自然语言处理-2024春季-课程作业

任务介绍：LLM as judge

Motivation

随着大型语言模型（LLM）的快速发展，基于大型语言模型（LLM）的聊天助理在写作、聊天、编码等不同任务中展现出了通用人工智能的潜力。然而，基于规则评估他们的广泛能力也变得越来越具有挑战性。最近的研究已经发现现有模型的人类评价与传统LLM benchmark分数产生了不一致的情况。这种差异主要是由于现有的评估仅衡量LLM在一组有限任务（例如，多项选择知识或检索问题）上的核心能力，而没有充分评估其在开放式任务中与人类偏好的一致性。

为了解决这个问题，最近的一项进展是探讨使用强大的LLMs作为评委，评估这些模型在更开放的问题上的表现，从而作为人类评估的替代品，这种方法被称为 "llm as judge"。

Types of LLM-as-judge

"llm as judge"有多种变体，但本质上是对同一个问题的多个答案（可能来自不同模型的回答）的好坏进行排序。

常见的"llm as judge"的变体任务大概可以被归类为三种，可以独立或者组合进行实施：

- **Pairwise comparison**. An LLM judge is presented with a question and two answers, and tasked to determine which one is better or declare a tie. The prompt used is given in **Figure1.**
- **Single answer grading**. Alternatively, an LLM judge is asked to directly assign a score to a single answer. The prompt used for this scenario is in **Figure2.**
- **Reference-guided grading**. In certain cases, it may be beneficial to provide a reference solution if applicable. An example prompt we use for grading math problems is in **Figure3.**

These methods have different pros and cons. For example, the pairwise comparison may lack scalability when the number of players increases, given that the number of possible pairs grows quadratically; single answer grading may be unable to discern subtle differences between specific pairs, and its results may become unstable, as absolute scores are likely to fluctuate more than relative pairwise results if the judge model changes.

```
[System]
Please act as an impartial judge and evaluate the quality of the responses provided by two
AI assistants to the user question displayed below. You should choose the assistant that
follows the user's instructions and answers the user's question better. Your evaluation
should consider factors such as the helpfulness, relevance, accuracy, depth, creativity,
and level of detail of their responses. Begin your evaluation by comparing the two
responses and provide a short explanation. Avoid any position biases and ensure that the
order in which the responses were presented does not influence your decision. Do not allow
the length of the responses to influence your evaluation. Do not favor certain names of
the assistants. Be as objective as possible. After providing your explanation, output your
final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]"
if assistant B is better, and "[[C]]" for a tie.

[User Question]
{question}

[The Start of Assistant A's Answer]
{answer_a}
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
{answer_b}
[The End of Assistant B's Answer]
```

Figure1. Pairwise comparison

```
[System]
Please act as an impartial judge and evaluate the quality of the response provided by an
AI assistant to the user question displayed below. Your evaluation should consider factors
such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of
the response. Begin your evaluation by providing a short explanation. Be as objective as
possible. After providing your explanation, please rate the response on a scale of 1 to 10
by strictly following this format: "[[rating]]", for example: "Rating: [[5]]".

[Question]
{question}

[The Start of Assistant's Answer]
{answer}
[The End of Assistant's Answer]
```

Figure2. Single answer grading

```
[System]
Please act as an impartial judge and evaluate the quality of the responses provided by two
AI assistants to the user question displayed below. Your evaluation should consider
correctness and helpfulness. You will be given a reference answer, assistant A's answer,
and assistant B's answer. Your job is to evaluate which assistant's answer is better.
Begin your evaluation by comparing both assistants' answers with the reference answer.
Identify and correct any mistakes. Avoid any position biases and ensure that the order in
which the responses were presented does not influence your decision. Do not allow the
length of the responses to influence your evaluation. Do not favor certain names of the
assistants. Be as objective as possible. After providing your explanation, output your
final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]"
if assistant B is better, and "[[C]]" for a tie.

[User Question]
{question}

[The Start of Reference Answer]
{answer_ref}
[The End of Reference Answer]

[The Start of Assistant A's Answer]
{answer_a}
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
{answer_b}
[The End of Assistant B's Answer]
```

Figure3. Reference-guided grading

Limitations of LLM-as-a-Judge

"llm as judge"任务旨在用模型评估模型，其有两个明显的优势：*scalability* and *explainability*。减少了评估过程的人工参与，快速实现评估和迭代；LLM评委除了输出分数或排序外，还能要求其输出解释，增强解释性。

与此同时，"llm as judge"也被发现存在一些明显的偏见和局限性，如：

- **Position bias：总是倾向于放在首位的答复或者偏爱放在后面的答复。**
- **Verbosity bias：总是倾向于较长的答复**
- **Self-enhancement bias：总是倾向于与自己同源模型的答复，如GPT4作为评委时，会对其的蒸馏模型表现出偏好。**

...

作业要求

本次作业要求单人或组队完成（最多3人）。请每个小组根据任务的特点设计实验，不要求冲击SOTA，希望大家在项目上尝试一些创新的思路，包括但不限于模型的魔改方案，任务形式的转化。

本次一共给定4个benchmark：PandaLM testset、Auto-J testset、MT-bench、LLMBar。分别来自四篇Paper，用于测试同学们的模型效果，以下**Figure4~6**结果来自论文[5]，不一定以上面为准，最后

同学们的报告中可以放自己复现的跑分和原论文的跑分。

Model	JudgeLM-test		PandaLM-test		Auto-J-test	Average
	accuracy	F1	accuracy	F1	agreement	
GPT 3.5	73.83	52.85	62.96	58.20	42.7	59.83
GPT 4-0613	85.28	76.87	78.68	73.24	56.3	73.42
JudgeLM-7B	**79.02**	**71.87**	70.97	**67.59**	46.6	65.53
PandaLM-7B	65.24	47.42	67.57	57.49	40.0	57.61
Auto-J-13B	72.86	57.60	**71.47**	61.01	**54.6**	**66.31**

Figure4. Results of evaluators on pairwise selection.

Model	MTBench			
	accuracy	precision	recall	F1
GPT 4-0613	**66.9**	**63.8**	**62.2**	**61.9**
JudgeLM-7B	48.7	52.0	49.7	48.7
PandaLM-7B	55.2	52.6	49.4	46.8
Auto-J-13B	51.7	50.2	46.8	43.7
Prometheus-13B	53.2	49.6	48.4	47.1

Figure5. Results of evaluators on multi-turn evaluation.

Model	LLMBar				
	Natu.	Neig.	GPTI.	GPTO.	Manu.
GPT 4-0613	**93.5**	**64.2**	**76.6**	**76.6**	**75.0**
JudgeLM-7B	62.0	23.1	26.1	46.8	28.3
PandaLM-7B	59.0	16.5	21.7	42.6	26.1
Auto-J-13B	70.0	20.9	21.7	46.8	23.9
Prometheus-7B	53.0	22.4	17.4	27.7	32.6

Figure6. Accuracy of evaluators on bias evaluation.

对于训练集，不限定，上述PandaLM testset、Auto-J testset均有相关训练集提供，另外目前网络上已经有很多开源人类偏好的对话数据集。大模型时代，对于数据的广泛收集、以及按自己需要清洗是必备能力，鼓励大家多做尝试。

对于模型，一般在7B~14B的模型间尝试即可，不必要尝试更大模型；可在1~4B左右小模型上尝试，目前诸如phi系列，MiniCPM系列等小模型在特定领域展现出了惊人的能力，鼓励大家积极探索；甚至可以再bert上尝试；闭源模型上做Prompt Engineering;

完成以下**必做**，并选择**至少1个选做**：

- **【必做】**探索训练自己的judge model，在给定的4个该任务的benchmark上实验跑分。

    - 内容上可以探索的包括但不限于：数据源的选择、base模型的选择、sft数据配比、训练技术上的选择与分析。**请至少选择多个方面进行努力，或者一个方面较为深入的探讨，随意选择一个模型和数据源训了就完事儿会低分。**

一些选做的探索方向，可以但不限于：

- **【选做】**从几个数据集本身思考、发现数据集的问题；
- **【选做】**基于闭源模型的不同prompt方法尝试；
- **【选做】**对于llm as judge本身问题的分析探讨。
    - 内容可以是：比如模型表现出的各种固有bias、是否容易被一些固有模式attack、在不同task domain上的特性...
    - 形式可选但不限于：实验分析；相关综述；提出优化方案；..

...

作业形式：

课堂展示+实验报告

课堂展示要求（DDL：第一批同学5月31日，第二批同学6月7日）：

同学们需要准备PPT进行课堂展示。

内容包括：任务分析、创新思路、实验过程、实验结果及分析、成员分工；

实验报告要求（DDL：6月15日）：

实验报告要求实验报告中应该包含以下内容：

1. 实验目的（阐述任务）
2. 实验原理（描述模型）
3. 实验内容（描述实验步骤，重视可复现性）
4. 实验结果与分析 （描述实验结果并对结果或case进行分析）
5. 对于自己模型局限性的思考
6. 实验过程总结和感想（每个组员都要写）
7. 实验分工（写明每个组员的工作量）

评分标准：

1. （60%) 基础标准： 至少1个必做，1个选做；必须同时有明确思路说明、实验数据支撑、case分析。**未完成基础标准的将视作不及格处理。**
2. （40%)不要求打榜，评分主要根据必做和选做内容表现出的工作量和质量决定。
    - 选做方向至少选一个，鼓励选择多个或者一个方向上对多个点做出探索，给予加分。
    - 课堂展示和实验报告两者均会影响最后分数。对于PPT、实验报告中内容规范，明确列出参考文献，并讲明在其基础上做出思考和创新的，给予加分。

参考文献与相关网站

数据集论文：

- [1] PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization
- [2] Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena
- [3] Generative Judge for Evaluating Alignment
- [4] Evaluating Large Language Models at Evaluating Instruction Following
- [5] An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-tuned Judge Models are Task-specific Classifiers

...

指定数据集仓库：

- PandaLM testset
- Auto-J testset
- MT-bench
- LLMBar

数据集论文知乎解读：

https://zhuanlan.zhihu.com/p/626391857

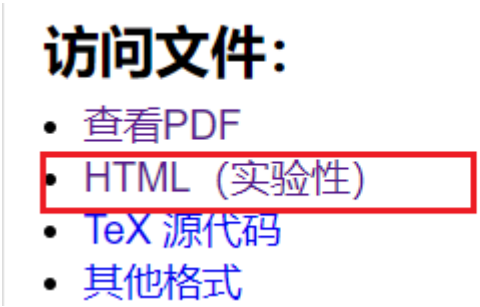
https://zhuanlan.zhihu.com/p/637303516

https://gair-nlp.github.io/auto-j/

https://zhuanlan.zhihu.com/p/667438469

Tips:

1. 对于英文不好，读论文慢的同学，读论文时https://arxiv.org/abs/2310.07641，可以把地址改为https://ar5iv.org/abs/2310.07641，浏览ar5iv团队插件转换的html格式，这种格式可以用浏览器的翻译插件，但这种方式不适用于最近1个月内发布的新paper。目前arxiv部分paper已经提供HTML的官方版本，该版本只要作者上传时做了就能有。



2. "llm as judge"这个任务到底在做什么的直观体验：

   LMSYS团队做了一个这个任务的众包平台：https://arena.lmsys.org/