大模型智能合同审查任务

1. 任务基础说明

同学你好，欢迎参加本次大模型智能合同审查任务！

要想让大模型以超高的准确率审查合同存在非常多的工程化手段，但这次我们时间有限，因此我们将任务进行了简化，只完成一个基础 demo，而且只从合同中提取“含税总金额”这一项信息。

你也许会在这个项目中接触到自己未尝试过的技术，但通过题目的引导，这些技术的实践都不会很复杂，你也会在这个项目的过程中对多种工程手段收获具象的感知。

本次任务可以分成以下几个步骤：

1. 版面分析 OCR
2. 版面恢复
3. 使用 Prompt 提取信息
4. 高亮关键信息

任务所需的合同图片在百度网盘，共四张经过马赛克脱敏的图片：

合同审查样本_免费高速下载|百度网盘-分享无限制 (baidu.com)

提取码：lbed

2. 版面分析 OCR

对于图片这种非结构化数据，我们要从中提取数据，首先要通过 OCR 识别文字。当前 OCR 不仅可以识别普通文字，还可以对整个图片的版面进行分析，识别出标题、页眉、页脚、表格、

印章等。本次任务我们主要需要 OCR 识别到的表格，因为含税总金额大多数都在表格中。

你可以调用百度智能云的办公文档识别接口识别这几张合同图片，新用户均有免费识别额度可以使用：

办公文档识别 - 文字识别 OCR (baidu.com)

OCR 看似是传统技术，但实际上百度的 OCR 水平在各家云厂商中遥遥领先，我们也与百度有非常深入的合作。在办公文档识别这个接口中，你可以看到它返回的参数中包含单字置信度（char_prob），这就是我们依靠当前项目推动百度在三个月内上线的一个关键新功能，它可以用于 OCR 识别可靠程度的回溯，也可以用于 NLP 矫正 OCR 结果，不过它不会用于本次任务中。

3. 版面恢复

OCR 接口返回的只有 JSON，如果要让大模型有效提取信息，还需要把 JSON 结果转换成格式化的文本。这个步骤就叫版面恢复，也就是将识别 JSON 结果恢复成原本版式或格式的纯文本。由于本次任务中的“含税总金额”都在表格中，所以我们只需要将表格转换成格式化的文本，不需要处理其他识别结果。

将表格处理成格式化的文本有两种方法：使用 Markdown 或者使用 HTML。使用 Markdown 非常简单，但不能表示合并单元格。而在合同中，合并单元格往往有重要的结构信息，因此我们选择将其转换为 HTML，因为 HTML 是可以表示合并单元格的。当前代码能力较强的大模型，都是有能力从 HTML 中提取信息的。

这一步只需要把识别结果中的表格部分转化成 HTML，每张图片保存一个 HTML 文件，不需要考虑识别出错（例如单元格被错误合并）的问题。contract_1.jpg 的示例 HTML 文本如下：

```
<table border="1"><tr><td rowspan="2">名称</td><td rowspan="2">规格型号</td><td rowspan="2">单位</td><td rowspan="2">数量</td><td colspan="3">不含增值税单价（元)</td><td rowspan="2">不含增值税总价（元）</td></tr><tr><td>出厂价</td><td>综合费用</td><td>综合单价</td></tr><tr><td>矿粉</td><td>S95 散装</td><td>吨</td><td>18240</td><td>368.00</td><td>50.00</td><td>418.00</td><td>7624320.00</td></tr><tr><td colspan="2">增值税</td><td>税率</td><td colspan="2">13%</td><td>税额</td><td colspan="2">991161.60</td></tr><tr><td colspan="2">含增值税合同总价（大写）：</td><td colspan="4">捌佰陆拾壹万伍仟肆佰捌拾壹元陆角整</td><td colspan="2">小写（元）：8615481.60</td></tr></table>
```

使用浏览器打开，上述 HTML 文本显示如下：

名称	规格型号	单位	数量	不含增值税单价（元）			不含增值税总价（元）
				出厂价	综合费用	综合单价	
矿粉	S95散装	吨	18240	368.00	50.00	418.00	7624320.00
增值税		税率	13%		税额	991161.60	
含增值税合同总价（大写）：		捌佰陆拾壹万伍仟肆佰捌拾壹元陆角整				小写（元）：8615481.60	

4. 使用 Prompt 提取信息

现在我们已经将完全非结构化的图片转换成了结构化的 HTML 文本，接下来我们就可以编写 Prompt 并使用大模型提取信息了。

请你编写 Prompt 模板，并将表格 HTML 拼在 Prompt 模板中，让大模型提取出“含税总金额”的纯数字，并让大模型以 JSON 格式输出提取结果，例如：

```
{
    "含税总金额": 8615481.60
}
```

大模型可以使用目前百度智能云免费的 ERNIE Speed 模型：ERNIE-Speed-8K - 千帆大模型平台 | 百度智能云文档 (baidu.com)

如果觉得这个模型性能不够，无法有效提取信息，也可以使用 ERNIE-4.0 模型：

ERNIE-4.0-8K-Latest - 千帆大模型平台 | 百度智能云文档 (baidu.com)

之所以让大模型以 JSON 返回结果，是因为如果让大模型随意地输出自然语言，后续程序依然无法有效提取信息，只有输出 JSON 才能衔接后续的程序。而且对于大模型来说，输出 JSON 也是它非常擅长的事情。当前 GPT 调用各种工具，例如查询网页、调用插件等等，也都是通过 JSON 传递信息。

在信息提取或信息判断比较简单时，我们直接使用一个比较完善的 Prompt 就可以完成，这样的好处在于开发迅速而且可以快速响应不同的需求。只有在任务比较困难，且对准确率要求极高时，我们才会微调大模型(SFT)。比如在另外一个项目中，我们通过编写 1700 余字的 Prompt 和使用数十万条数据微调大模型，将一项很复杂的信息提取任务的准确率从传统语言模型的 89%(该模型由某家知名 NLP 人工智能公司交付)，提升到了惊人的 99.97%！

5. 高亮关键信息

完成信息提取后，还需要高亮合同中的“含税总金额”，方便人工校对提取结果。需要你用程序根据提取到的“含税总金额”从 OCR 返回的 JSON 中找到对应的像素区域，并使用程序在图片中高亮该区域。

至此，你已成功创建了一个基本的大模型智能合同审查应用，初步领略到了大模型在 ToB 领域的巨大潜力。如果你成功通过本次笔试面试，并决定成为我们团队的一员，你将见证更多大模型和其他人工智能技术如何在 ToB 领域落地生根。

6. 提交要求

请将程序封装为.py 文件，输入为提供的四张图片，输出为 4 个 HTML 文件和 4 个经过高亮的图片。代码的规范程度是我们关注的指标。此外，我们也欢迎你使用大模型（包括 Copilot）

辅助编程，如果能很好地使用大模型辅助编程会是加分项，可以将与大模型的对话记录或使用方法与其他文件一并提交。**请在收到笔试题目的 5 天内提交结果。如果有特殊情况无法及时提交（如期末考试），也可以邮件告知我。发送邮箱为 liuzhengze@yljr.com。邮件主题为“姓名+笔试提交”，整个文件夹压缩为压缩包，压缩包同样命名为“姓名+笔试提交”。**

具体提交格式如下：

1. **代码文件**:

- 包含主程序逻辑的 `main.py` 文件

2. **输入**:

- 提供的四张合同图片应放置在 `input_images` 文件夹下

3. **输出**:

- 4 个转换后的 HTML 文件，应放置在 `output_html` 文件夹下，命名为 `contract_1.html`, `contract_2.html`, `contract_3.html`, `contract_4.html`
- 4 个经过高亮处理的图片，应放置在 `output_images` 文件夹下，命名为 `contract_1_highlighted.png`, `contract_2_highlighted.png`, `contract_3_highlighted.png`, `contract_4_highlighted.png`

4. **大模型对话记录（可选加分项）**：

- 与大模型的对话记录或使用方法文档，请放置在 `ai_assistance` 文件夹下。

提交格式示例:

```
- project_folder/
    - main.py
    - input_images/
        - contract_1.jpg
        - contract_2.png
        - contract_3.png
        - contract_4.png
    - output_html/
        - contract_1.html
        - contract_2.html
        - contract_3.html
        - contract_4.html
    - output_images/
        - contract_1_highlighted.png
        - contract_2_highlighted.png
        - contract_3_highlighted.png
        - contract_4_highlighted.png
    - ai_assistance/ (可选加分项)
```

期待你的精彩表现，祝你好运！