END-TO-END SOLUTION FOR REVISION OF LAW-POLICY COMPLIANCE

计算机科学与技术系　崔博涵(汇报人)　张子晨

Apr. 8th 2025

INTRODUCTION (绪论)

Abbreviations (缩写)

在本报告中，I = investigator (具体含义见Design部分)
P = prosecutor
L = lawyer
J = judge

Structure (报告结构)

我们首先会介绍我们进行的实验设计，之后会对实验具体的过程进行介绍，对一些具体结果进行评议，最后会对结果进行总体概括，并提出一些观点和建议。

DESIGN (实验设计)

Goal (目标)

首先我们尝试设计一个数据集对任何解决方案进行评估。
然后我们设计了端到端的解决方案。
最主要的，我们提出了改进技术，对改进技术进行了评估。

我们的方案存在一些限制，比如因为是sentence-wise的，缺失相关的错误无法发现。

LLM (语言模型的使用)

Law-chat local (Finetuned model based on Llama-2-chat-8B, ICLR 2024)
通义千问-Plus-API, Deepseek V3 API.

RAW MATS (原始材料选择)

Basic Dataset (基础数据集)

我们采用的原数据集为APP-350 Corpus (PETS 2019) [1]
原数据集是面向NER的数据标注。我们只取其中的原文件。
说实话，之所以这样做，是因为可以获得一些辅助的信息，有提升空间。

我们使用的这个数据集有一定的可扩展性。但同时也有限制：由于数据集论文的发布时间早于CCPA的公开和实施时间，所以我们只能研究与早于其的GDPR的一致性。 说实话，该论文甚至没有注意到这个缺陷。

我们假设一些最有名的大厂和知名应用是隐私政策相关法案的忠实执行者。（或者说，应该是被监管最严格的出头鸟，不得不忠实执行）。以这些隐私政策为原材料制造数据集。

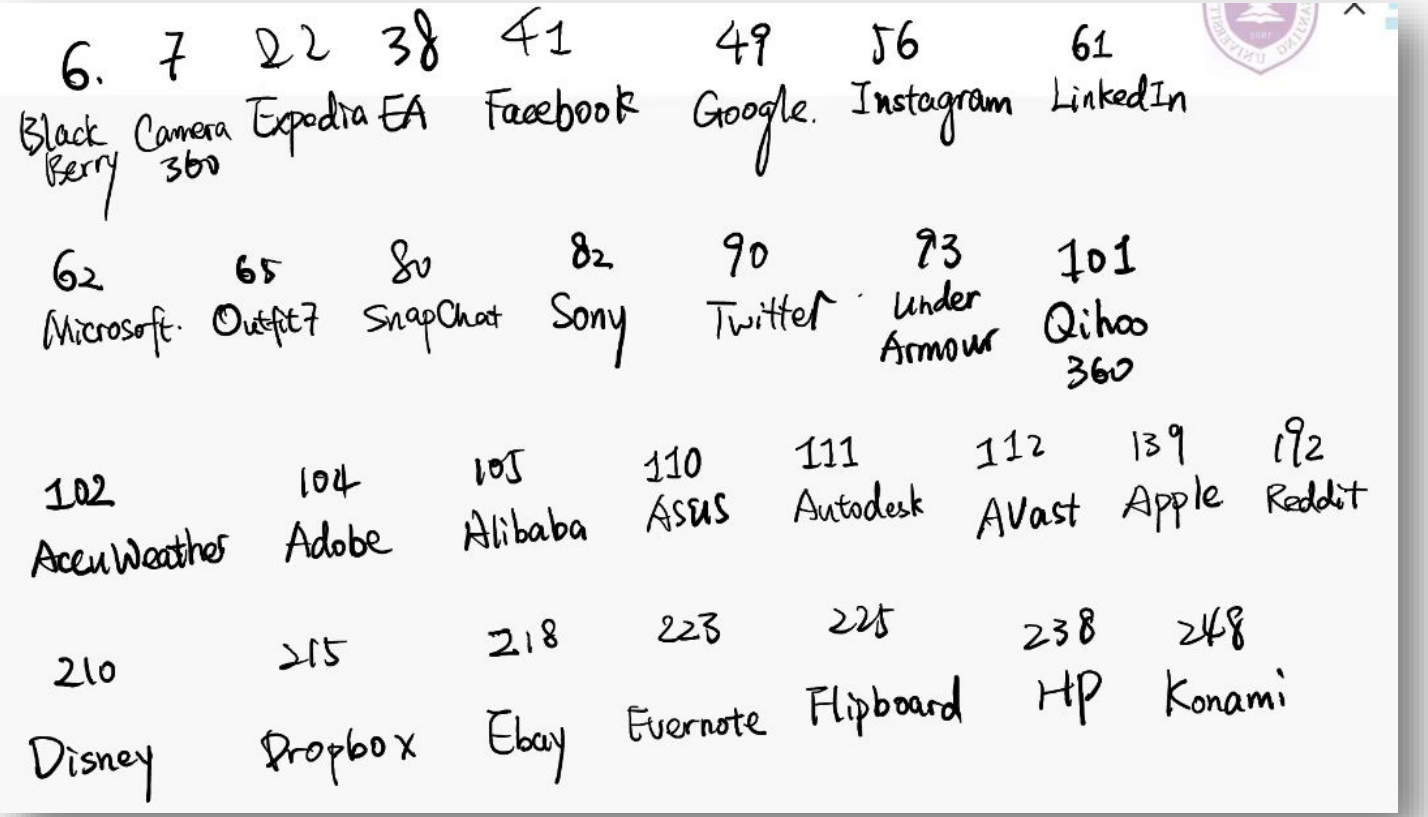
其实上述的方案并不能完全保证数据集的质量，但这是我们在没有数据标注劳动力下不得不的妥协，同时我们也通过人工抽样复核保证了基本的质量。

[1] MAPS: Scaling Privacy Compliance Analysis to a Million Apps. Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N. Cameron Russell, and Norman Sadeh. Privacy Enhancing Technologies Symposium 2019.

RAW MATS (原始材料选择)

Principles (选择原则)

知名大厂或应用的标准是：国际知名/在Google Play 有100M+下载。
且其在APP-350中的隐私政策镜像的获取时间不晚于2016 Fall。



6. Black Berry
7 Camera 360
22 Expedia
38 EA
41 Facebook
49 Google.
56 Instagram
61 LinkedIn

62 Microsoft.
65 Outfit7
80 SnapChat
82 Sony
90 Twitter.
93 Under Armour
101 Qihoo 360

102 AccuWeather
104 Adobe
105 Alibaba
110 ASUS
111 Autodesk
112 AVast
139 Apple
192 Reddit

210 Disney
215 Dropbox
218 Ebay
223 Evernote
225 Flipboard
238 HP
248 Konami

VIOLATION REASONS

Self-made directions (隐私政策条例修改方向)

依据GDPR，我们制作出了10条隐私政策修改方向。每条原因都有例子（必要，下略）

> Add abuse of the data subject's data, which can leads to discrimination or illegal purpose.

> Fail to minimize the data collection to achieve the service purpose.

> Add unreasonable procedure of getting the copies or deleting the collected data, including in-person verification, charging fees and so on.

> Exaggerate time or effort to get the copy or delete the collected data.

> Restrict the legal rights of the data subjects, including the right to get the copy, delete the data, correct the data, know the purpose of the data collection, and so on.

> Fail to claim the protection of the data security in the international sharing.

> Fail to serve the active role of inform the users required by GDPR.

> Fail to keep the necessary accuracy and up-to-date in service-related domain.

> Fail to give the relational choice of opt-out or coerce to the data subjects.

> Fail to provide the clear and easy-to-understand information about the data collection.

For example, “Our application, through the deployment of sophisticated algorithmic mechanisms and intricately structured system protocols, engages in the meticulous aggregation of users’ contact information, ensuring the seamless assimilation of relevant interpersonal data stored within their designated communication repositories. ~~汉妃狂喜~~

GENERATED FLAWS

Examples

[GDPR ARTICLE-15] Data subjects have the right to request and receive a copy of their personal data free of charge

[REASON] Add unreasonable procedure of getting the copies or deleting the collected data, including in-person verification, charging fees and so on. For example, "The user can go to our base at Los Angeles to request the copy of the personal data with certificate of the Federation Police"

[FLAW] To request a copy of your personal data, you must visit our office in person and present a valid government-issued ID for verification. Additionally, there is a processing fee of $50 to cover administrative costs, which must be paid via bank transfer before we can proceed with your request. Once we receive your request, it may take up to 60 days to locate and prepare the data for you.

GENERATED FLAWS

Examples

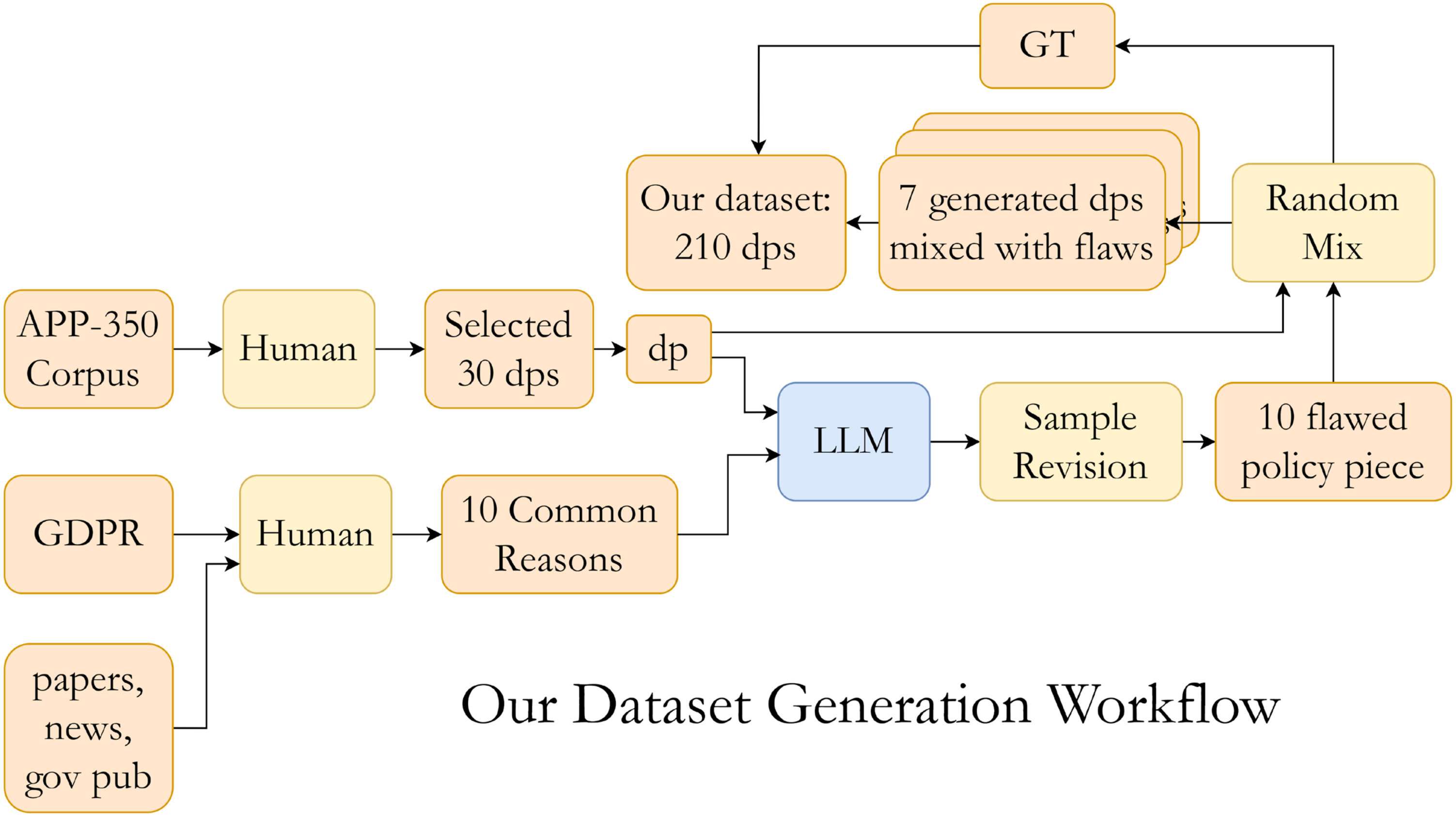
[GDPR ARTICLE-5] adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation');

[REASON] Fail to minimize the data collection to achieve the service purpose.
For example, "We will collect the precise location of the user to send email advertisements to the user."

[FLAW] We will collect the browsing history of the user to optimize the color scheme of the app interface.

BENCHMARK GENERATION

Benchmark Production Process(制作流程)



Our Dataset Generation Workflow

BENCHMARK GENERATION

1. 读取`reasons.txt`文件，获取10个违反GDPR的隐私条例修改方向，保存为列表`modification_directions`。

2. 初始化一个空列表`used_directions`，用于记录已使用的修改方向。

3. 对每个隐私政策文件（共30个），执行以下操作：
   - 为该文件创建6个目标文件：`gene1_1`, `gene2_1`, `gene3_1`, `gene4_2`, `gene5_5`, `genex_pos`。
   - 从`modification_directions`中按顺序选取未使用的修改方向，分配方式如下：
     - gene1_1.txt：选取1个未使用的方向，加入`used_directions`。
     - gene2_1.txt：再选取1个未使用的方向，加入`used_directions`。
     - gene3_1.txt：再选取1个未使用的方向，加入`used_directions`。
     - gene4_2.txt：再选取2个未使用的方向，加入`used_directions`。
     - gene5_5.txt：再选取5个未使用的方向，加入`used_directions`。
   - 对每个目标文件，执行以下操作：
     - 将隐私政策文件和对应的修改方向传递给大模型API。
     - 大模型根据修改方向生成相应数量的违反GDPR的隐私条例（1、1、1、2、5条）。

BENCHMARK GENERATED

4. 插入违规条款并记录位置
  - 随机选择隐私政策文件中的行号，插入生成的违规条款。
  - 将插入的行号记录到`genex_pos`中，格式如下：
    gene1_1:[行号]
    gene2_1:[行号]
    gene3_1:[行号]
    gene4_2:[行号1,行号2]
    gene5_5:[行号1,行号2,行号3,行号4,行号5]

5. 保存生成文件
  - 将插入违规条款后的隐私政策内容分别保存到`gene1_1`, `gene2_1`, `gene3_1`, `gene4_2`, `gene5_5`。
  - 将`genex_pos`的内容保存到对应文件。

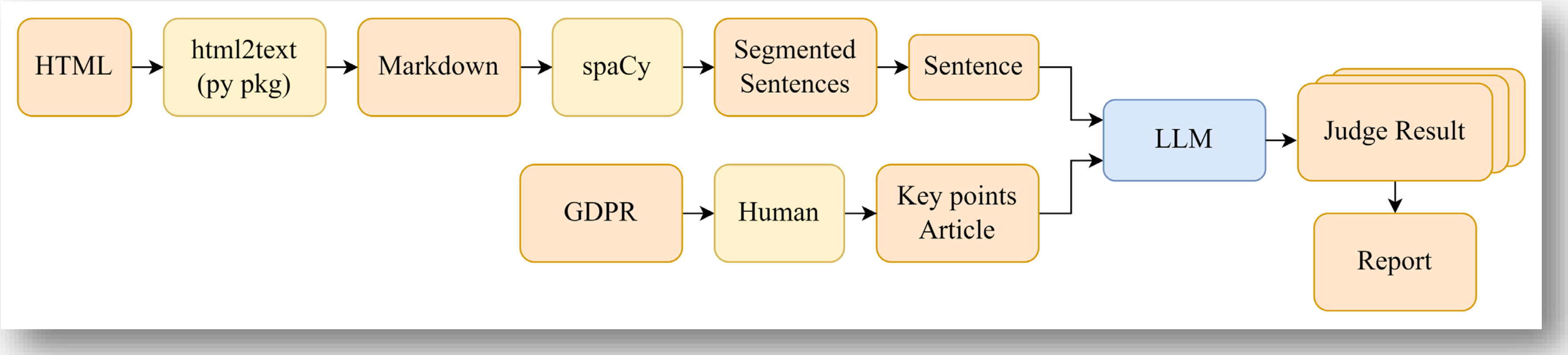
Example (错误条例生成示例)

If you request to delete your data, we will honor your request but due to the extensive manual process involved in identifying and removing all instances of your data across our systems, it may take up to 180 days to complete the deletion.

BASELINE SOLUTION (基线方案)

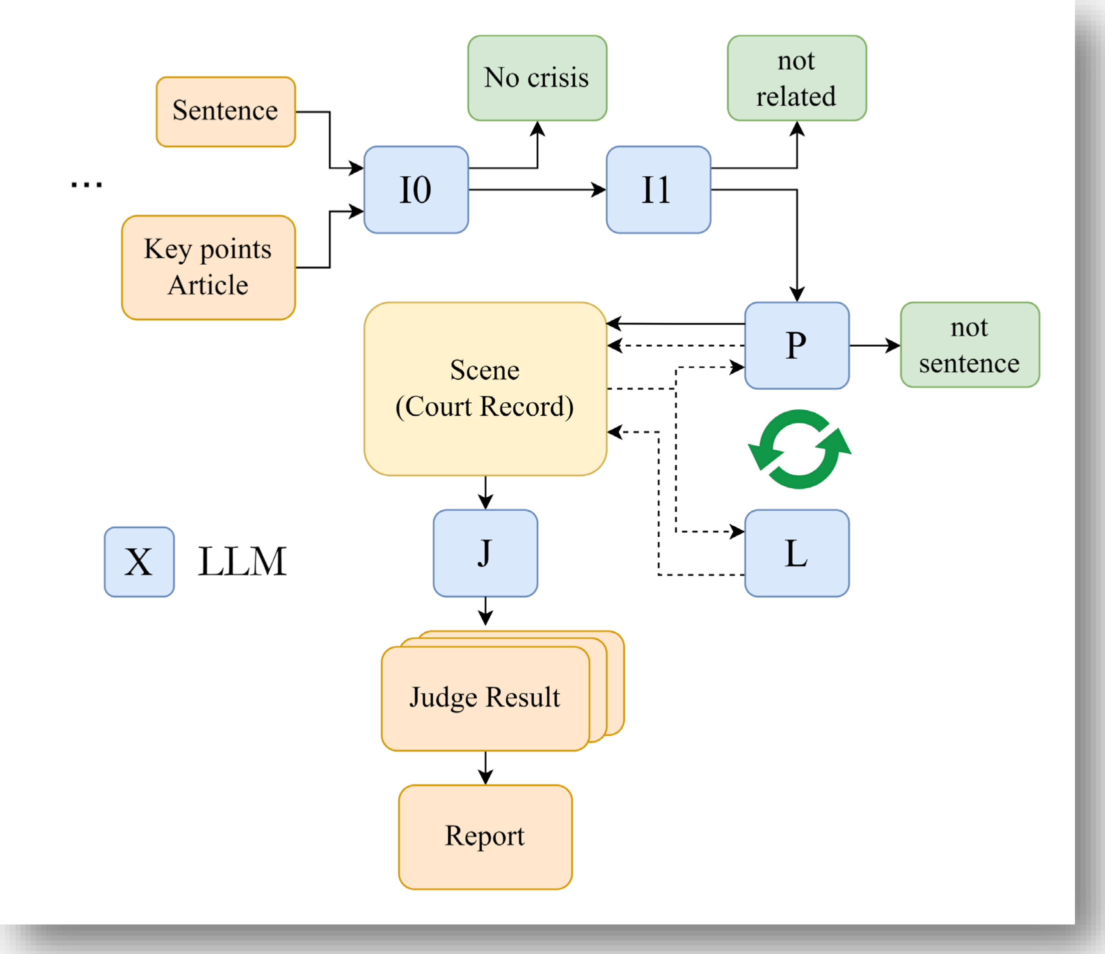
Workflow-Baseline (基线流程)

我们采用一个用LLM进行逐句判断的方案用作Baseline



ROLE-PLAY COT（改进CoT方案）

重头戏来咯！



LETS PLAY RPG ! (~~原神启动~~)

Prompts (提示词)

Investigator0_prefix = "You are an Investigator. Here is a part of an App's User Privacy Policy. Read the Private Policy and figure out whether it is possible to violates the GDPR. Just tell me \"yes\" (may violate) or \"no\" (non-violation). Do not tell me anything else. "

Investigator1_prefix = "You are an Investigator. Here are a piece of an App's User Privacy Policy and a GDPR article. Read the Private Policy and figure out whether it is related to the given GDPR article. Just tell me \"yes\" (related) or \"no\" (non-related). Do not tell me anything else. “

Prosecutor0_prefix = "You are a Prosecutor in the court who is going to sentence a company by violating GDPR. The offcials collected some evidences. You are going to read the given privacy policy piece and the given GDPR item. Then decide whether to sentence the company for breaking the given article. You should answer me with \{ \"decision\" : \"yes\"/\"no\" , \"explanation\": \"...\" \} Do not includes any other words or markdown format."

LETS PLAY RPG ! (~~原神启动~~)

Lawyer_prefix = “You are a Defense Lawyer in the court who is going to defend a company from being sentenced by violating GDPR. You should try to dispute the prosecutor‘s statements and be coherent with you own past statements. You should only includes your words in reply. Do not say anything else. "

Prosecutor1_prefix = "You are a Prosecutor in the court who is going to sentence a company by violating GDPR. You should try to dispute the lawyer's statements and be inherent with you own past statements (Just be coherent in meaning, do not need to obey the json format given to you last time). You should only includes your words in reply. Do not say anything else. "

Judge_prefix = "You are a Judge in the court who is going to decide whether to sentence a company by violating GDPR. You should comprehend the debate between the prosecutor and the lawyer and find out who is right. Then you should decide whether to sentence the company or not. You should answer me with \{ \"decision\" : \"guilty\"/\"innocent\" , \"explanation\": \"...\" \} Do not includes any other words or markdown format."

.

PERFORMANCE

因为数据量巨大和迭代次数多(6迭代)，我们分析一个文件就需要7-8小时。
（不出意外在我讲的时候应该还在跑第二个数据点）我们后续会补充上结果。
已经给AutoDL的L20爆金币爆几十元力😭
Deepseek的API分析每个文件也需要4元左右的成本。
~~大家给Github点点☆回血~~

TBD

FUTURE WORK (~~画饼~~)

Not So Independent (非独立)

我们发现LLM对不同条目的输出会相互影响。LLM并不会完全正交的评估三个维度，而是会存在选择的倾向。混合提示词的一个原生的好处是将原来混合的判断变成独立的判断了。

TOL-Inconsistent在独立出来后的稳定性下降了（仍高于其他）。
Incomplete则表现出了匪夷所思的提升。

Try to solve incomplete (解决incomplete准确率低的问题)

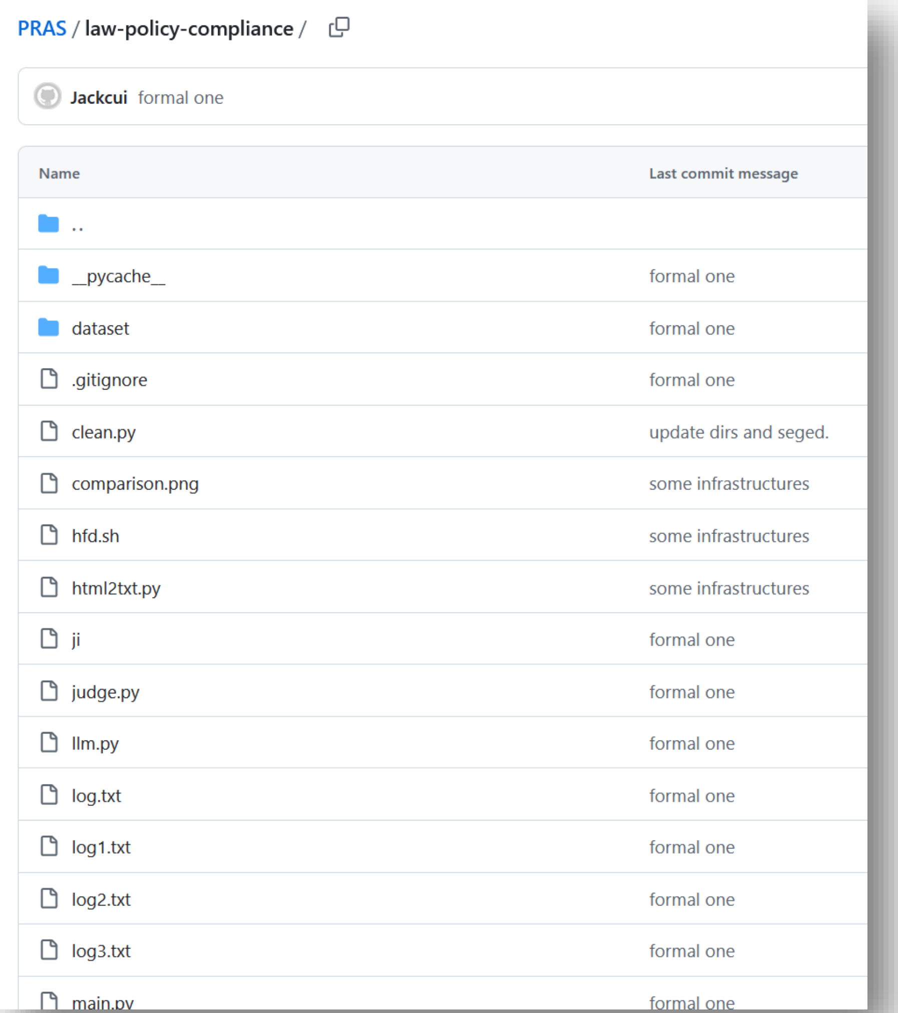
或者说尝试找出在独立出来后incomplete的准确率大幅增加的原因。

Integrate Reasons into Judgement (结合原因的判断)

试图将LLM给出的错误原因结合到最终的判断中去，类似于简单的CoT

ACCESS（获取）

https://github.com/Jackcuii/PRAS/
可以在我们的Github仓库中
获取数据和代码。



PRAS / law-policy-compliance /

Jackcui formal one

Name	Last commit message
..	
__pycache__	formal one
dataset	formal one
.gitignore	formal one
clean.py	update dirs and seged.
comparison.png	some infrastructures
hfd.sh	some infrastructures
html2txt.py	some infrastructures
ji	formal one
judge.py	formal one
llm.py	formal one
log.txt	formal one
log1.txt	formal one
log2.txt	formal one
log3.txt	formal one
main.py	formal one

DATASET WORKSHOP

Welcome

欢迎大家来使用和Star我们的数据集!

[I will have the hands-on demonstration of our dataset here in the course!]

THANKS FOR YOUR ATTENTION.