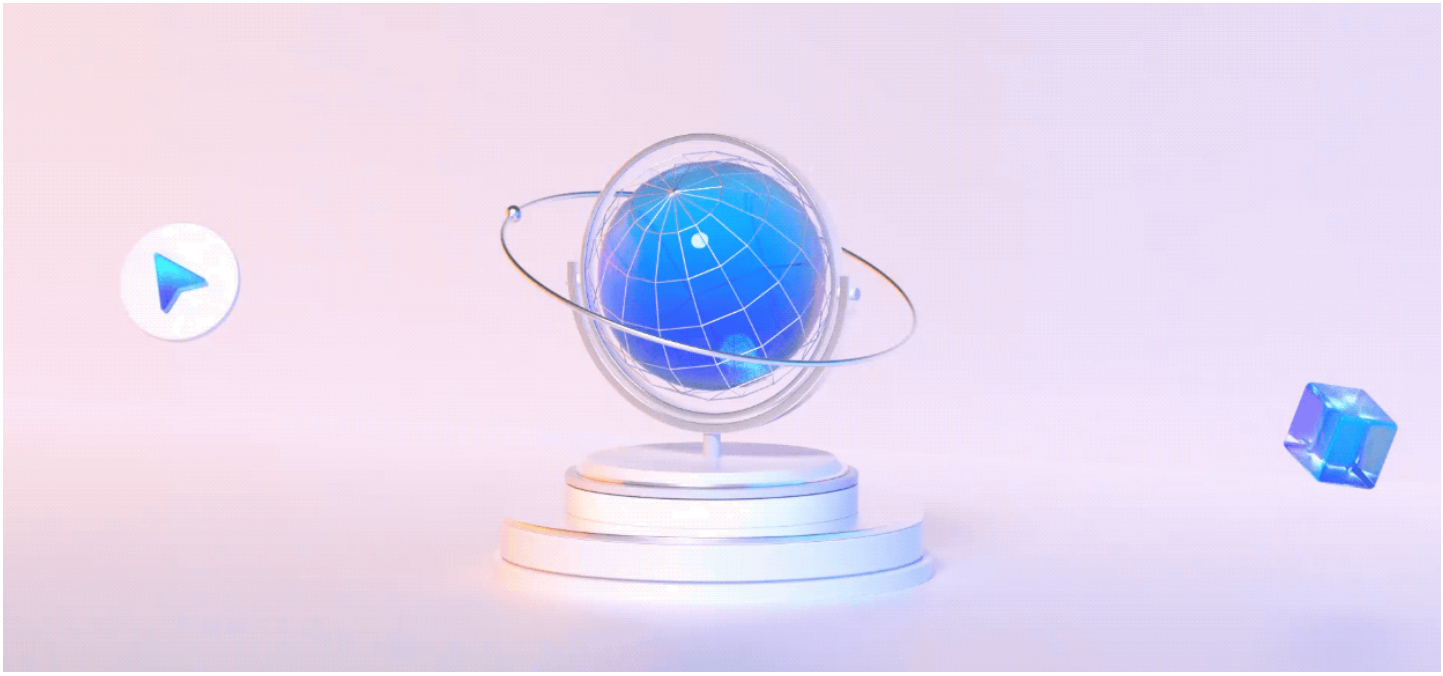


数据标注SOP

一、业务背景与各流程呈现方式

将整套PDF试题处理成单题目问答形式的图片与代码，用于丰富数据库与模型训练，方便用户在APP题库中精准的查找相关信息

1）原始数据呈现方式：

一、填空题（本题共 5 小题，每小题 2 分，共 10 分）

得　分	

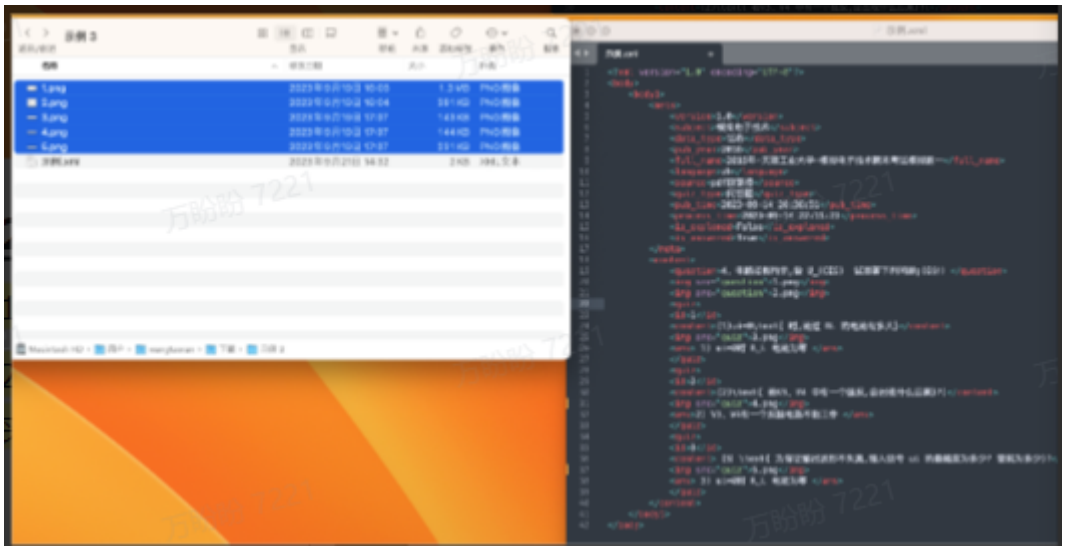
1．若 $f(x)=\begin{cases} e^{x}, & x<0 \\ a-bx, & x\geq 0 \end{cases}$ 在 $x=0$ 处可导，则 $a=$_____，　$b=$_____.

2．设函数 $y=y(x)$ 由方程 $y=\sin(xy)$ 确定，则 $y'=$ _____.

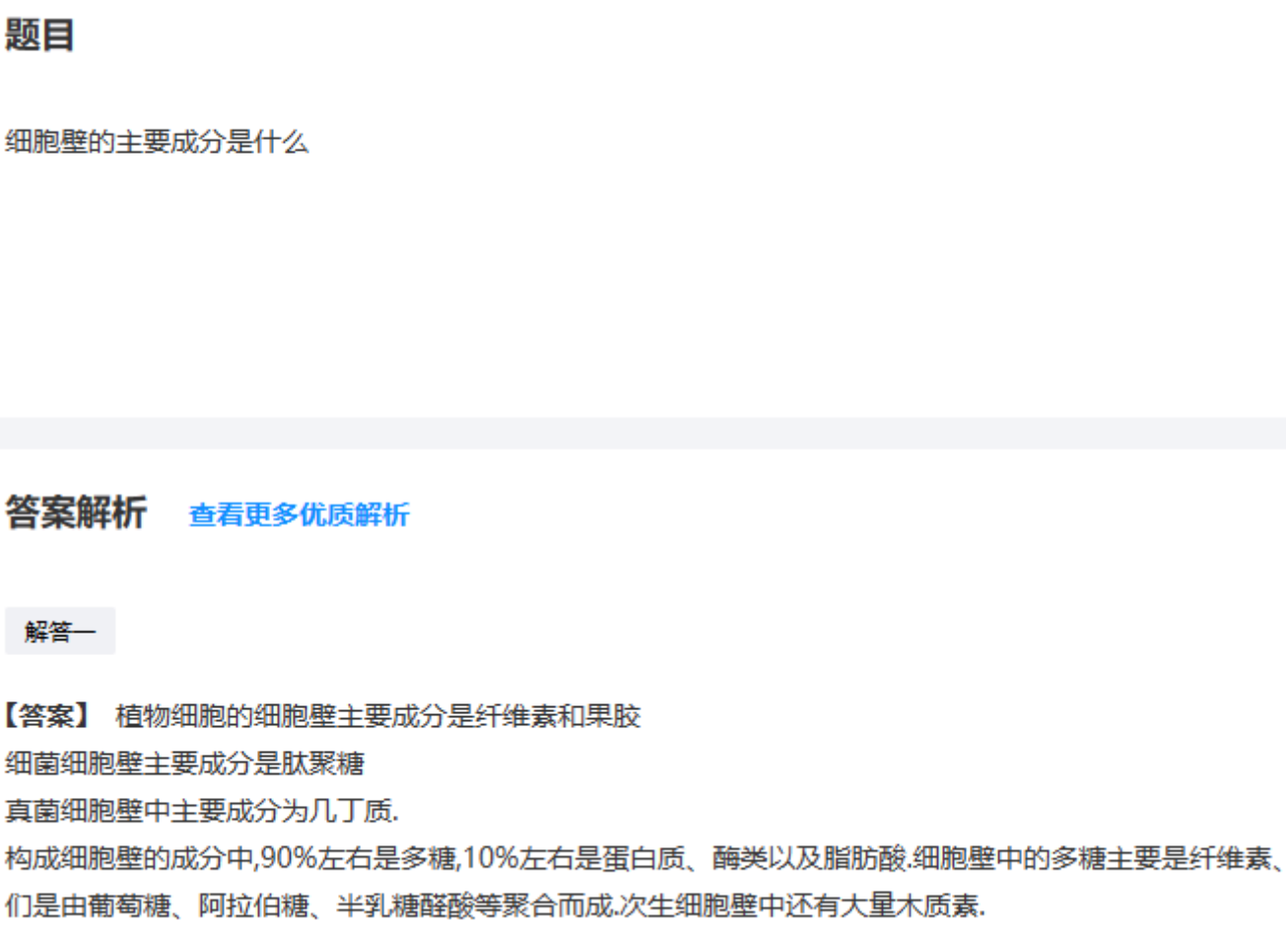
3．若函数 $f(x)$ 为连续奇函数，则 $\int_0^x f(t)\mathrm{d}t$ 为_____函数．（填 “奇”或“偶”

4．若曲线 $y=\dfrac{x^2+3x+k}{x^2-1}$ 恰有两条渐近线，则常数 $k=$ _____.

2）业务交付要求：单题目切图+代码展示



3）用户端展示



二、标注要求

1）切图：

通过截图的方式将整套题目进行拆分，需注意保证题目的完整性和独立性（重点注意大题多问需要进行整题截图PS:完整性与小题截图PS:独立性）以及判断大题多问独立切图的逻辑性（是否通过题干可

判断为独立作答题目）

2）标注结构：

在原有代码的基础上，判断题目的属性，并将题目信息准确录入代码中

基础代码示例：

2、漂移电流是（反向　　）电流，它由（ 少数 ）载流子形成，其大小与（ 温度
）有关，而与外加电压（ 无关 ）。

```xml
<meta>
    <version>1.0<version/>
    <subject>模拟电子技术</subject>
    <data_type>试卷</data_type>
    <pub_year>2016</pub_year>
    <full_name>2016年-天津工业大学-模拟电子技术期末考试模拟题一</full_name>
    <language>zh</language>
    <source>xxxxxx</source>
    <quiz_type>填空题</quiz_type>
    <is_explaned>False</is_explaned>
    <is_answered>True</is_answered>
    <pub_time>2023-09-08 19:00:00</pub_time>
    <process_time>2023-09-08 20:00:00</process_time>
    <additional_info>
        <附加字段示例1></附加字段示例1>
        <附加字段示例2></附加字段示例2>
    </additional_info>
</meta>
<content>
    <question>2、漂移电流是（ ）电流，它由（ ）载流子形成，其大小与（ ）有关，而与外加电压（ ）。</qu
    <ans>反向[space]少数[space]温度[space]无关</ans>
    <img src="question">/path/to/x.png</img>
</content>
```

重点

代码不需要重新编辑，仅需要判断题目的分类，并将各个字段信息填入标准代码中；

判断每种题型，并匹配对应的填写标准；

判断每个字段的含义，并将字段填写到对应的位置；

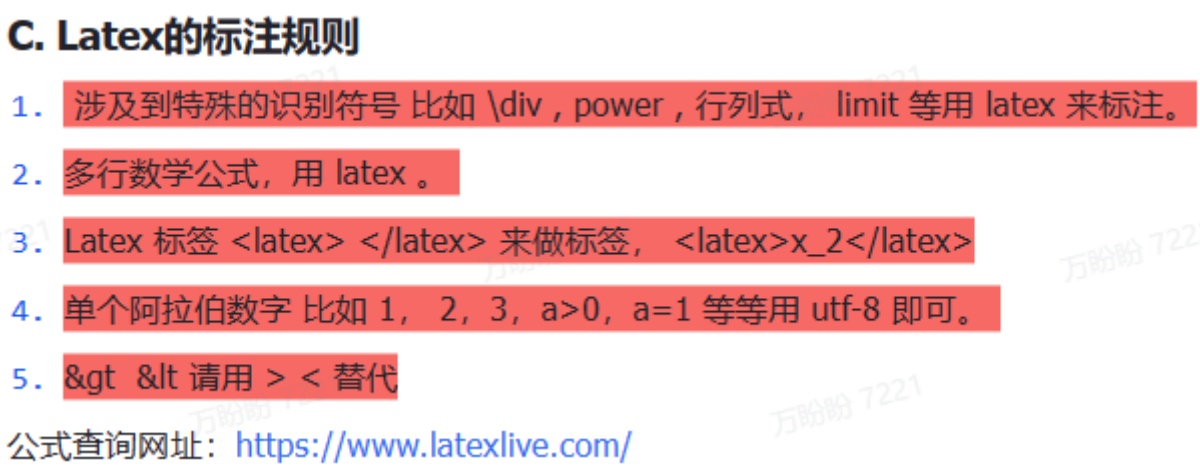
科目必须按照标准的大学专业分类填写（需判断）；

难点

a.单题与大题多问中大小题的判断方式（题型判断与填写标准匹配）

b.图片的插入方式（格式与细节）

c.公式的转换细节



d.是否有quiz的answer填写方式

3)交付标准

交付形式：xml数据及图片数据；按照交付的压缩包ZIP（XXX-交付-年月日）---PDF文件夹---PDF中每道题的文件夹---图片和xml

北极星指标：完成率100%，正确率95%

注意：每个xml文件夹仅放置一道题目数据，并确保不出现多切图、少切图、错切图、缺少字段、缺少转写、错别字、字段错误、答案题目不匹配、图片来源格式错误等问题

三、实操难点问题记录与解决方案

🥛 提前熟悉VS2010的使用方法，这个很重要！！！

高数

1）公式转换：复杂公式需要手动输入之后在在线LATEX插件中转换；公式转换完成之后需要在代码中加入LATEX标签754.分析下面的语言材料是复句还是句群?

英语

1）图片较多，核对难度大，可在切图过程中直接标注题号，方便快速识别

2）可提前建文档标注各类题型代模板，尽量减少手动输入代码

3）如果quiz部分有答案，正常标注，如果quiz部分没有答案，那就在最后加<ans></ans>，别忘了最后的content（针对每个quiz及question）

4）每个content需对应结尾的content，但是quiz中的content仅代表小题部分的结束，每大题结束的content与开头对应

5）注意代码中的空格

C语言

1）解析的呈现方式：在ans的后面写<exp>解析内容</exp>

2）答案中包含图片的呈现方式：需要截图，单独按照图片插入的方式引入

3）源文件内容无法复制时需要手动输入，可截图识别，注意错别字检查

四、培训难点问题及问题发现

1）切图中若有答案，答案需要抹除，截图之后白色笔迹直接抹除即可

2）转换类型多变，尽量在文档中找到对应的问题与答案对应进行标注，若仅有问题无答案，按照无答案标注，若仅有答案没有问题，进行答案舍弃

eg:非简单问答形式的题库，例如实验方案类的切片方式示例

3）关于着重号、尖括号及各种公式的转写规范示例