企业数据资产的安全。

【用户期望】

1. **数据标注**

- 信息抽取标注

信息抽取标注是指对文本中的实体、关系、事件等信息进行标注的过程。信息抽取标注是自然语言处理中的重要任务之一，其目的是从文本中提取出有用的信息，为后续的信息处理和分析提供基础数据。信息抽取标注的流程包括语料导入、抽取信息定义、预标注、人工标注等步骤。

语料导入：业务人员将需要标注的文档导入到平台中，文档格式为：pdf 或者 word，去页眉、页脚、批注，并解析其文字内容；

抽取信息定义：以文档原格式展示语料，业务人员可根据实际业务对样例文档做抽取信息定义，支持实体、长文本、关联关系、枚举分类等标注任务。业务人员可通过在语料上作划选和批注完成抽取信息定义；

预标注：根据业务人员定义的抽取信息，提供自动标注功能，对语料进行初步的标注，以便后续的标注工作更加高效和准确；

人工标注：提供标注界面，标注人员可看到业务人员定义的抽取信息以及待标注语料，并可通过划选和批注的形式完成语料标注。

- 文本分类标注

文本分类标注是指对文本进行分类的过程，其目的是将文本划分为不同的类别，为后续的信息处理和分析提供基础数据。文本分类标注的流程包括语料导入、标签定义、预标注、人工标注等步骤。

语料导入：业务人员将需要标注的文本数据导入到平台中，文本格式为：短文本（字数少于100）；

标签定义：业务人员可根据实际业务对业务标签做定义和描述；

预标注：根据业务人员定义，提供自动标注功能，对语料进行初步的标注，以便后续的标注工作更加高效和准确；

人工标注：提供标注界面，标注人员可以看到待标注的语料，并可通过标签选择、证据划选等形式完成语料标注。

- 图像文本标注

图像文本标注是指框选出图像中出现的文本，其目的是框选出各种场景图片中出现的文本，标注结果会作为文本检测模型的训练数据来优化模型效果。

语料导入：业务人员将需要标注的图片数据导入到平台中，图片格式为：jepg、jpg、bmp、png、pdf 等；

预标注：根据业务人员定义，提供自动标注功能，对图片进行初步的标注，以便后续的标注工作更加高效和准确；

人工标注：对当前图片中的文本行进行手动绘制标记框框选单行文本，使用四点标注模式，依次点击 4 个点来确定一个标记框，需要支持倾斜文本行标记。标注的结果信息需要包含图片名称、图片上的所有标记框及标记框的 4 个点的坐标。

- 图片分类标注

图片分类是指对图片进行分类和打标签的过程，其目的是将图片按照一定的类别进行归类，为后续的信息处理和分析提供基础数据。在此处图片分类标注的流程包括确定标注标签、对图片进行标注等步骤。在进行图片分类标注时，需要根据具体的应用场景和需求，确定标注类别和标注规范。

语料导入：业务人员将需要标注的图片数据导入到平台中，图片格式为：jepg、jpg、bmp、png、pdf 等；

标签定义：标签将用于标注人员对图片的分类操作，以便后续的标注工作更加高效和准确。标签名称应该具有明确的含义和界限，避免出现歧义和重复；

预标注：根据业务人员定义，提供自动标注功能，对图片进行初步的标注，以便后续的标注工作更加高效和准确；

人工标注：标注人员可对图片进行标注，并能根据标签组，对图片进行分类和打标签。

2. 数据审核

标注审核是指对标注结果进行审核和修正的过程，其目的是确保标注结果的准确性和一致性。标注审核的流程包括对标注结果进行初步审核、对标注错误进行修正、对标注结果进行二次审核、对标注结果进行统计和分析等步骤。在标注审核过程中，需要严格按照标注规范进行操作，对标注错误进行及时修正，并对标注结果进行统计和分析，以便优化标注质量和效率。

3. 数据发布

对于已通过审核的标注数据集，可以进行发布。数据集多次发布，数据集每次发布都会生成一个唯一的发布版本号，发布时为数据集生成当前的数据快照，并进行存储，发布的数据快照不可进行修改，具备权限的用户可以导出已发布的数据集。

4. 项目管理

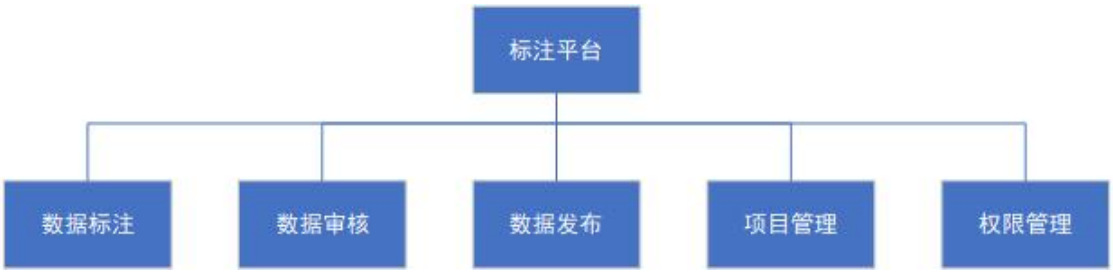
项目是数据标注的基础管理维度，所有的标注数据都归属于一个项目，不同项目之间数据相互隔离。用户只有拥有项目的权限，才可以对项目下属的数据进行标注、审核以及发布。

5. 权限管理

要求系统能够提供角色、用户、权限管理相关功能，由管理人员进行相关的权限分配。用户可以属于多个项目，当用户属于项目时，才可以看到该项目，以及该项目的数据。权限管理需要能对系统的菜单、按钮进行权限控制。

3、任务要求

1. 系统总体结构



1.2.2. 项目要求

- 开发语言：Java/js等
- 系统采用流行的B/S体系结构
- 应用服务器：Tomat / Jetty

- 数据库服务器：mysql

语料文件需要采用分布式文件系统存储，需要支持T级别数据量的语料数据管理。

2. 服务模型

2.1. 可行性服务模型

可行性分析是在启动项目或实施新计划之前进行的一项关键活动。它旨在评估项目的可行性，确定项目是否值得投资时间、资源和资金。可行性分析是一个系统性的过程，它有助于组织和决策者更好地理解项目的环境、挑战和机会。这种分析是项目管理过程中的关键一步，有助于确保项目在开始之前就有最大的成功机会。分析结果可行与否，直接关系到后期工作的实施决策。该部分中，项目组对于各个不同的方面分别进行了技术可行性，使用可行性，市场可行性的分析。相应的可行性分析模型图如图所示：

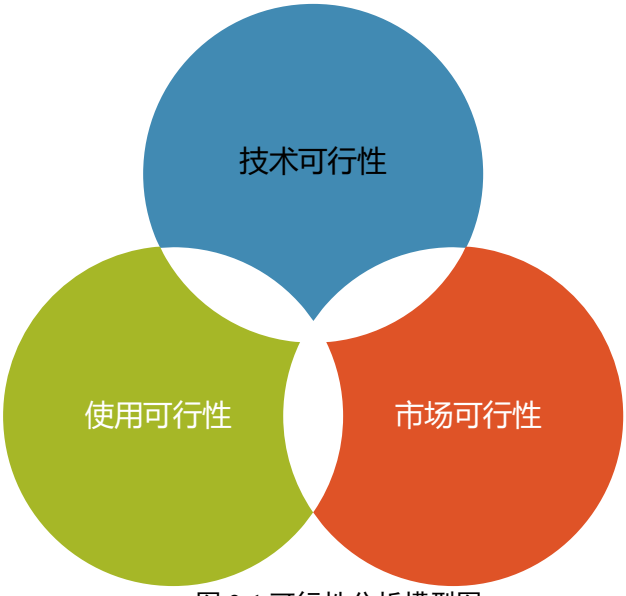


图 2-1 可行性分析模型图

2.1.1. 技术可行性

- 系统采用流行的B/S体系结构：

B/S (Browser/Server)体系结构在系统设计中具有多个优势。首先，它实现了跨平台的应用，用户只需通过浏览器访问，无需安装特定的客户端软件，提高了系统的灵活性和可访问性。其次，B/S 架构在服务器端进行数据处理和逻辑运算，降低了客户端的负担，减轻了终端设备的性能要求，有利于提高系统的整体性能。此外，B/S 模式便于系统的维护和升级，只需更新服务器端的代码，而无需每个客户端都进行更新。总体而言，B/S 体系结构简化了系统的部署和管理，提高了系统的易用性和可维护性，适用于广泛的应用场景。

- 核心算法准确度高

本系统经准确性验证表明，实际处理准确度较高，可以精确地把需要标注的文本或图片信息进行语料导入、抽取信息定义、预标注、标签定义、人工标注等步骤，另外可以对标注结果进行初步审核、对标注错误进行修正、对标注结果进行二次审核、对标注结果进行统计和分析。

- **技术选取合理、成熟度高**

现阶段的计算机技术、人工智能技术发展极快，完全支持我们系统核心的开发和运行。我们采用了 html 设计 B/S (Browser/Server)体系结构和 UI 界面用 Javascript 等核心技术处理信息，它们充分支持了我们核心模块的实现，取得了不错的成效。

- 机器学习与模型训练：

在机器学习与模型训练方面，建立标注数据集用于监督学习，选择适当的机器学习模型，如深度学习或传统算法，以完成表格结构化解析任务。

- 算法泛化性：

为了确保算法在不同文件格式和结构下的适应性，需要注重模型的泛化性。不同格式和结构的文件具有一定泛化性，可以适应不同金融公司发布的数据标注平台的多样性。

2.1.2. 使用可行性

- 用户友好性：

为了提高用户友好性，本系统的设计过程中充分考虑到用户的实际使用情况，尽可能直观地提供可视化界面、操作界面。用户能够轻松上传文本文件和图片，并获取清晰的解析结果。使用户容易理解和使用，进一步提升了用户体验。

- 易用性：

系统追求简单操作和即时反馈以增强易用性。用户可以通过简单的操作完成文件上传和数据审核发布，无需繁琐的设置步骤，降低了使用门槛。同时，系统提供即时的反馈，让用户在解析过程中能够及时了解系统的运行状态，增强用户对系统的掌控感。

- **拓展性强，维护性好**：

本系统页面上还有许多空间位置，如果有需求和时间还可以不断增加功能。本项目当前只针对恒生电子金融科技公司。本项目不需要经过复杂的修改，只需要获得数据，就可以扩展到不同的公司，为不同公司服务。

2.1.3. 市场可行性

- 市场需求：

市场上存在明显的金融行业需求和用户需求。金融机构急需一个方便、快捷、高效的标注工具，使得标注过程更加高效、准确和可靠的解决方案，以提高工作效率和减少人工错误率。同时，投资者和分析师对获取金融产品关键信息的需求不断增加，系统的解决方案正好能够满足这一迫切需求。

- 竞争分析：

在竞争分析中，目前市场上缺乏专门针对文本、图片、文件数据标注的解决方案，系统因此具有先发优势。其差异化优势体现在高准确性的抽取、用户友好的界面和良好的集成性，使其在激烈的市场竞争中脱颖而出。

- 潜在用户群体：

潜在用户主要涵盖金融机构和投资者、分析师两大领域。金融机构，包括公募基金公司、保险公司、银行等，是系统的主要潜在用户，因为系统提供了一个方便、快捷、高效的标注工具。投资者和分析师也是系统潜在的用户群体，他们需要分析大量数据，系统的解决方案为他们提供了强大支持。

数据标注的用途十分广泛，如果对当前数据标注平台进行优化，在社会中将有更多的潜在用户。

- 法规和政策环境：

系统的市场可行性与法规和政策环境密不可分。为了确保系统的合规性，必须符合金融行业的相关法规和政策，特别是数据处理和存储方面。同时，政府对金融科技的支持和推动也为系统提供了积极的市场环境，为其发展创造了有利条件。

2.2. 需求调研模型

- 需求调研

本项目组首先进行初步调研，主要方式是查询和使用各大现有的国内外数据标注系统，发现其优缺点，通过搜集政府投诉痛点、询问发包方掌握确实需求。并通过科学的需求分析技术一一列出需求矩阵，对需求分析进行完善。

- 快速原型

本项目的快速原型旨在迅速展现数据标注平台系统的核心功能和用户交互。通过草图绘制、交互元素设计和基本页面链接，我们将创建一个简化的系统模型，突出关键功能如指标定位和结构化解析。通过用户反馈和多次迭代，快速原型将逐步完善，最终为项目的设计和开发提供具体可视化的基础。

该原型将呈现用户通过数据标注中关键指标进行快速定位和结构化解析的流程。用户能够模拟浏览系统，体验基本交互，并通过收集的反馈不断优化细节，确保原型反映项目的核心功能和用户期望。

该快速原型工作机制如下图所示：

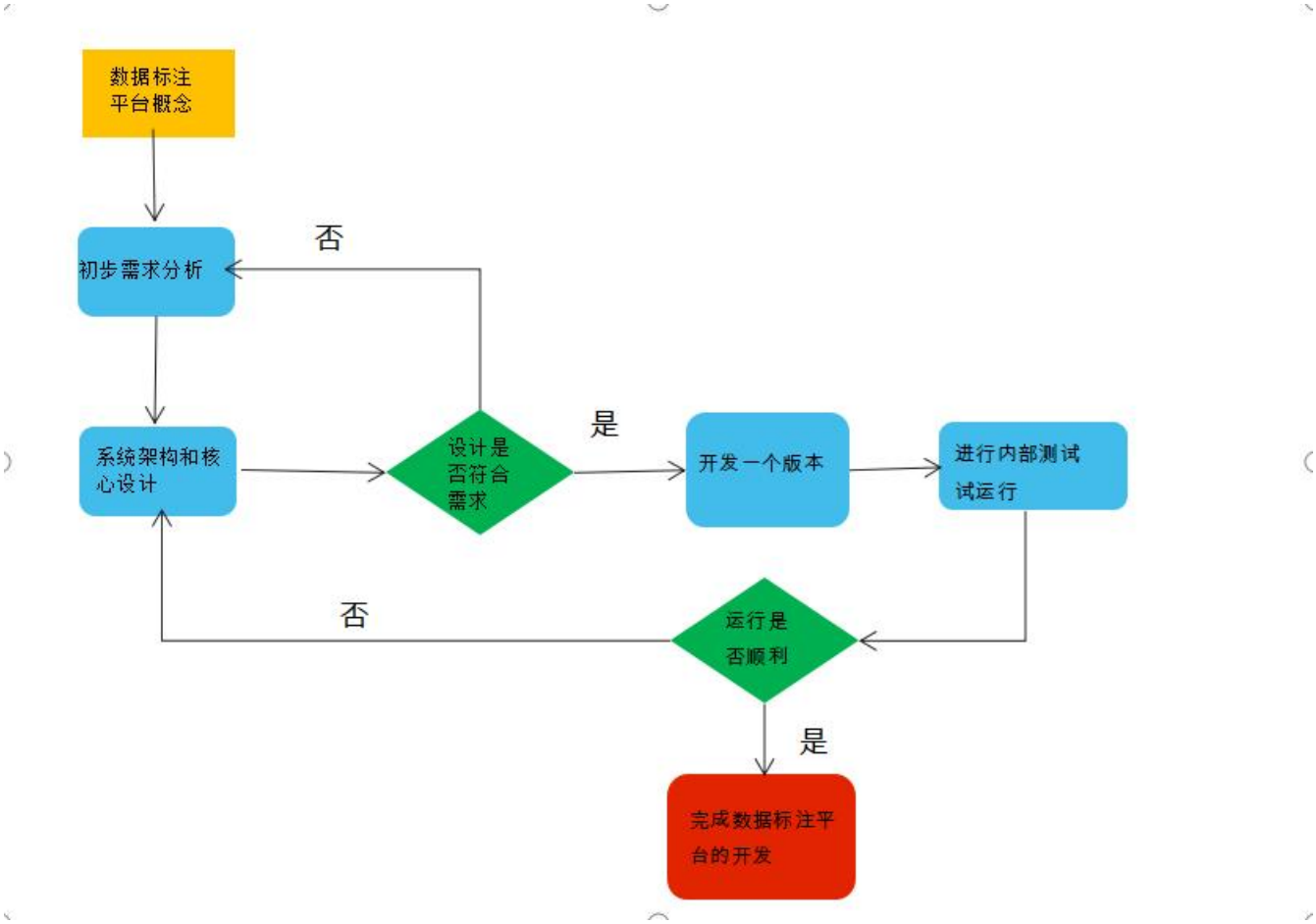
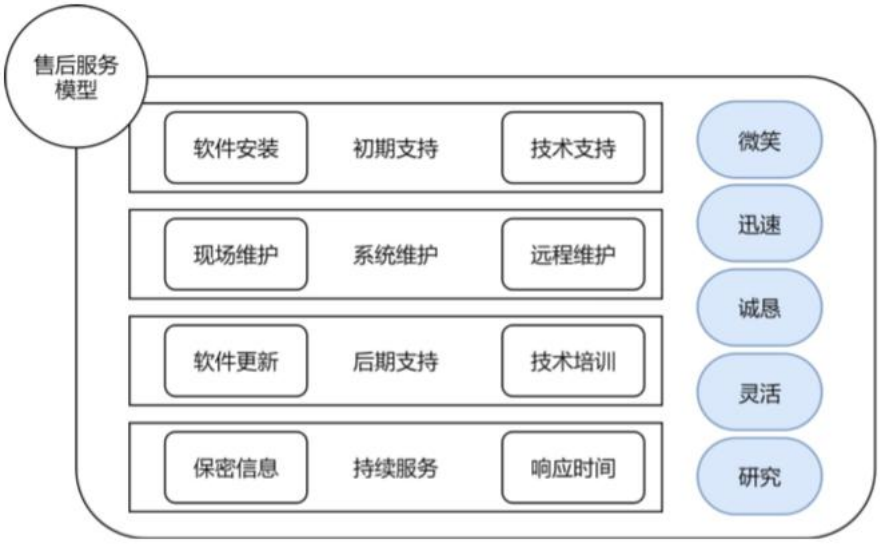


图 2-2 快速原型工作流程图

本系统设计，项目组采取附加策略的原型开发。项目中的 UI 设计人员，通过前期的需求分析，在开发前期迅速做出一份系统的快速原型。并由随机抽取的工作人员来测评，并根据工作人员的反馈，对系统架构和核心设计进行相应调整，并进行新的版本开发，直到能够符合工作人员的相应需求，完成最终版本的提交。

数据标注系统旨在为研究者和企业提供一个方便、快捷、高效的标注工具，使得标注过程更加高效、准确和可靠。有助于金融工作者从大量繁琐的数据标注流程中脱身出来，专注于其他更为重要的工作。在金融工作者的工作中起到了协助作用。

2.3. 售后服务模型



图表 2-3 售后服务模型图

团队为相关行政机构初期制定一套售后服务协议方案，包括：系统安装、技术支持、远程维护、现场维护、系统更版、技术培训、响应时间、保密信息等。供客户进行审核，与客户协商，删除不必要的服务选项，增加客户提出的合理要求，为外包方量身打造一款专属自己的售后服务体系，承诺将尽团队最大的努力实现客户的一切合理要求，制定相关售后服务协议，严格遵守协议规定。

3. 项目价值

3.1. 项目难点分析

该项目面临一些挑战和难点，主要涉及技术、数据和业务方面：

- 标注质量管理：

  标注过程可能受到标注者主观判断的影响，导致标注结果的不一致性。确保标注者之间有一致性的标准和培训是关键的。对于复杂的标注任务，标注者可能会遇到困难，导致标注质量下降。这可能需要更详细的标注指南和支持。

- 标注效率和成本：

  大规模数据集的标注可能需要大量的时间和资源。标注者需要不停优化标注流程、使用自动化技术以及合理安排。人力成本和工时管理也是一个挑战，需要平衡标注的速度和质量，以最小化成本。

- 数据隐私和安全：

处理金融文件时，涉及的信息往往非常敏感，包括个人身份信息、财务数据等，这些信息一旦泄露或被不当使用，可能会对用户造成严重的困扰和损失。因此，建立一套高度安全的数据处理流程是至关重要的，以确保用户隐私得到充分保护，防止标注数据的泄露或滥用，特别是当标注工作外包给第三方时。

- 技术挑战：

处理同时包含文本、图像、音频等多模态数据的标注任务可能更为复杂，需要整合不同类型的标注工具和专业知识。特别是在处理复杂任务和非结构化数据时，尝试使用自动化技术进行标注可能面临挑战。

- 数据多样性：

针对特定领域的标注可能需要专业知识，而这可能不容易获取。使用标注平台需确保标注者具有足够的领域专业知识是一个挑战。

解决这些挑战需要综合考虑人员培训、技术工具的选择、有效的项目管理和沟通等因素。在项目启动前，充分了解和规划这些方面可以帮助提前预防一些潜在的问题。

以上挑战需要团队充分理解金融领域的特点，采用前沿的技术手段，并进行不断的迭代和优化，以确保系统在复杂的金融环境中稳健运行。

3.2. 项目优势分析

- **需求定位准确**

现在市场上的数据标注系统尚且不多，金融科技公司能有效利用起来的更是少之又少，手动处理工作量大，效率低下，因此此处需求空缺严重，需要填补。本项目则恰当的满足了金融科技公司处理大量繁琐数据标注时效率不高的痛点。

- 强大的数据标注能力：

项目专注于解决使得标注过程更加高效、准确和可靠的难题，这对于金融文档中广泛标注分析数据是至关重要的。通过设计强大的信息识别和精确标注，项目能够准确地从复杂的数据结构中提取关键信息进行标注。

- 实时性和高效性：

金融行业对信息实时性的要求意味着项目需要具备高效的处理能力。通过优化算法和系统设计，确保在短时间内标注大量数据，满足金融领域对实时性的严格要求。

- 机器学习模型泛化性：

项目对抽取模型的要求包括高泛化性，即使在面对新的文档结构和内容时也能保持准确性。这种泛化性能够使项目适应金融行业不断变化的数据和文档格式，为未来的应用提供了可靠的基础。

4. 创新点

1. 自动标注和半自动标注：

引入机器学习算法实现数据的自动标注，并提供半自动标注工具以加速标注过程。

2. 多模态数据标注：

支持图像、文本等多种数据类型的标注，以满足不同领域和任务的需求，让标注人员能够有效地处理不同类型的信息。

3. 标签的不确定性处理：

支持对标签的不确定性建模，特别是在模糊或困难情况下，使系统能够更灵活地处理标注结果。

4. 元学习和主动学习：

在之后的功能中打算利用元学习算法，通过标注人员的反馈不断改进模型，以更好地适应新的标注任务。并引入主动学习，让系统能够主动选择最具信息量的样本，以降低标注的成本。同时提供标注人员之间的反馈机制，以不断改进标注准确性和效率。

5. 安全和隐私保护：

采取措施保护敏感信息，例如对图像中的个人身份进行模糊处理。遵循隐私法规，确保用户数据的安全性和隐私保护。

6. 开放性和可扩展性：

提供 API 和插件系统，使得其他应用能够方便地集成和扩展标注平台的功能。支持自定义标注任务和工作流，以适应不同行业和应用领域的需求。