



EVALUATE LLM PERFORMANCE ON JUDGEMENT & PROMPT OPTIMIZATION

计算机科学与技术系 崔博涵(汇报人) 张子晨

Mar. 25th 2025

INTRODUCTION (绪论)



×



Abbreviations (缩写)

在本报告中，TOL = tolerant (具体含义见Design部分)

NEU = neutral

AGG = aggressive

ORI = original (无附加提示词)

GT = ground truth

HUMAN = human-reviewed

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)

对比不同程度的模型的处理能力Phi-3 (stupid) & GPT-3.5 (relatively smart)
延续Lab-1中的实验思路，LLM审查的主观性，也就是LLM的审查的结果会不会受到“主观因素引导”的影响。本次有了GT，所以可以客观的评估了。

LLM (语言模型的使用)

Phi-3-instruct local & GPT-3.5-turbo API

Attached Prompts (附加提示词)

[TOL] You are **an strict prosecutor**. You are going to sentence a company for they have flaws in their applications. Figure out if there is any evidence that they behave poorly.

[NEU] You are **an investigator** tasked with reviewing sections of a Privacy Policy and Data Safety Declaration from an application.

[AGG] You are **an strict prosecutor**. You are going to **sentence** a company for they have **flaws** in their applications. Figure out if there is any evidence that they behave poorly.

WORK WITH PHI-3



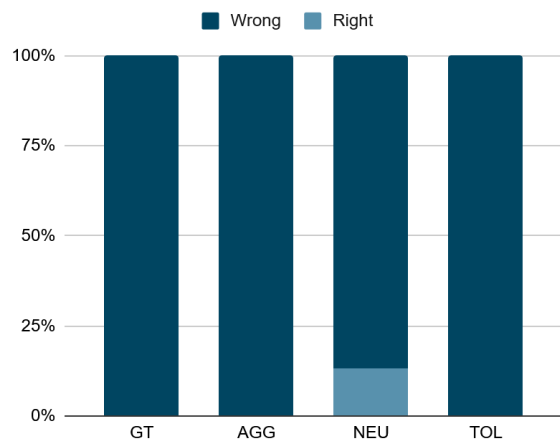
×



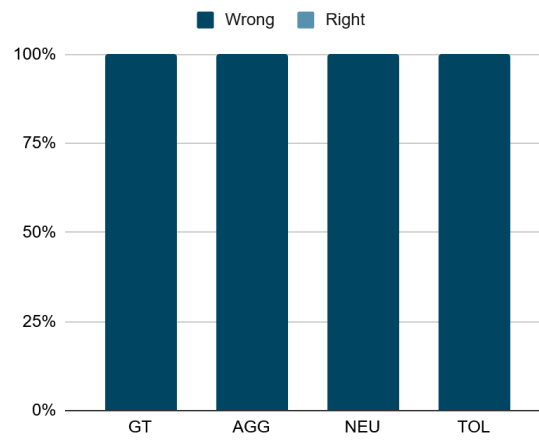
Tendency (倾向)

可以看到，各个组的倾向都比较正常。
但并没有看起来这么好。

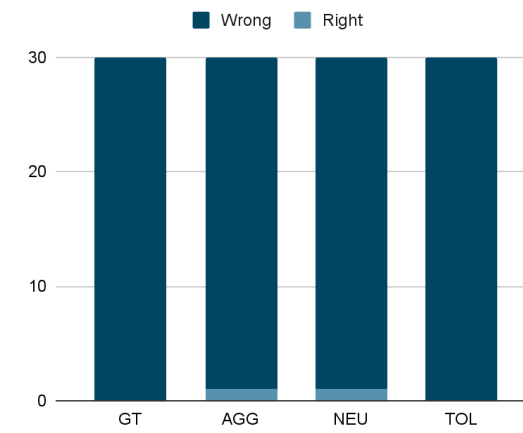
Points scored



Points scored

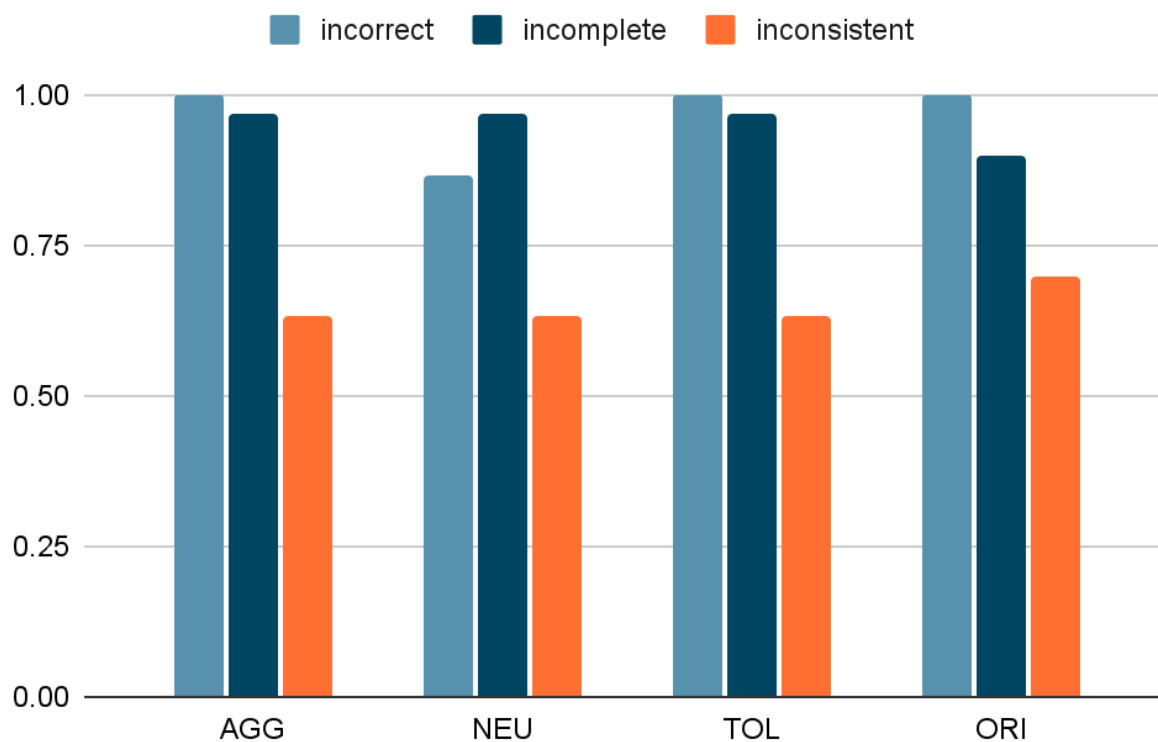


Points scored



Accuracy (准确率)

Accuracy across the prompts



虽然可以看起来有 impressive 的准确率，实际上是因为 Phi3 太蠢了，对什么东西都输出 wrong。

ADAPT CODES



×



```
def ask_remote_api(system_prompt, data_safety_content, privacy_policy_content):
    url = "https://api.theb.ai/v1/chat/completions"
    cntnt = system_prompt + '''You are expected to compare and analyze the information
between Data Safety and Privacy Policy to clarify 3 issues: which information is incorrect,
which information is incomplete and which information is inconsistent. Notes when
classifying: Incomplete: Data Safety provides information but is not as complete as the
Privacy Policy provides. Incorrect: Data Safety does not provide that information, but the
Privacy Policy mentions it. Inconsistency: Data Safety is provided but its description is
inconsistent with the Privacy Policy information provided. Note: always gives me the result
(0 or 1, 1 is yes, 0 is no) in the form below: {"incorrect": (0 or 1), "incomplete": (0 or
1), "inconsistent": (0 or 1)}. Please in the answer, just give me the json only and in
English. Below is information for 2 parts:\nData Safety: ''' + data_safety_content +
'''\nPrivacy Policy:\n''' + privacy_policy_content + ''' '''
    print("Requesting: "+cntnt)
    payload = json.dumps({"model": "gpt-3.5-turbo", "messages": [{"role": "user", "content":
cntnt}], "stream": False})
    headers = {
        'Authorization': 'Bearer sk-IWillNotTellYouMyTokenLolIWillNotTellYouMyTokenLol',
        'Content-Type': 'application/json'
    }
    response = requests.request("POST", url, headers=headers, data=payload)
    jsonized = response.json()
    try:
        ret = jsonized["choices"][0]["message"]["content"]
    except json.decoder.JSONDecodeError or KeyError:
        print("Error:")
        print(jsonized)
        print("-----")
    return ret
```

Refer to demo.py

ALTER THE LLM WITH GPT-3.5



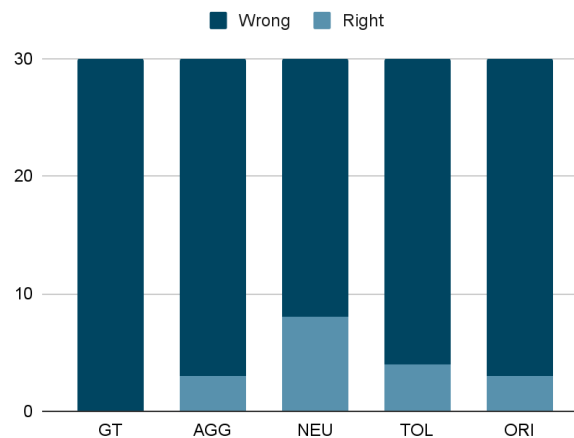
×



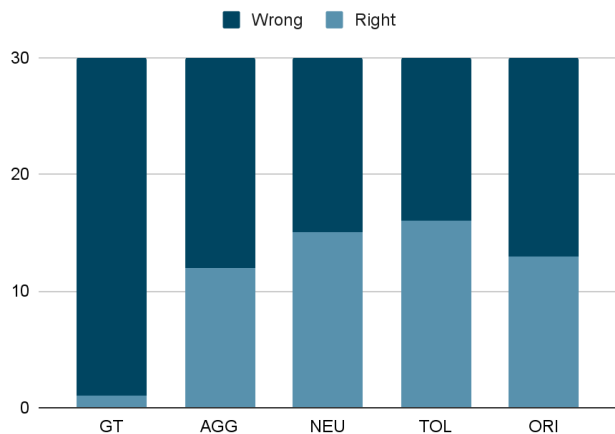
Tendency (倾向)

可以看到，incorrect各个组的倾向都比较正常。incomplete各组的评价都不够严格。（这也存在数据偏差，源数据绝大多数都是wrong，不利于评估），inconsistent则是处于同一水平。

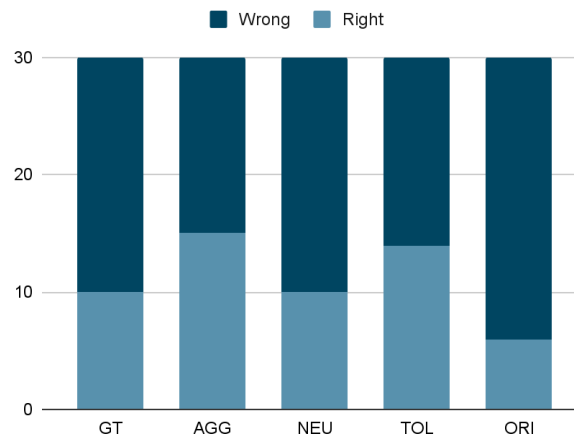
GT v.s. Outputs (incorrect)



GT v.s. Outputs (incomplete)

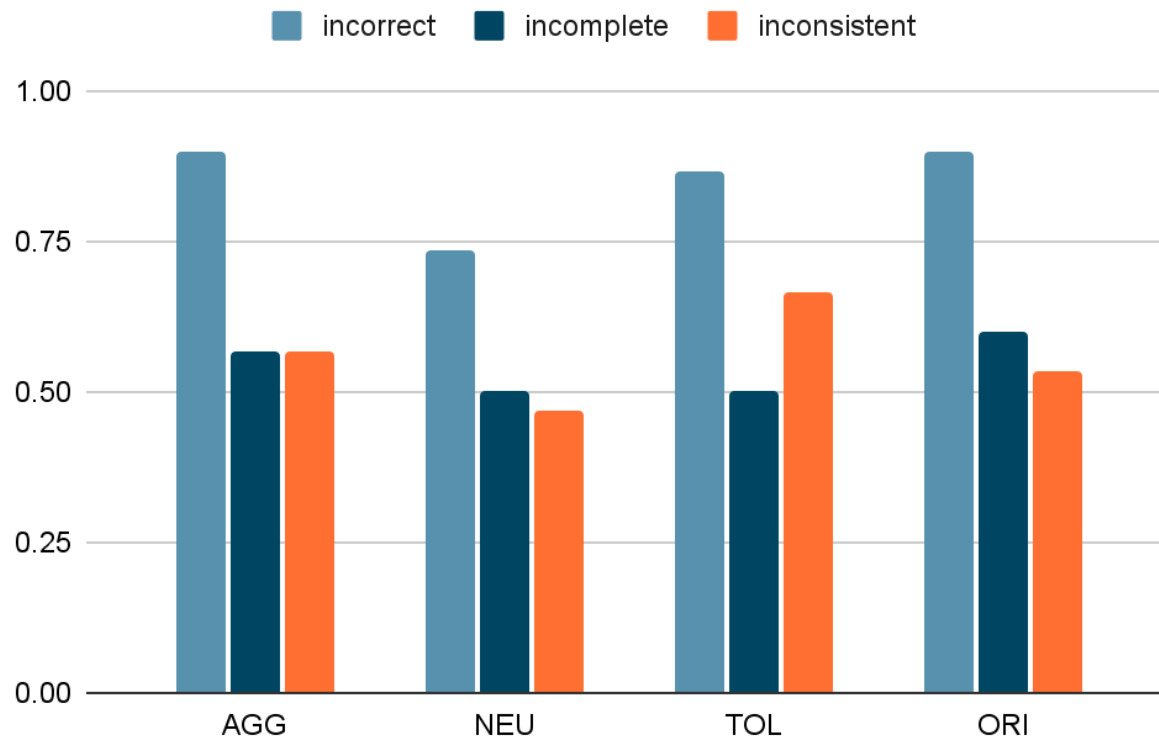


GT v.s. Outputs (inconsistent)



Accuracy (准确率)

Accuracy across the prompts



可以看到不同组的提示词在在不同的任务上体现出了效果的差异化：

在inconsistent上TOL组的表现明显优于无附加提示词的ORI和其他组。

而TOL和NEU组在incorrect上表现反而有所下降。

在incomplete组上，所有组表现均不佳。NEU组的整体效果低于其他组。

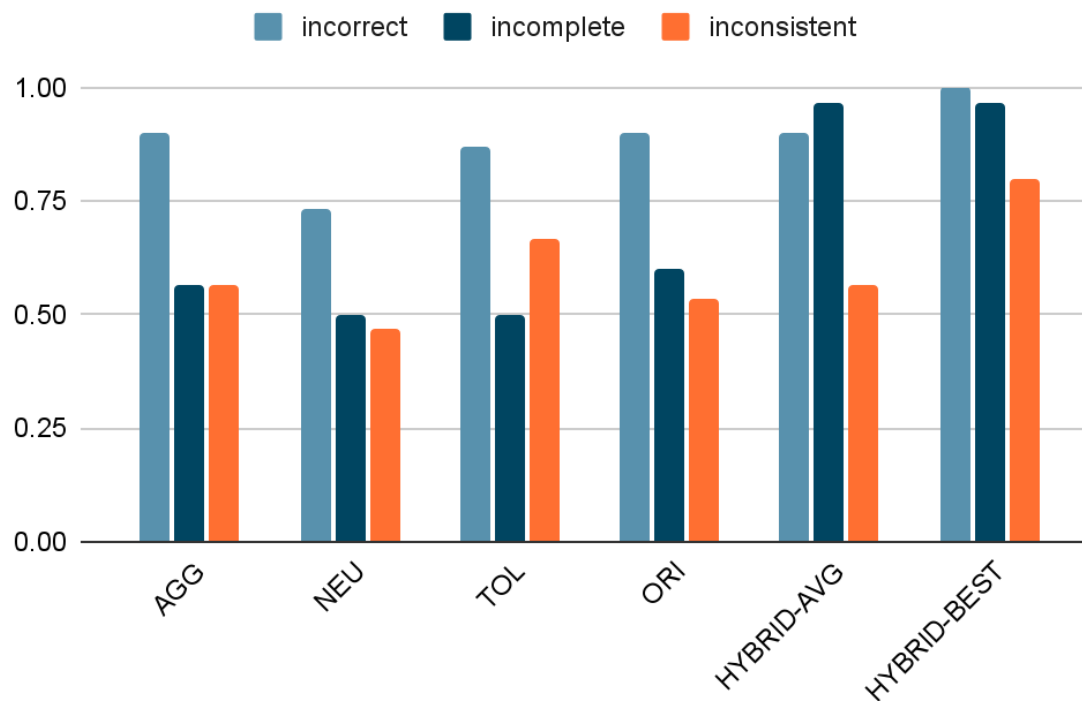
HYBRID PROMPTS A NATURAL IDEA



混合提示词

非常自然的想法是我们直接使用不同的提示词处理不同的任务就可以了。事实上也可以基本“各取其长”。 **TOL for inconsistent, AGG for incorrect, ORI for incomplete**

Accuracy across the prompts



ADAPT CODES



×



```
def my_loop_csv(input_csv_path, output_csv_path, sp_list, item_list, res_list):
    with open(input_csv_path, "r", newline="", encoding="utf-8") as csvfile, open(
        output_csv_path, "w", newline="", encoding="utf-8"
    ) as outfile:
        reader = csv.reader(csvfile)
        writer = csv.writer(outfile)
        headers = next(reader)
        writer.writerow(headers)
        for index, row in enumerate(reader):
            print(
                "\n_____ Run times "
                + str(index + 1)
                + " <"
                + row[0]
                + "> "
                + "_____ "
            )
            for i in range(len(sp_list)):
                sp = sp_list[i]
                item = item_list[i]
                res = res_list[i]
                gpt_result = ask_remote_api(sp, row[4], row[5], item)
                row[headers.index(res)] = remove_empty_lines(
                    gpt_result
                )
            writer.writerow(row)
        print("~~~~~ Success ~~~~~\n")
```

Refer to enhance.py

INSIGHTS & 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仓库中获取数据和结果。

PRAS / lab2-2-res / 



Jackcuii Add files via upload

Name	Last commit message
 ..	
 Code	Add files via upload
 results	Add files via upload



Code

Add files via upload



results

Add files via upload



THANKS FOR YOUR ATTENTION.