



# PRIVACY POLICY RELATIONSHIP DIVISION WITH LLM & CLUSTER

计算机科学与技术系 崔博涵(汇报人) 张子晨

Apr. 1<sup>st</sup> 2025















WE DID NOTHING  
THANKS FOR YOUR ATTENTION.





×



# HAPPY APRIL FOOLS' DAY !

四月は君の嘘



# INTRODUCTION (绪论)



×



## Abbreviations (缩写)

在本报告中, AP = Amazon photos

Bili = Bilibili

CMOS = CityMallOnlineShopping

x-Instruct = 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)

我们尝试使用大语言模型进行分类(Qwen-Plus), 同时用人工审查得出了一组GT。另外, 我们分析用KMeans进行聚类分析的效果, 并且尝试使用一些方法对其进行改进。

## LLM (语言模型的使用)

Qwen Plus & GPT-4o API & GPT-3.5-turbo API & Llama 3.1-405B

# FAILED TRY (失败的LLM预处理)



我们发现基于句号的分割预处理效果很差。我们尝试用LLM对最开始的句子处理进行预处理。我们尝试让LLM删去一些非完整的句子。

## Prompts (提示词)

Try to **divide (or combine) the whole text into lines**. Each line should only be one complete sentence. i.e. If there is any line is only one **isolate word of phrase** (for example, the titles of the original text), you can just delete them. only give me the result, do not add any extra information. mark the end of each line with tag string `'</next>'` Remember you must try your best to include all the sentences in the original text. `\n` Target text:

**Break some of the lines** in the original text. Make sure each line in your output should only be one complete sentence. The content of the output lines should be exactly the same as the original text except for the breaks. Only give me the result, do not add any extra information. mark the end of each line with tag string `'</next>'` Remember you must try your best to include all the sentences in the original text. `\n` Target text:

效果令人失望，对于GPT-4o（CloseAI 又偷偷降智了？）。无论是否使用删除部分的提示词，再三强调，都只会有选择的生成一小段甚至是直接进行总结。

对于Llama-405B是可以正常输出的。但输出太长会被中途截断。

另外就是太贵了！上述两个跑两次就要 \$0.8 !!!!!



## Final Solution (最终方案)

我们最后采用将换行和 ‘.’ 共同作为切割的标识符。

同时我们为了防止其中的 i.e. 和 e.g. 等被切割，在切割前进行替换，再进行恢复。可以获得相对良好的效果。

在切割完之后，每一个单元再用 ‘ ’ 切割，如果单词数少于6个就直接扔掉。

# MANUAL CALIBRATION ONE



×



Private Policy :

AmazonPhotos\_processed\_original

Total Sentences : 203

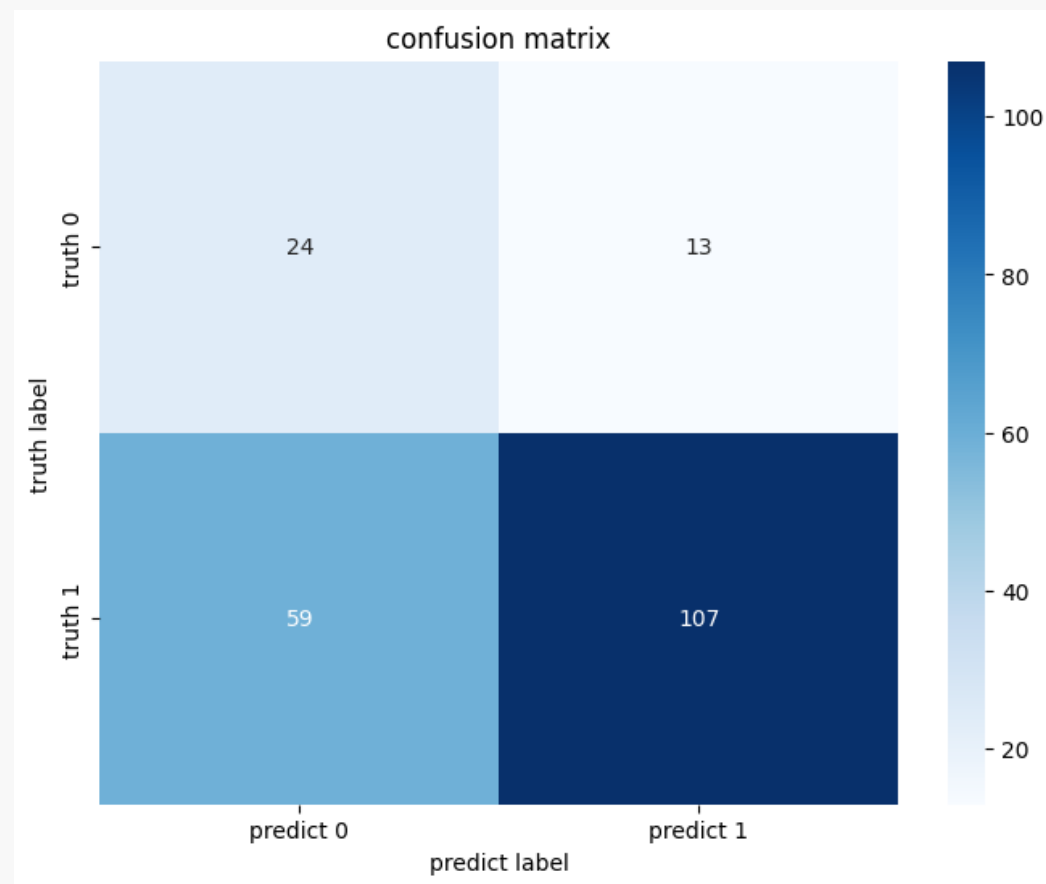
Accuracy : 0.65

Example :

[TEXT] You can choose not to provide certain information, but then you might not be able to take advantage of many of our Amazon Services.

[LLM] True

[MANUAL] False



模型倾向于将“涉及到个人信息”的条款看作与“个人信息利用”相关，实际上该条款只与“个人信息”相关，与“利用”无关，LLM在该隐私政策文件上的大多数错误都如此。

# MANUAL CALIBRATION TWO



Private Policy :

BilibiliHDAimeVideos\_processed\_original

Total Sentences : 265

Accuracy : 0.97

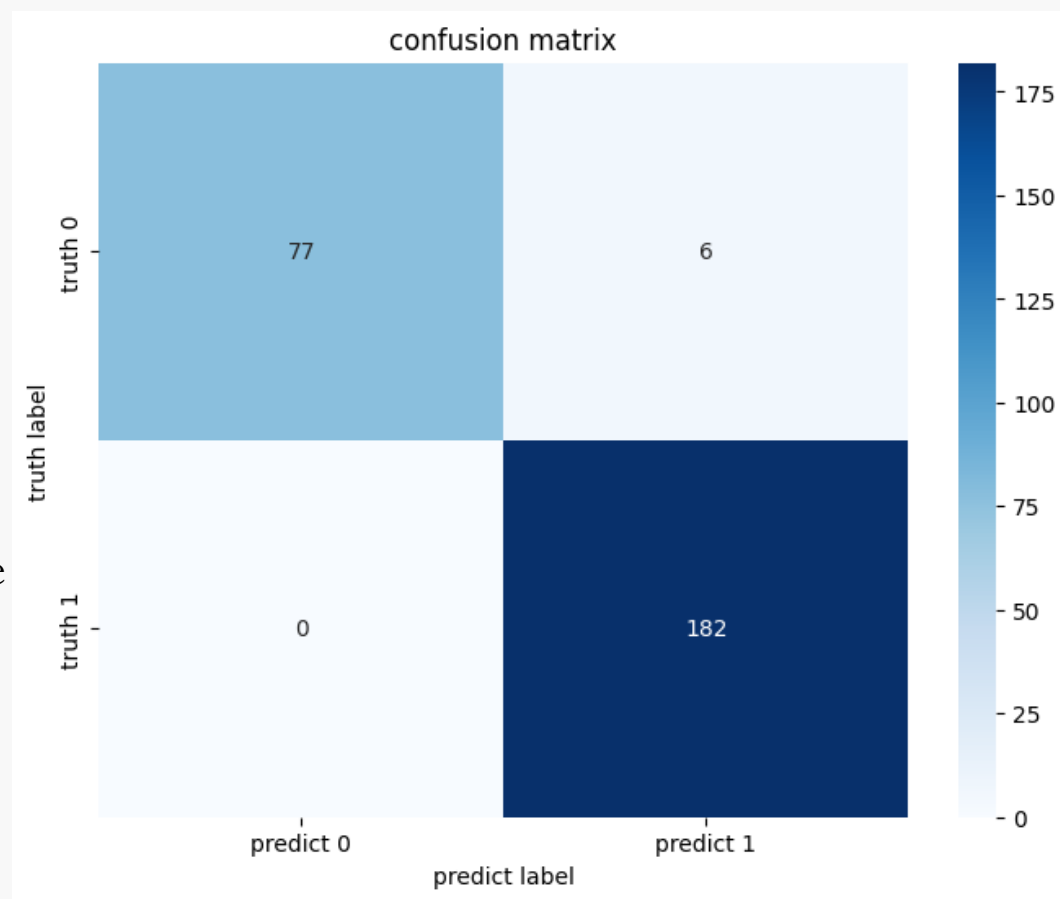
Example :

[TEXT] If you choose to register to use the Platform using your social network account details, you will provide us or allow your social network to provide us with your username and public profile.

[LLM] False

[MANUAL] True

未直接提及“个人信息”该名词，但username和public profile与个人信息明显相关，LLM忽略了条款中对个人信息的隐式表达方式。



# MANUAL CALIBRATION THREE



×



Private Policy :

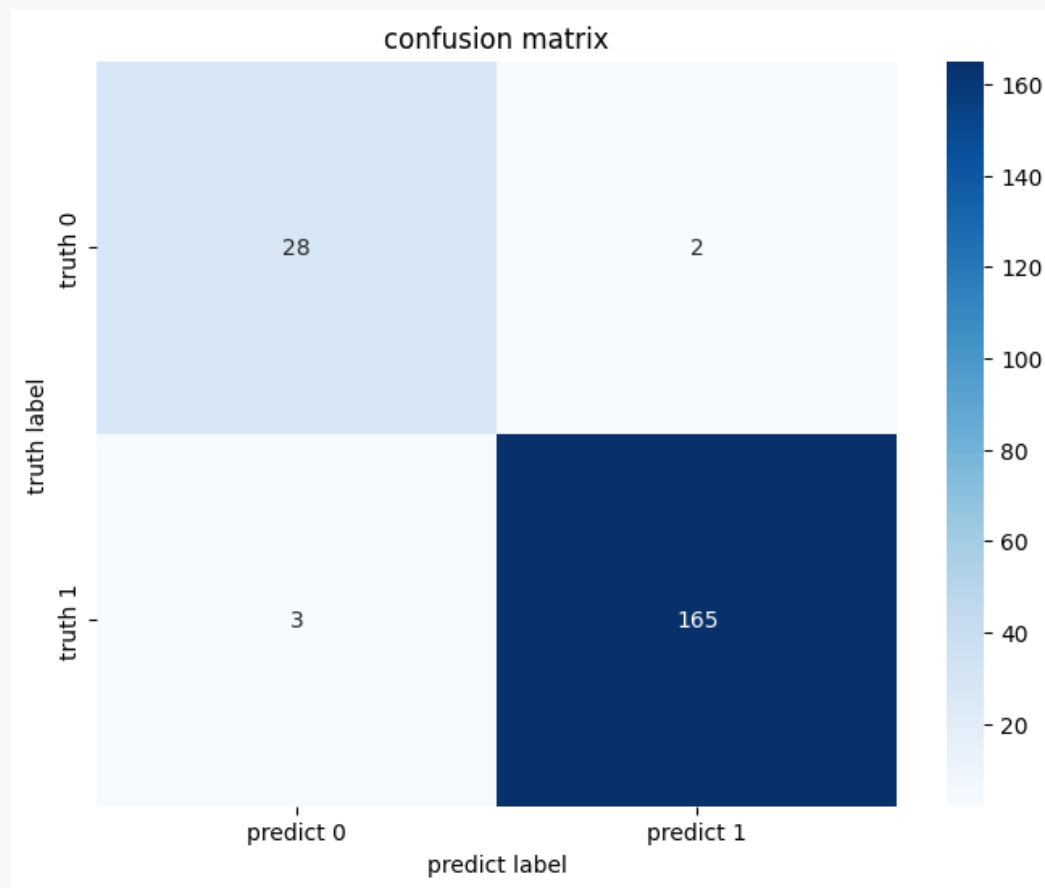
CityMallOnlineShoppingApp\_processed\_original

Total Sentences : 198

Accuracy : 0.97

Example :

[TEXT] This means that We can contact customers to draw attention to Our services.





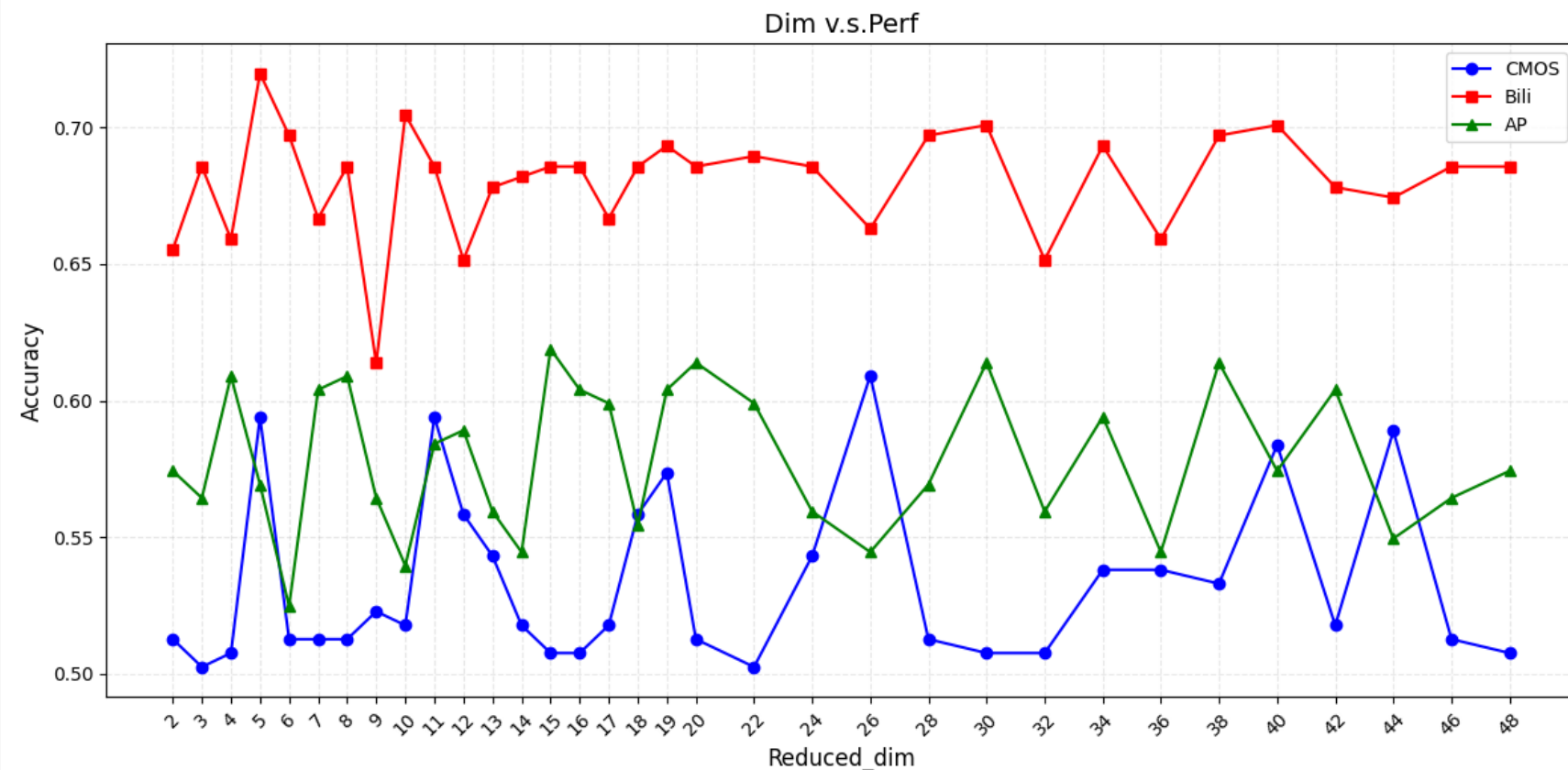
# THE HYPER-PAR DIM



×



可以看到，几乎没有什么规律，我们就暂且将其认为是一个无关量。  
而且不同的文本之间有巨大的差距。Bilibili可能是同一句式较多。



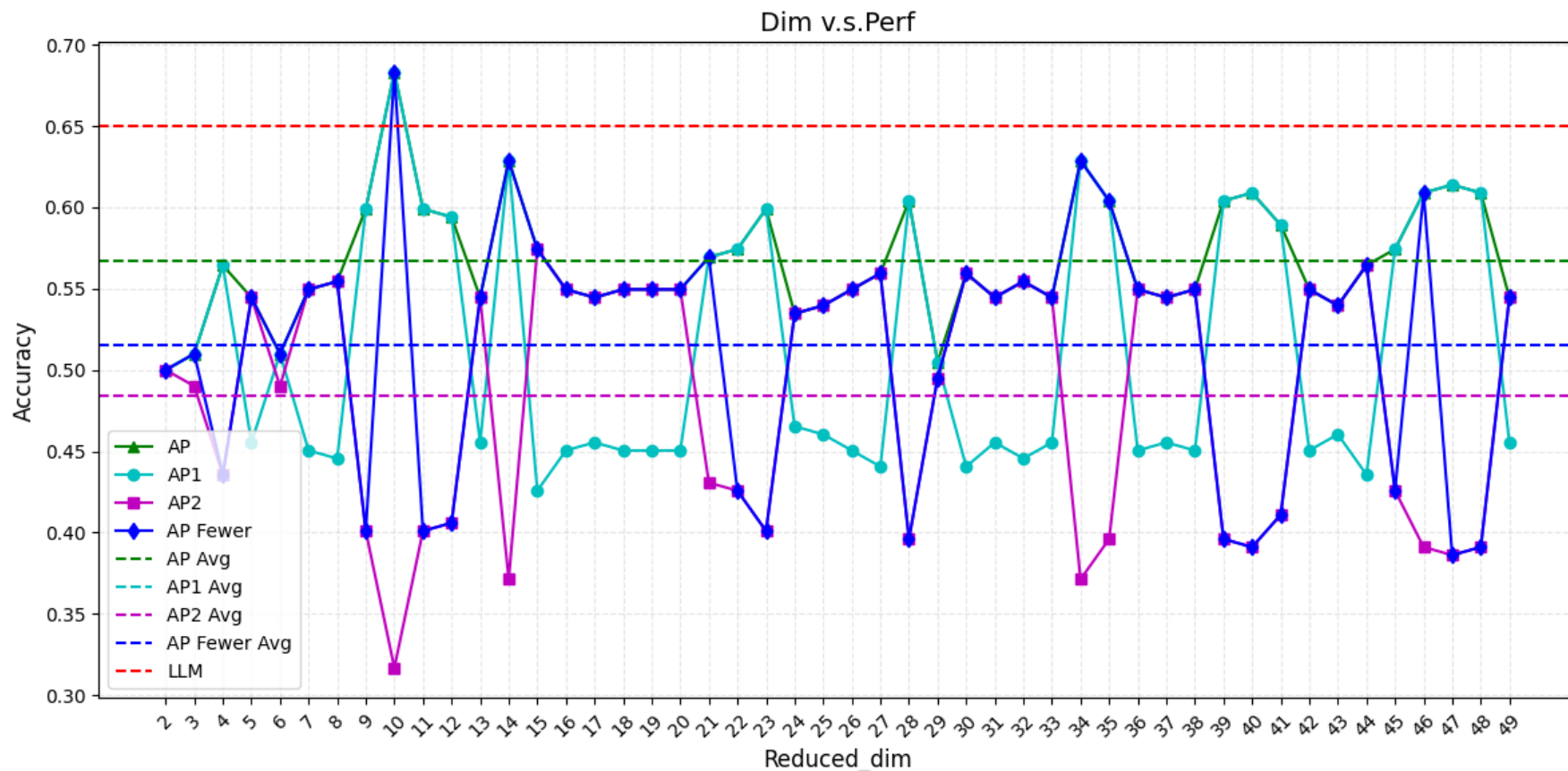
# THE PERFORMANCE OF KMEANS



×



可以看到，可以看出效果比较拉跨。  
而且最重要的是“怎么知道要哪一类？”，这里面展示了几种策略。



# IMPROVE THE PERFORMANCE



×



根据老师上一次的提示:

我们统计了一下其中被LLM分割出的数据中yes的数据的占比。  
这三组中 从 15%到40%不等。

一个最直接的办法就是就是在里面添加一些已知的组的数据。  
我们将Review过的CMOS组数据30组放入AP组。

一个附加的好处就是我们可以通过这些额外组数据的分布判断分组。

```
Unique GT(LLM) Entries: 192
Cluster Entries: 202
Combined and ordered. Total Entry: 202
There is 38 zeros and 164 ones in LLM.
There is 118 zeros and 84 ones in Cluster results.
True Positive: 21
False Positive: 17
False Negative: 97
True Negative: 67
Accuracy: 0.43564356435643564
True Positive: 17
False Positive: 21
False Negative: 67
True Negative: 97
Accuracy: 0.5643564356435643
```

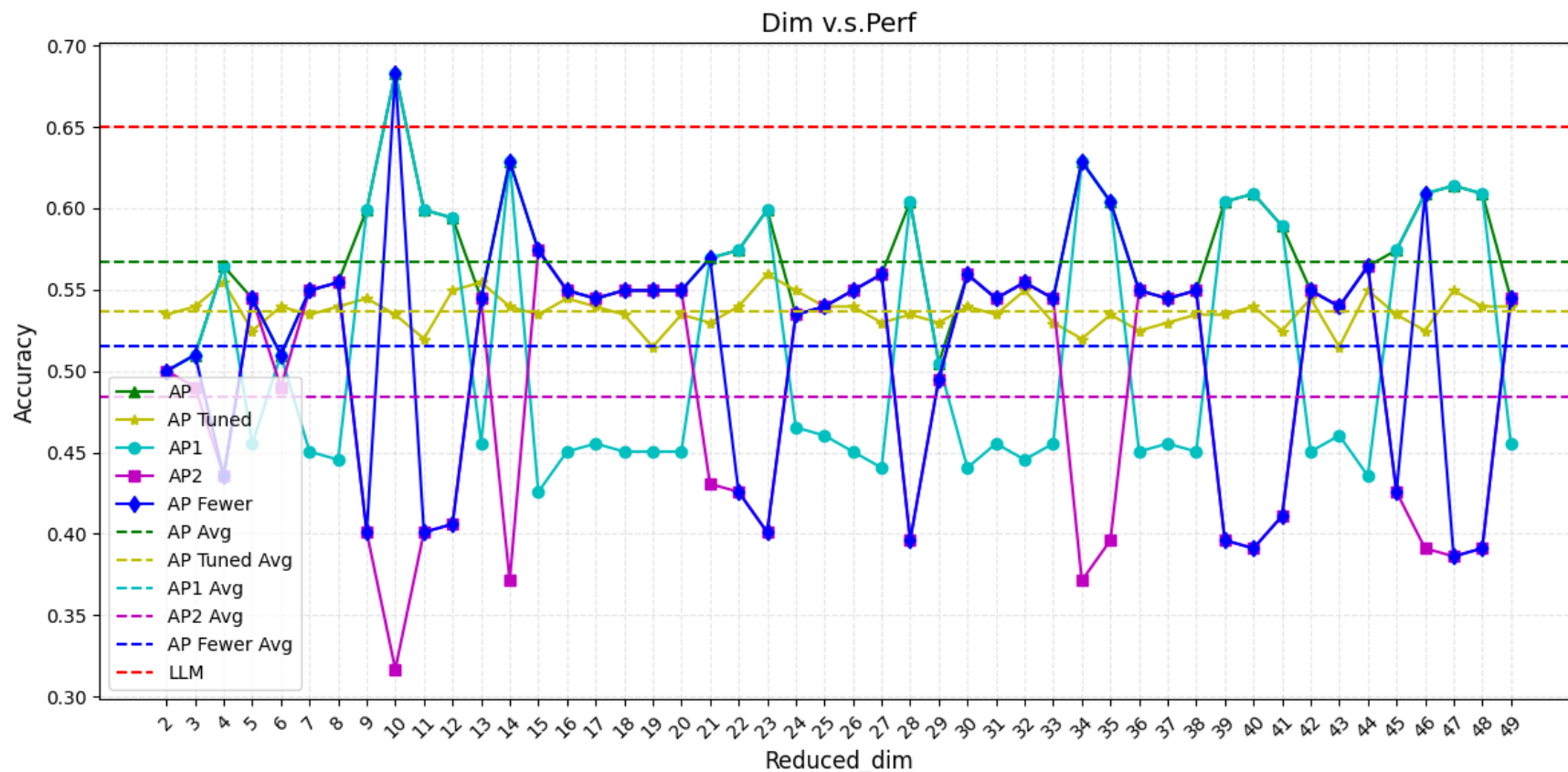
# IMPROVE THE PERFORMANCE



×



可以看到准确率有两个百分点的提升，最重要的是，其稳定程度明显高于其他比较naïve的策略。还算比较喜人(?)



# IMPROVE THE PERFORMANCE

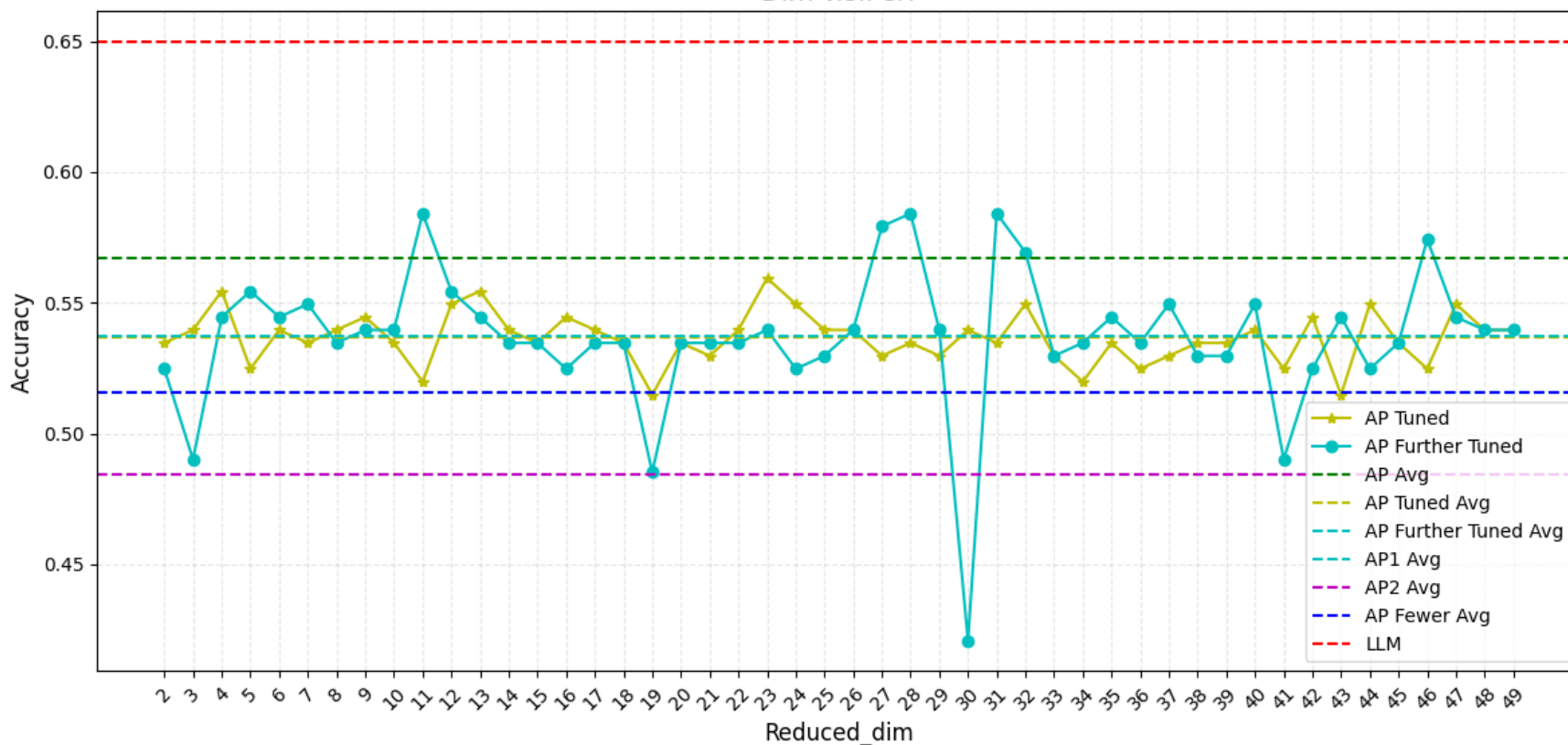


×



但是这种策略边际效应明显，就是再扩大组数到60组，效果有限。  
一致性也下降了。更极端的说，如果加上无数组，那就相当于是一个分类模型了。

Dim v.s.Perf




# 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



results



THANKS FOR YOUR ATTENTION.