**数据工程师笔试题F题**

**长文本翻译与信息抽取**

有一通过 OCR 获得的 Markdown 文档，**结构相对整洁，含图片，约 11w tokens**，其主要内容是 2024 年国际数学奥林匹克竞赛的预选题。在文档中，一道题目一般可以被划分为 problem solutions answer三个部分，现要对其中的所有题目进行**结构化解析并翻译成中文**。

|  |
| --- |
| Markdown 待处理文档示例（详情请看 📎 附件） # Problem Selection Committee  ![](images/4a25c17d8e5d644bc1fc27cad1ac18895c2d2a125e774ac92c0bd73f8adc551c.jpg)  Aron Thomas, Yinghua Ai, Andrew Ng, Géza Kos, Ivan Guo, Alice Carlotti, James Aaronson, Sam Bealing, Adrian Agisilaou, James Cranch, Joseph Myers (chair),Harvey Yau, Maria-Romina Ivan,Michael Ren, Elisa Lorenzo Garcia  # Problems  # Algebra  # A1.  Determine all real numbers $\alpha$ such that the number  $$ \lfloor \alpha \rfloor + \lfloor 2 \alpha \rfloor + \cdot \cdot \cdot + \lfloor n \alpha \rfloor $$  is a multiple of $n$ for every positive integer $n$ . (Here $\lfloor z \rfloor$ denotes the greatest integer less than or equal to $z$ ） |

已知 2024 年 Short list 包含 31 道不同的题目，每道题依序在全文中出现两次。Problems 章节只含题干，Solutions 章节则完整包含一个或多个题解，参考答案作为可选字段出现。因此候选人须重点关注 Solutions 章节下的内容，完成以下两个任务：

**任务一**

全文翻译：需要构建中文语料，因此要对整篇文档进行全文翻译；

* 要求：须保留 Markdown 格式，并且更关注语义正确性，过程结果命名为 full\_zh.md一并提交；
* 提示：可自行选择合适的翻译工具（例如通用 LLM api 等）；

**任务二**

文档解析：需要构建 QA 对形式的预训练语料，所以还需要将通篇的文档转化为程序友好的结构化形式；

* 要求：按题目顺序实施结构化解析，返回一个 JSON数组，其中每个条目都代表一道原题，清除无关内容，只保留问题题干、题解和最终答案；
* 提示：原题样例可参考附件中的 PDF，可以尝试滑动窗口、XML 注入等手段；

|  |
| --- |
| JSON 目标格式.json [  {  "problem": "求出所有实数 $\alpha$...", // 题干部分  "solutions": [  "从一系列的实数开始: ...", // 题解 1  "我们发现确实存在一些函数", // 题解 2/3/4/... 如果有  ],  "final\_answer": "$n$" // 未明确声明则留 null  },  ... ] |

**加分项**

* 将图片通过 Base64 的形式，合理地存储在 JSON 中；
* 如若使用 LLM，最大限度地节省 Tokens 消耗；

**附件**

* 下载与说明：
* [📎 full.pdf](https://molar-public.oss-cn-hangzhou.aliyuncs.com/temp/yy/Python/full.pdf) # 原文档，仅作参考
* [📎 full.md](https://molar-public.oss-cn-hangzhou.aliyuncs.com/temp/yy/Python/full.md) # 待处理 OCR 全文
* [📎 images.zip](https://molar-public.oss-cn-hangzhou.aliyuncs.com/temp/yy/Python/images.zip) # 相对路径下的图片资源
* 01.jpg
* 02.jpg

**成果要求**

|  |
| --- |
| 答案不需要完美，但是要有完整的思路 |

1. 解题思路：README.md，简明扼要地阐述你的设计、操作中遇到的问题及解决，好的文档往往让人眼前一亮；
2. 过程代码：虽然代码不是核心关键，但却是工作开展的基础，我们希望你有清晰的思路和整洁的编码风格；
3. 翻译结果 + 目标 JSON：字段合理、结构清晰、内容完整地安放你的最终结果；

|  |
| --- |
| Python 成果目录样例.zip \* src/\* # 假如你用的是 src layout \* results/  \* full\_zh.md # 翻译结果  \* full\_zh.json # 最终结果 \* README.md # 解题思路 \* main.py # 如有 |