

## B题 ESG报告数据智能提取与分析：基于大模型的定量与定性指标识别

### 一、赛题题目

#### ESG 报告数据智能提取与分析：基于大模型的定量与定性指标识别

### 二、背景介绍

随着全球对可持续发展的重视日益增强，环境、社会与治理（ESG）已成为衡量企业长期价值和社会责任的重要标准。企业在披露 ESG 信息时，通常以年度 ESG 报告或社会责任报告的形式发布，内容涵盖温室气体排放、能源消耗、员工培训时长、董事会多样性、反腐败措施等多项指标。这些报告多以 PDF 格式发布，内容冗长、结构多样，且包含大量非结构化文本和表格信息。

ESG 报告的阅读和分析对投资者、评级机构、监管部门和学术研究者具有重要意义。然而，当前主流的 ESG 信息提取仍依赖人工阅读和手动录入，效率低下且易出错。尤其是在需要同时提取定量指标（如碳排放量、女性高管比例）和定性指标（如 ESG 战略描述、治理结构说明）时，传统规则模型难以覆盖多变的语言表达和排版方式。

### 三、问题提出

基于上述背景，现向各参赛队提出如下任务，要求参赛团队结合大模型技术，设计并实现一套从 ESG 报告中高效、精准提取定量与定性指标的智能系统。

#### 1、数据采集与预处理

在合法合规的前提下，自主采集不少于 100 份 A 股或港股上市公司发布的 ESG 报告（PDF 格式）。对采集到的文档进行格式统一、文本提取、表格识别等预处理操作，确保后续算法可有效处理。

#### 2、指标体系构建与标注

构建一个包含不少于 50 个 ESG 关键指标的抽取体系，指标应覆盖 E（环境）、S（社会）、G（治理）三个维度，并明确区分：

- 定量指标：如温室气体排放总量（吨 CO<sub>2</sub>e）、能源消耗总量（MWh）、员工培训总时长（小时）、女性高管占比（%）等；
- 定性指标：如 ESG 战略目标、气候风险管理措施、员工权益保障政策、反腐败机制描

述等。

鼓励参赛团队结合实际应用场景扩展指标范围。

### 3、基于大模型的结构化提取算法开发

利用开源或商用大语言模型（如 Deepseek、通义千问、Llama 等），结合提示工程、微调、检索增强生成（RAG）等技术，开发高精度的 ESG 指标提取算法。具体要求包括：

- 对定量指标，提取数值及其单位，并保留上下文以便验证；
- 对定性指标，提取原文表述段落，保留其语义完整性；
- 若报告未提及某指标，字段值应设置为空；

支持对表格、图表中数据的识别与提取，提取结果以结构化格式（如 JSON、CSV）存储，便于后续分析与展示。

### 4、成果深化与应用系统开发（可选）

完成上述任务视为成功参赛。鼓励参赛团队进一步开发可视化分析系统，支持用户对 ESG 指标进行查询、对比、趋势分析等功能。系统可展示企业在不同年度的 ESG 表现，支持多维度筛选，辅助投资决策与政策研究。

## 四、作品评价标准

### 1、撰写规范性（20分）

（1）参赛报告结构完整，包括以下内容

- a) 项目概述：项目背景、应用行业、算法或模型优势等；
- b) 解决方案：方案设计、方案功能、关键技术、算法实现过程、结果分析等；
- c) 应用价值：经济效益、社会效益分析等。

（2）参赛报告格式规范，术语、图表、公式准确规范。

（3）所提交代码清晰、详尽、可读性强，能准确帮助评审人理解算法的工作原理和实现过程。

### 2、算法效果与效率（50分）

（1）参赛报告需要清晰展示算法的效果，包括使用合适的评估指标对模型进行量化评估，测试算法的准确性。

（2）参赛报告需要清晰展示算法在不同条件下的表现，以及对异常输入的处理能力，

给出测试报告。

(3) 评估算法的运行时间和资源消耗，包括计算速度和内存使用情况。

### **3、文档可读性和用户指导性（20分）**

(1) 除参赛报告外，提交的算法、系统实现的源代码、可用于验证的原始数据集以及必要的算法、系统运行说明文档等附件的可读性强，逻辑连贯，论述层次清晰，语言表达简洁明了。

(2) 参赛报告应详细解释算法的工作原理、使用方法以及可能的配置选项，确保用户能够无障碍地安装、配置和运行算法。并能为用户提供足够的背景信息和示例，帮助用户深入理解算法的应用场景。

### **4、结果展示形式的丰富性（10分）**

评审将关注作品结果展示形式多样、创新，鼓励参赛团队使用多种方式来呈现结果。结果展示可以包括图表、动画、交互界面、软件系统等多种形式，以直观、生动地展现算法的效果和优势。展示形式应清晰易懂，能够突出关键信息，帮助评审人和用户更好地理解作品。

（出题单位：上海华证指数信息服务有限公司）