

# 隐私保护的人工智能生成内容研究 与设计

指导老师：唐屹教授

班级：信安201





姓名：李文杰



# 背景

相比传统网络空间，AI 模型面临的攻击面是不同的、全新的。

## 存在问题

-  **隐私数据泄露** - OpenAI 会将用户的训练数据提供给其他公司
-  **成员推断攻击** - 判定某些特定数据是否在指定的训练集中
-  **数据投毒** - 攻击者向 AI 模型注入恶意训练数据
-  **提示注入攻击** - 构造特定的措辞引诱模型作出回答

## 数据来源

- **ChatGPT:** ChatGPT 是由 OpenAI 开发的一款自然语言处理模型
- **ChatGLM:** ChatGLM 是国内清华大学AI团队打造一款新一代认知智能大模型

在本毕业设计中，选择ChatGPT大模型平台作为第三方模型接口

## 例子

下图是使用 ChatGLM 作为演示例子

👤 用户\_Oplado

现在有二十位小学六年级同学的体重：33, 35, 37, 37, 38, 39, 40, 41, 41, 41, 43, 43, 44, 45, 45, 45, 45, 46, 47, 48，帮我计算他们的平均值。

🌀 ChatGLM

☑ 代码生成：已完成 ✓

结果

41.65

这组二十位小学六年级同学的体重的平均值为41.65千克。

👤 用户\_Oplado

刚刚我说的一个学生的体重是多少

🌀 ChatGLM

您之前提到的体重数据是：33, 35, 37, 37, 38, 39, 40, 41, 41, 41, 43, 43, 44, 45, 45, 45, 45, 46, 47, 48。这些数据中，没有单个学生的体重被明确指出。如果您需要计算特定学生的体重，请提供更多信息或指定一个特定的数据点。

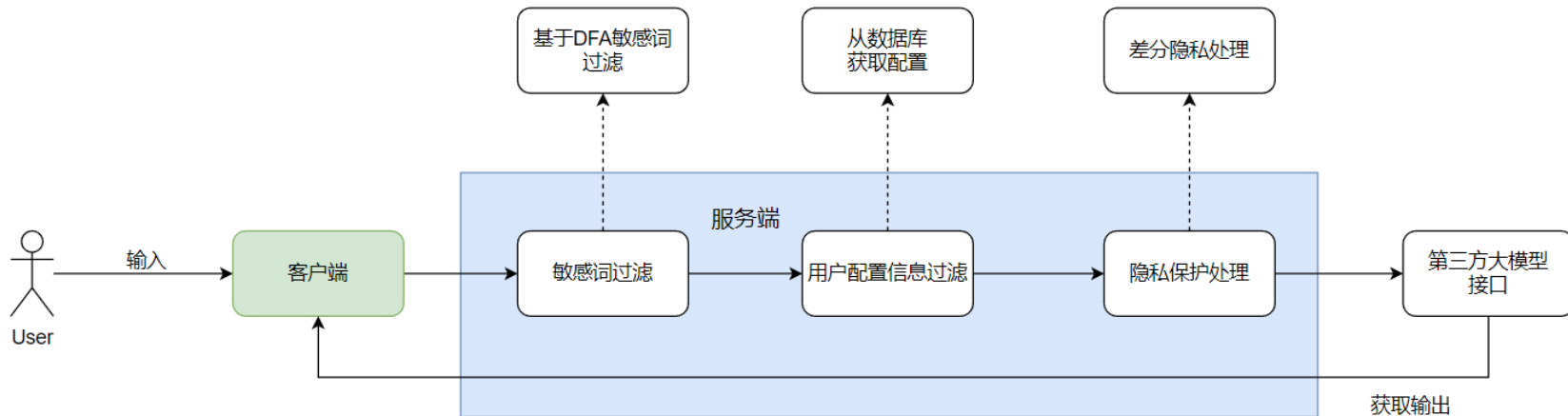
# 系统设计与实现

技术选型与系统架构设计

## 技术栈

- 前端：React
- 后端：SpringBoot + MySQL

## 系统架构设计图



# 后端功能设计——敏感词处理

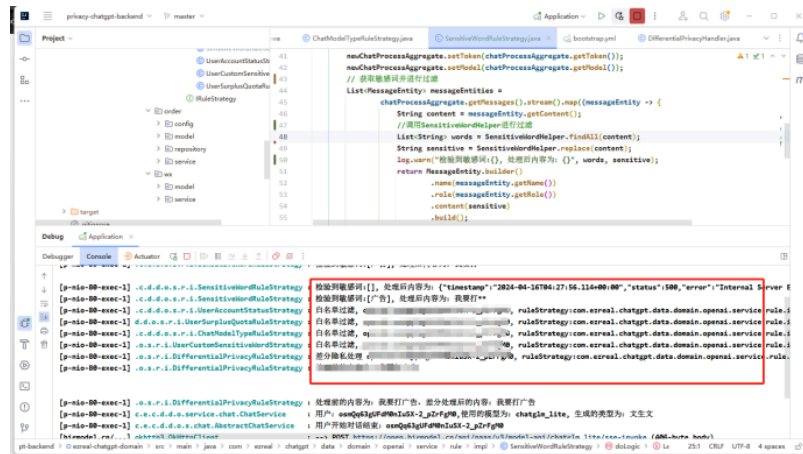
基于开源敏感词库sensitive-word实现敏感词过滤

## 第三方敏感词库依赖

```
<dependency>
<groupId>com.github.houbb</groupId>
<artifactId>sensitive-word</artifactId>
<version>0.16.1</version>
</dependency>
```

- 基于DFA算法，性能为 7W+ QPS，应用无感
- 支持敏感词的判断、返回、脱敏等操作

## 实现效果



在上述图片中，向系统输入敏感词广告，该敏感词被准确地识别出来

# 后端功能设计——处理指定隐私数据

过滤用户自定义隐私保护数据

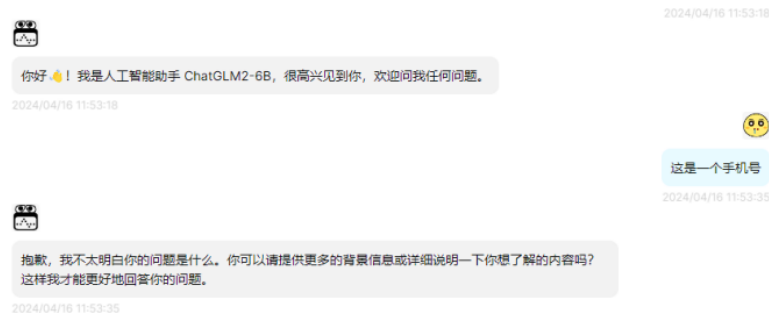
## 实现思路

系统读取用户的敏感词，配置到MySQL数据库中，系统在处理前端输入时会从数据库中获取用户指定的配置信息。这步的作用主要是过滤字符类型的输入。

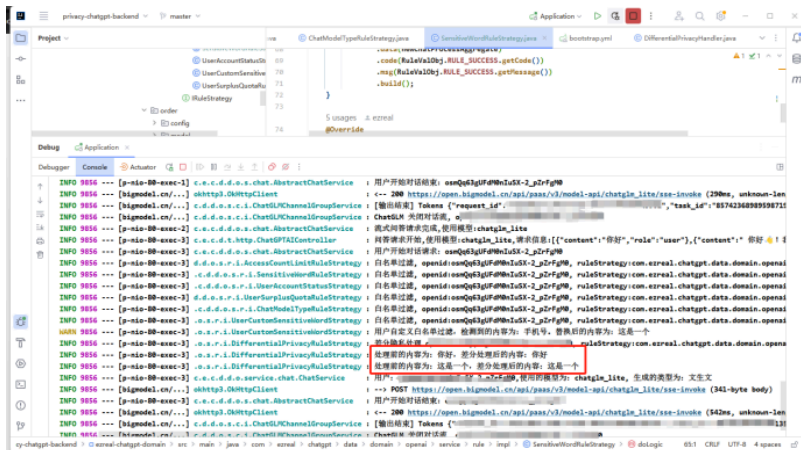
id	openid	sensitive_word	create_time	update_time
1	osmQq63gUfdM0nIuSX-2	手机号	2024-03-30 10:21:50	2024-03-30 12:04:35

- 上图用户配置的敏感词为**手机号**

## 实现效果



- 第三方平台并没有读取到我们输入的内容



- 通过查询日志可以发现，手机号这个用户指定的敏感词被成功处理了，有效保护用户的敏感信息

# 后端功能设计——差分处理隐私数据

对于**整型**或者**浮点型**的数据，我们将通过基于差分隐私的隐私保护技术进行处理

## 差分隐私代码

```
/**
 * sensity: 敏感度 (Sensitivity) , 表示对单个记录的最大影响力。
 * eps: 隐私参数 (Epsilon) , 表示隐私预算, 值越小隐私保护越强, 但会影响数据的精度。
 */
public static double dPLaplace(double sensity, double eps) {
    double beta = sensity / eps; //计算拉普拉斯分布中的尺度参数 $\beta$  (Beta)
    double u1 = Math.random(); // 生成两个介于0和1之间的均匀随机数u1和u2。
    double u2 = Math.random();
    // 根据u1的值选择拉普拉斯分布的负方向或正方向, 计算噪声。
    if (u1 < 0.5) {
        return (-1 * beta * Math.log(1.0 - u2));
    } else {
        return (-1 * beta * Math.log(u2));
    }
}
```

- 实现了一个基于**拉普拉斯机制**的差分隐私 (Differential Privacy) 算法, 用于在处理敏感数据时**添加噪声**以保护隐私。

# 后端系统设计——差分处理隐私数据

向系统输入一组小学生的体重数据计算平均数来检验差分隐私的效果

## 效果展示

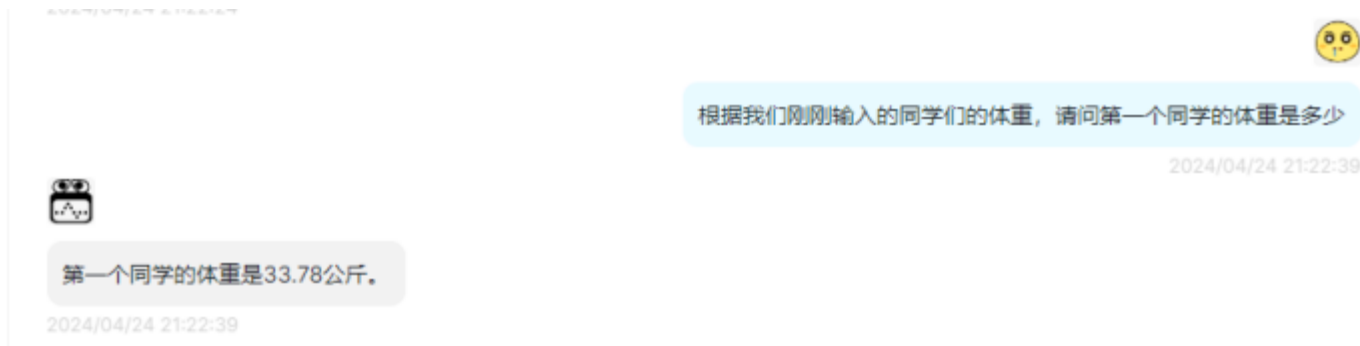


结论: 每一个体重数据都经过不同程度的噪声处理, 最终得到的结果是42.4245kg, 与使用计算机计算出来的结果41.65kg, 差距为41.65, 相差了0.7745kg。有效保护了用户信息, 避免隐私泄露。

# 后端系统设计——差分处理隐私数据

接着，我们向ChatGPT获取这组体重数据的第一个值

## 效果展示



结论：第三方模型平台的回答是33.78，不是33，说明我们的真实的数据33是被后台系统处理了，第三方拿不到我们真实的数据，成功防止成员推断攻击。



# 系统缺陷与不足

## 问题现状

由于我们处理了客户端的传入的数据，会引起下面两个问题：

- 第三方大模型平台返回的内容不符合我们的预期；
- 第三方大模型平台计算的结果不准确；

## 解决办法

1. 选用更加先进的第三方大模型平台，比如说使用OpenAI的ChatGPT大模型代替国内的ChatGLM大模型；
2. 对数据进行合理的处理，对于用户指定的隐私数据或者敏感词，我们可以采用同类词替换来保证；
3. 对计算类的结果，可以根据值的范围返回评价结果；

演示完毕，请评委老师批评指正

导师：唐屹教授

答辩学生：李文杰