

# Kaggle - LLM Prompt Recovery

---

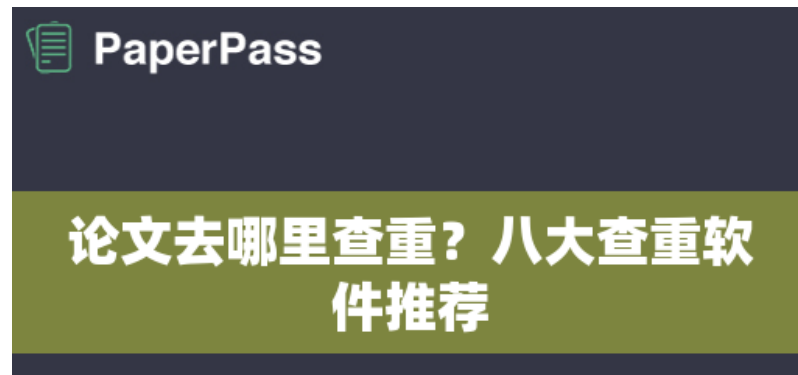
赛后讲解

主讲人：H老师

# 赛题背景

## 竞赛概览

近年来，大型语言模型LLM通常用于重写文本或对文本进行风格更改。本次比赛的目标是**恢复用于转换给定文本的 LLM 提示语 (prompt)**。



NLP工作流程越来越多地涉及重写文本，但关于如何有效地改进LLM 提示语 (prompt)，仍有很多东西需要学习。这次竞赛旨在以一种新颖的方式深入研究这个问题。

挑战：恢复用于重写给定文本的 LLM 提示语。您将针对包含 1300 多个原始文本的数据集进行测试，每个文本都与来自Gemma（Google 的新开放模型系列）的重写版本配对。

竞赛概览

```
test.head()
```

id		original_text	rewritten_text
0	1	Memo\n\nTo: All Staff\nFrom: Keith\nSubject: Fish Tank Maintenance\n\nPlease be reminded to clean the fish tanks in the office every Friday. This ensures a healthy environment for our beloved fish. Thank you for your cooperation.	Memo\n\nTo: All Staff\nFrom: Keith\nSubject: Fish Tank Maintenance\n\nLet's dive into a fin-tastic Friday tradition! Show our beloved fish some love by giving their tanks a sparkling clean. Together, we'll create a vibrant underwater oasis. Thanks for making a splash with your cooperation!
	1	Hey Suzann! Just wanted to let you know that Grandma and Grandpa are coming over for dinner tonight. Can you pick up some dessert on your way home? Thanks!	Hey Suzann! I understand it's uncertain, but could you possibly grab dessert on your way home? Grandma and Grandpa are coming over for dinner tonight. Thanks!
	2	Amy plucked the strings of her zither, filling the room with enchanting melodies. The delicate instrument, adorned with intricate carvings, transported her to a world of tranquility and beauty.	Amy plucked the strings of her zither, finding balance and stability in its cleat-adorned design. The enchanting melodies transported her to a world of tranquility and beauty.
	3	Betty's peony bloomed beautifully, filling her garden with vibrant colors and a sweet fragrance.	Betty's parser parsed perfectly, processing data efficiently and providing accurate results.
	4	The cornet, a brass instrument similar to a trumpet, was invented by Michael in the early 19th century. It is commonly used in jazz and brass bands for its bright and mellow sound. The cornet is played by pressing the valves to change the pitch and produce different notes.	Summary: The cornet, invented by Michael in the 19th century, is a brass instrument with a bright and mellow sound. It is widely used in jazz and brass bands, played by pressing valves to change pitch and produce various notes.

# 赛题背景

评价指标: Sharpened Cosine Similarity (锐化余弦相似度)

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

$$\text{Score} = \text{similarity}^3$$

使用sentence-T5 base模型计算每个提交的预测提示语与真实提示语的嵌入向量 (embeddings) , 通过锐化余弦相似度计算得分, 采用3次方减弱错误答案的得分过高问题。

# 数据概览

## 数据分析：官方数据集

竞赛官方只提供了1个训练样本和1个测试样本，无法分析得到有效信息

train.csv (1.4 kB)

⬇️ ⌂ ➡️

DetailCompactColumn


4 of 4 columns ▾

🔗 id	🔗 original_text	🔗 rewrite_prompt	🔗 rewritten_text
1 total values	1 unique value	1 unique value	1 unique value
-1	The competition dataset comprises text passages that have been rewritten by the Gemma LLM according ...	Convert this into a sea shanty: ""The competition dataset comprises text passages that have been re...	Here is your shanty: (Verse 1) The text is rewritten, the LLM has spun, With prompts so clever, they...

No more data to show

## 数据分析：辅助数据集

本次比赛官方并不提供足够的样本数据集，额外补充一个类似的生成数据集，即辅助数据集，从此数据集入手分析。

 KAGGLE AND 1 COLLABORATOR · UPDATED 2 MONTHS AGO


▲ 24

New Notebook

Download (81 kB)

⋮

### Supplementary Texts Rewritten by GenAI Models



[Data Card](#) [Code \(9\)](#) [Discussion \(0\)](#) [Suggestions \(0\)](#)

---

#### About Dataset

Supplementary data that can be used in LLM-centric Kaggle Competitions such as <https://www.kaggle.com/competitions/llm-detect-ai-generated-text> and <https://www.kaggle.com/competitions/llm-prompt-recovery>

**Usability** ⓘ  
4.71

**License**  
Unknown

**Expected update frequency**  
Not specified

**Tags**  
Computer Science



## 文本类型构成

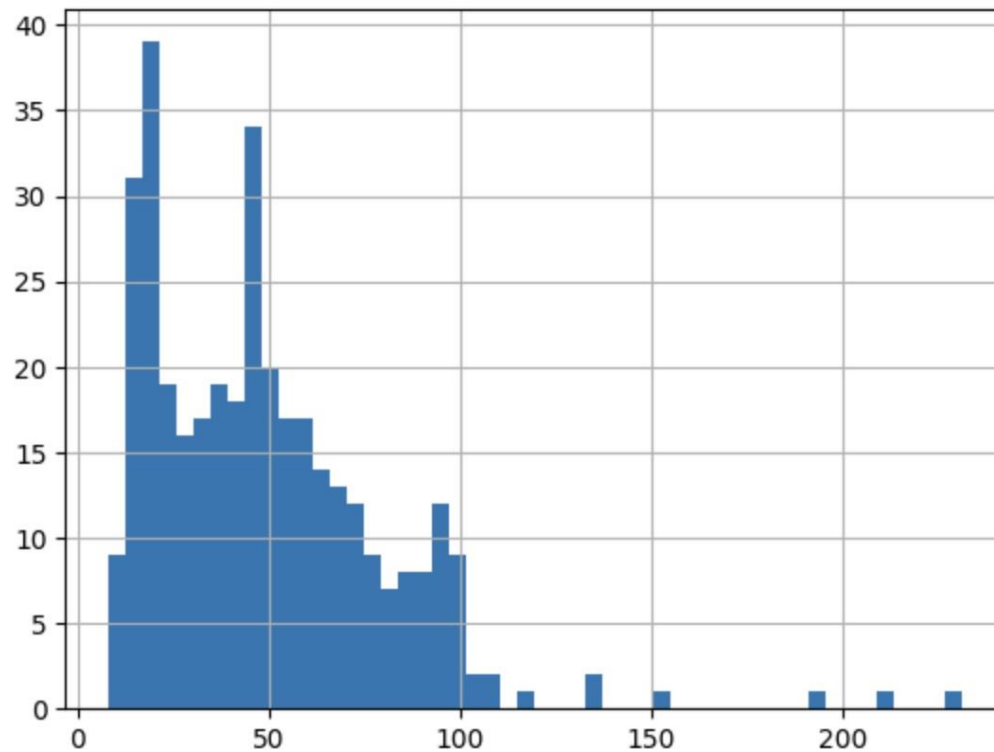
test.head()

	id	original_text	rewritten_text
Memo	0 1	Memo\n\nTo: All Staff\nFrom: Keith\nSubject: Fish Tank Maintenance\n\nPlease be reminded to clean the fish tanks in the office every Friday. This ensures a healthy environment for our beloved fish. Thank you for your cooperation.	Memo\n\nTo: All Staff\nFrom: Keith\nSubject: Fish Tank Maintenance\n\nLet's dive into a fin-tastic Friday tradition! Show our beloved fish some love by giving their tanks a sparkling clean. Together, we'll create a vibrant underwater oasis. Thanks for making a splash with your cooperation!
Message	1 2	Hey Suzann! Just wanted to let you know that Grandma and Grandpa are coming over for dinner tonight. Can you pick up some dessert on your way home? Thanks!	Hey Suzann! I understand it's uncertain, but could you possibly grab dessert on your way home? Grandma and Grandpa are coming over for dinner tonight. Thanks!
Narrative	2 3	Amy plucked the strings of her zither, filling the room with enchanting melodies. The delicate instrument, adorned with intricate carvings, transported her to a world of tranquility and beauty.	Amy plucked the strings of her zither, finding balance and stability in its cleat-adorned design. The enchanting melodies transported her to a world of tranquility and beauty.
Narrative	3 4	Betty's peony bloomed beautifully, filling her garden with vibrant colors and a sweet fragrance.	Betty's parser parsed perfectly, processing data efficiently and providing accurate results.
Informative	4 5	The cornet, a brass instrument similar to a trumpet, was invented by Michael in the early 19th century. It is commonly used in jazz and brass bands for its bright and mellow sound. The cornet is played by pressing the valves to change the pitch and produce different notes.	Summary: The cornet, invented by Michael in the 19th century, is a brass instrument with a bright and mellow sound. It is widely used in jazz and brass bands, played by pressing valves to change pitch and produce various notes.

## 文本长度分布

```
test['original_text_len'].hist(bins=50)
```

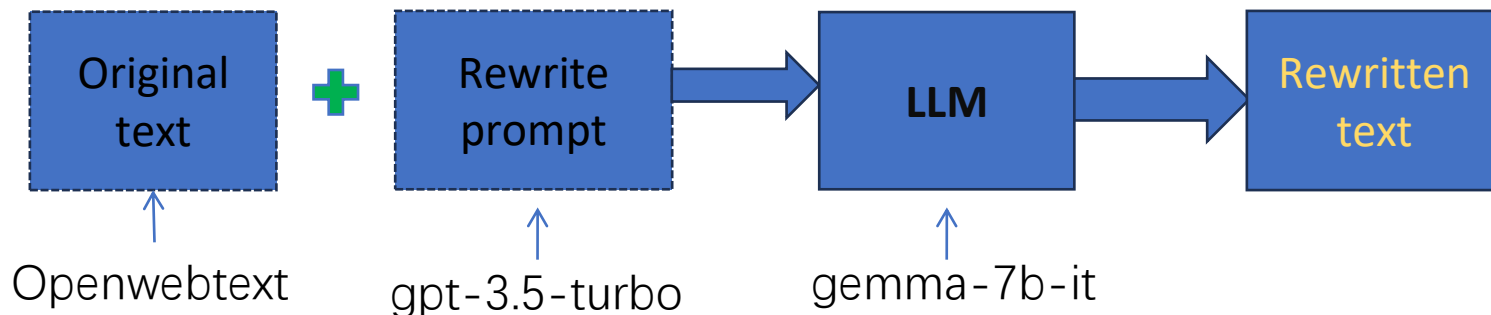
<Axes: >





# 算法详解

## 数据生成



Original text:原文， 由一个开源的互联网搜集文本Openwebtext数据集提供， 移除词数长度大于512

Rewrite prompt: 重写文本的提示语， 调用openai的大模型gpt-3.5-turbo的API批量生成

LLM: 重写文本的大模型， 使用限定的gemma-7b-it， 是谷歌最近开源模型的指令微调版本

<https://www.kaggle.com/models/google/gemma/pyTorch/7b-it-quant>

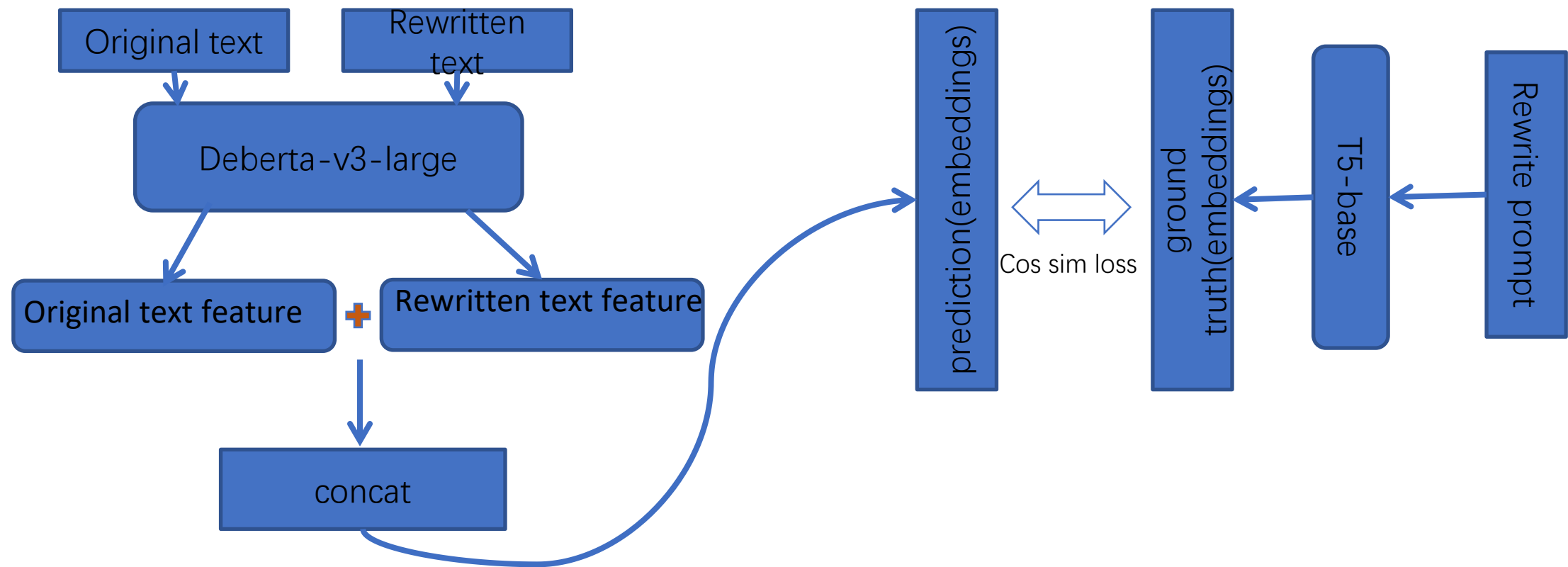
将原文随机与重写提示语结合， 组成文本对， 输入到重写大模型， 生成重写文本Rewritten text， 进行清洗与过滤， 得到最终训练数据集。

数据生成样本

	original_text	rewrite_prompt	rewritten_text	strlen
0	Your dryer can do more than what you think it can. It can also handle soft toys, kids clothes, and down jackets — it's quicker and easier than hanging them to dry.\n\nFind out more about our UltimateCare 800 here: <a href="http://bit.ly/ElectroluxMYEDH903BEWA">http://bit.ly/ElectroluxMYEDH903BEWA</a>	"I was a detective with a haunted past, and she was a mysterious beauty with a past shrouded in darkness. Together, we danced on the edge of a knife, the line between love and betrayal blurring with each whispered confession."	Summary:\n\nThe text describes a captivating story about a detective and a mysterious beauty with a haunted past. It emphasizes the blurred line between love and betrayal, hinting at a tense and thrilling narrative.	248
1	my acronym J Joyful I Influential N Nerdy Name / Username: urlLink Name Acronym Generator From urlLink Go-Quiz.com J Joyous I Industrious N Natural N Normal E Exhausting R Radical S Stunning Name / Username: urlLink Name Acronym Generator From urlLink Go-Quiz.com	Implement a peer coaching program using the insights provided in this content, to foster mentorship and skill development among team members.	Peer Coaching Program Proposal\n\nBased on the content provided, here's a proposed peer coaching program:\n\nOverall Goal:\n\nFoster mentorship and skill development among team members\n\nCreate a collaborative learning environment\n\nImprove overall performance and effectiveness\n\nTarget Audience: All team members\n\nDuration: 6 months\n\nStructure:\n\n1. Identify Coaches and Mentees:\n\nSelect experienced team members to be coaches and match them with junior team members who need guidance.\n\nConsider individual strengths and areas where improvement is needed.\n\n2. Establish Ground Rules:\n\nDefine clear roles and responsibilities for both coaches and mentees.\n\nSet expectations for communication, feedback, and support.\n\nEstablish boundaries to ensure a professional and respectful environment.\n\n3. Set Goals and Objectives:	390
2	We are about halfway through the process of moving house. We've got our mortgage approved, and had the survey done. This resulted in having further checks carried out on the wiring and gas appliances, plus checking for damp. This has resulted in us asking for Â£7K off the price! We are waiting for a response on that. In the meantime, the solicitor acting for our buyers has spotted that we don't appear to have permission from the freeholder for the extension. The extension had been built by the previous owners, so our solicitor should have spotted this. We are now having to apply for permission retrospectively, and I have told the old solicitor I will be charging our costs to them.	Take a poem and rewrite it in the style of a different poet, such as turning a Shakespearean sonnet into a contemporary free verse poem.	A symphony of moving, a melody of strife,\n\nA halfway point etched upon the life.\n\nWith mortgage approved and a survey complete,\n\nThe wiring and gas dance to a sinister beat.\n\nThe damp whispers secrets to the walls,\n\nThe checks reveal a tale of the past,\n\nThe extension stands as a testament to memories formed,\n\nBut missing the sanction that has been torn.\n\nThe solicitor's pen casts doubt upon the deed,\n\nA missing permit a puzzle to be freed.\n\nThe application beckons a journey to the past,\n\nRetrospective permission to be won at last.\n\nIn the midst of this	695
3	Wonder if my concert going buddy wants to go to Machine Gun Kelly\u002c Rick Ross\u002c Meek Mill\u002c and Wale Nov 20 in Indy ?? @user	Create a news report or headline summarizing the main points of the text.	Music Lovers Rejoice: Machine Gun Kelly, Rick Ross, Meek Mill, and Wale Set to Rock Indy on Nov.	137
4	I NEED HELP! No, I haven't relapsed into a drunken-drugged out stupor... I need help finding a location for the birthday party I'm throwing with urlLink Miss Laura "I'm An Honorary Asian Jew" ModernAge and Dan Byrne. We were hoping to have it at this one place, but that's being closed for renovation. Now, I need to know if anyone knows of any bar that is typically empty and could really use the business that the three of us would be bringing them. There are so many bars in Lower Manhattan, one of them should love to have 100+ retards celebrating and singing karaoke on February 7. Please email me if you can help. Thank you.	Create a meme or comic based on the events described in the paragraph.	The Party Fails\n\nPanel 1: Miss Laura, Dan Byrne, and the narrator are huddled on a couch, looking despondent.\n\nSpeech bubble above narrator: "No, I haven't relapsed! Just...cancelled."\n\nPanel 2: The narrator's phone rings. They pick it up excitedly.\n\nSpeech bubble above narrator: "Someone has a suggestion!"\n\nPanel 3: The narrator's eyes light up and they smile ominously.\n\nSpeech bubble above narrator: "The perfect place..."	634

# 算法详解

## 1. Seq2Seq模型



## 2. Few-shot LLM

The three settings we explore for in-context learning

### Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French:  ← task description
2 cheese =>                   ← prompt
   .....
```

### One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French:  ← task description
2 sea otter => loutre de mer    ← example
3 cheese =>                    ← prompt
   .....
```

### Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1 Translate English to French:  ← task description
2 sea otter => loutre de mer    ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese =>                    ← prompt
   .....
```

Task description:

Provide the new text and I will tell you what new element was added or change in tone was made to improve it - with no references to the original.

Example one:

original\_text: "Hey there! Just a heads up: our friendly dog may bark a bit, but don't worry, he's all bark and no bite!"

rewritten\_text: "Warning: Protective dog on premises. May exhibit aggressive behavior. Ensure personal safety by maintaining distance and avoiding direct contact."

prompt: "Please improve this text using the writing style of a warning."

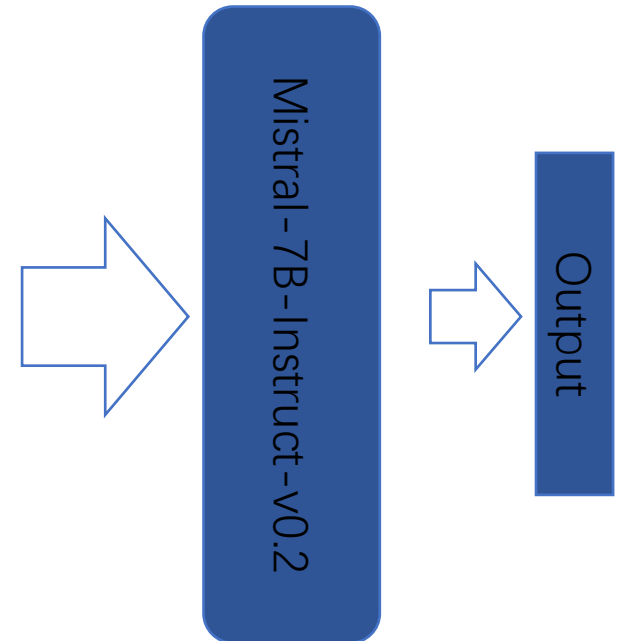
Example two:

original\_text: "The park was empty, save for a solitary figure sitting on a bench, lost in thought. The quiet of the evening was punctuated only by the occasional rustle of leaves, offering a moment of peace in the chaos of city life."

rewritten\_text: "Beneath the cloak of twilight, the park transformed into a realm of solitude and reflection. There, seated upon an ancient bench, was a lone soul, a guardian of secrets, enveloped in the serenity of nature's whispers. The dance of the leaves in the gentle breeze sang a lullaby to the tumult of the urban heart."

prompt: "Please improve this text using the writing style to be more poetic."

Example xx:...



3. Phi2



LUMOS · COPIED FROM LUMOS +121, -55 · 1MO AGO · 3,028 VIEWS

▲

48

Copy & Edit

272



[0.61+]LLMPR phi2 sft model training

Python · [llm\\_whls](#), [llmpr\\_public\\_10k\\_unique](#), [LLM Prompt Recovery](#) +1 · [Phi](#) (2)

- Notebook
- Input
- Output
- Logs
- Comments (10)



Competition Notebook

[LLM Prompt Recovery](#)

Run

3.5s

Version 1 of 1

Phi2 SFT training baseline

data

use all public data, but I dropped the dupilcate rewrite prompts.

hyperparamters

epoch: 5

<https://www.kaggle.com/code/mozhiwenmzw/0-61-llmpr-phi2-sft-model-training>

<https://www.kaggle.com/code/mozhiwenmzw/0-61-llmpr-phi2-sft-model-generate-infer>

## 三模型集成

