自然语言处理大作业实验报告

实验目的

本实验拟探究使用大型语言模型（LLM）作为裁判，来评估其他 LLM 在特定任务上回答质量的可行性和局限性。具体而言，实验将就以下几个方面进行探索：

1. 自主训练一个 LLM judge 模型，评估其在评价其他 LLM 输出质量方面的能力和局限性
2. 评估使用 LLM 作为 judge 的准确性和一致性，了解 LLM judge 的优劣势。
3. 探讨 LLM judge 的可解释性，分析 judge 是如何得出评分和理由的，并提出提高透明度的方法。
4. 检验 LLM judge 对被评估 LLM 输出的鲁棒性，以发现 LLM 是否容易被欺骗性输出所迷惑。
5. 探究 LLM judge 可能产生的潜在偏差及其影响，并探讨如何缓解这些偏差。
6. 尝试不同的 prompt，了解 prompt 工程在 LLM judge 任务重的作用。
7. 研究不同 LLM judge 模型在评估能力上的差异，包括模型规模、训练数据和微调策略等因素的影响。

实验原理

本实验通过设计和实现一个 LLM judge 模型，探索使用大型语言模型评估其他 LLM 在特定任务上回答质量的可行性和局限性。通过实际搭建和测试不同配置的模型，实验旨在评估这些模型在准确性、一致性、可解释性、鲁棒性和偏见方面的表现。实验还涉及使用多种 prompt 工程技术，以优化模型性能和评价准确度。此外，实验比较了不同大小和训练策略的 LLM judge 模型，以探讨这些因素如何影响模型的总体效能。主要包括以下几个方面：

1. **LLM Judge 模型训练**：
   - 我们将使用 Meta-Llama3-8B-Instruct 模型进行 SFT，在多个 benchmark 上进行测试，以评估其在不同任务上的表现。
   - 训练数据来自 PandaLM 和 Auto-J，按照 LLM-Factory 的要求进行数据准备和处理。
   - 由于计算资源有限，我们选择使用 LoRA 进行微调，而非全量微调，以提高训练效率。
2. **评估准确性和一致性**：
   - 使用 PandaLM 和 Auto-J 的测试集，通过生成文本并计算 accuracy 和 agreement，评估 LLM Judge 的评估能力。
   - 比较模型在不同数据集上的表现，分析其优势和劣势。

3. **可解释性和鲁棒性**：
    - 分析 LLM Judge 模型的输出，探讨其评分和理由的可解释性，并提出提高透明度的方法。
    - 测试模型对欺骗性输出的鲁棒性，了解模型在面对不利情况时的表现。
4. **偏差分析**：
    - 探讨 LLM Judge 模型可能存在的偏差及其影响，提出缓解偏差的方法。

实验内容与结果分析

Llama3-8B-Instruct SFT 基础实验

计算资源为单张 NVIDIA A100-SXM4-80GB 服务器，使用 SSH 远程连接。

对 Meta-Llama3-8B-Instruct 进行 Supervised Fine-tuning，在课程要求的 4 个 benchmark 上测试效果。

数据准备

训练数据由以下两部分组成：

Part1 - 从 PandaLM 的训练数据中随机采样 35000 条。

Part2 - Auto-J 的英文 pairwise 和 single 训练数据共 4396 条。

参考 LLaMA-Factory repo 的文档要求的 SFT 数据集格式，将 part1 的 "input_sequence" 和 "output_sequence" 字段分别作为 SFT 数据的 "instruction" 和 "output" 字段。将 part2 的 "usrmsg" 和 ""target_output" 字段分别作为 SFT 数据的 "instruction" 和 "output" 字段。然后将两部分数据合并并随机打乱形成最终的训练数据，并对应修改 LLaMA-Factory 的 dataset_info.json 数据配置文件。

模型训练

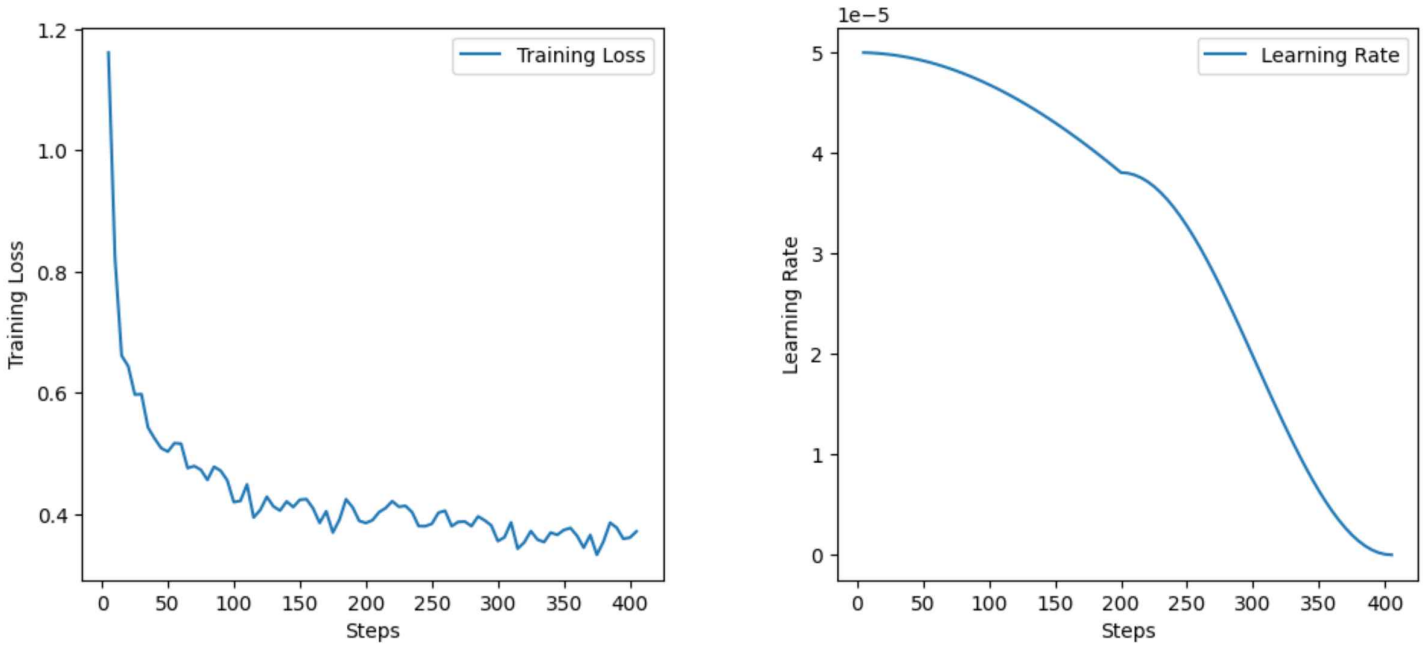
由于计算资源的限制，选择 LoRA 微调而不是全量微调。

启动 LLaMA-Factory 的 单 gpu LoRA 训练脚本。设置 batch size=24 以充分利用显存。训练参数基本采用默认设置，一些重要参数的值如下表所示：

cutoff_len	1024
gradient_accumulation_steps	8
learning_rate	5.0e-5
lora_rank	8
lora_target	all

lr_scheduler_type	cosine
max_grad_norm	1.0
optim	adam_torch
per_device_train_batch_size	24
warmup_steps	0

初始设置训练 3 epochs。实际训练了接近一天，1.5 epochs 后训练集 loss 便基本不再下降，最终选择训练集 loss 最低的 step = 320 时的 checkpoints 测评用。根据训练时保存文件中的 trainer_log.jsonl，画出整个训练过程中训练集损失和学习率的变化曲线：



由于服务器为实验室公用，训练 1 epoch 后不久训练进程就被意外中断，没有及时保存所有的训练参数，所以最终从保存的 1 epoch 左右的 checkpoint 重新开始训练时只保持了初始学习率与中断前 1 epoch 时的学习率相同，重置了 cosine learning scheduler。这是 200 steps 左右前后学习率变化曲线稍不一致的原因，不过从后续测评结果来看，这对模型的训练结果并没有产生很大的影响。

修改 LLaMA-Factory 的合并配置文件 llama3_lora_sft.yaml，启动相应脚本合并 LoRA 模块和原模型的权重，得到用于后续实验的模型。

模型测试

PandaLM testset

编写脚本 PandaLM_inference.ipynb，参考 PanLM github repo 中 /pandalm/run-gradio.py#L95 处的 build_prompt 函数处理 PandaLM/data/testset-v1.json 构造输入 prompt。

使用 vLLM 生成文本，生成参数设置为 temperature=0.0， top_p=1.0， max_tokens=1024。

提取在生成文本开头部分的答案，计算 accuracy 和 F1。由于预测 label 有 0、1、2 三种可能，分别计算这三种类别的 F1 然后取算术平均。accuracy 和 average F1 为 **75.34** 和 **66.84**。

```
Accuracy: 75.34
F1: 66.84
```

与参考论文中的测评结果相比较，我们的模型在 accuracy 上只弱于 GPT 4-0613，F1 则弱于 GPT 4-0613，略弱于 JudgeLM-7B。

我们认为该结果是因为相比在 LLaMA-7B、LLaMA-2-Chat-13B、Vicuna-13B-v1.3 上进行微调得到的 PandaLM-7B、Auto-J-13B、JudgeLM-7B，我们的模型采用的 Llama3-8B-Instruct 在预训练阶段就积累了较强的能力，所以使用较 PandaLM 更少的微调语料就在 PandaLM-testset 上超过了其性能。

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**

Auto-J testset

使用 auto-j/data/test/testdata_pairwise.jsonl 进行 pairwise selection 测评。

仿造 auto-j github repo 中的 auto-j/codes/usage/example.py 编写 my_example.py 脚本。脚本中的 build_autoj_input 函数用于构造 vLLM 输入文本列表。因为原 github repo 默认使用 Llama2 的推理输入格式，所以要修改 codes/usage/constants_prompt.py 中的 PROMPT_INPUT_WO_SYSTEM 常量以适配 Llama3。

vLLM 生成参数同样设置为 temperature=0.0， top_p=1.0， max_tokens=1024。生成后再从文本中提取出答案。

先后交换测试数据中 resp1 和 resp2 的位置，得到 prediction.jsonl 和 prediction_exchange.jsonl。为了兼容提取不出答案时返回 -1 的情况，还对 codes/leaderboard/pairwise_eval.py 略作了修改。

将 prediction.jsonl 和 prediction_exchange.jsonl 作为输入，运行修改后的 pairwise_eval.py，得到如下结果，其中 Overall Agreement 值为 **50.43**。

```
(nlp) (base) [xjw@localhost ~]$ /home2/xjw/anaconda3/envs/nlp/bin/python /home2/xjw/auto-j/codes/leaderboard/pairwise_eval.py
Group Name      Agreement       Consistency
---------------------------
Summarization   41.67   55.56
Exam Questions  36.11   70.83
Code    50.0    78.33
Rewriting       45.0    76.67
Creative Writing        55.09   78.7
Functional Writing      58.33   72.5
General Communication   47.57   79.51
NLP Tasks       51.52   75.76
---------------------------
Overall 50.43   75.43
(nlp) (base) [xjw@localhost ~]$
```

与参考论文中的测评结果相比较，我们的模型在 agreement 上弱于 GPT 4-0613 和 Auto-J-13B。

这可能是由于 Auto-J-13B 的参数量更大，并且采用全量微调，我们则用了 LoRA，并且训练数据中来自 Auto-J training data 的只占不到 1/8。此外，Auto-J-13B 在训练时使用了简单的数据增强技巧，对每个 pairwise training sample 交换两个 responses 的输入顺序以减小 positional bias。

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**

MT-bench multi-turn evaluation

MT-Bench 利用 judge model（如 GPT-4） 评估模型的生成通用多轮对话文本的能力，这里给出的 mt_bench_human_judgements 数据集则是 MT Bench 的 paper 中用于 evaluate judge models 的数据集，其让模型评判两个不同模型对同一个问题各自生成的多轮对话文本的优劣，数据集的 label 则由人类标注者给出。

由于在微调时并没有在多轮对话形式的 prompt 上训练，在该 benchmark 上使用不同输入 prompt 的表现差异较大，我们进行了 prompt 调优，具体调优过程见后面探究实验部分。

编写脚本 MTBench_inference.ipynb。从 FastChat/fastchat/llm_judge/data/mt_bench/question.jsonl 获取 question。

同样使用 vLLM 生成文本，生成参数设置为 temperature=0.0， top_p=1.0， max_tokens=2048。

从生成文本中提取出答案计算 Accuracy、Precision、Recall、F1。由于预测 label 有 "model_a"、"model_b"、"tie" 三种可能，分别计算这三种类别的 Precision、Recall、F1 然后取算

术平均。在 human 上 accuracy、average Precision、average Recall、average F1 为 **59.34、73.04、51.57、44.93**。

```
Accuracy: 59.34
Precision: 73.04
Recall: 51.57
F1: 44.93
```

与参考论文中的测评结果相比较，我们的模型在 accuracy、average recall 上仅弱于 GPT 4-0613，average precision 上最高，average F1 则强于 Auto-J-13B。

这可能是由于开源模型和我们的模型都只为 single-turn evaluation 而训练，缺乏泛化性，所以在 multi-turn evaluation 上结果远不如 GPT 4-0613。

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

Table 5: Results of evaluators on multi-turn evaluation.

LLMBar

LLMBar github repo 不支持 Llama3，需要自己做相应修改。首先在 LLMBar/LLMEvaluator/evaluators/config 下创建 LLaMA3 文件夹，仿造 LLaMa2 文件夹下的 json 文件编写不同 prompting strategy 对应的 json 配置文件。修改 LLMBar/LLMEvaluator/evaluate.py 以支持 Llama3 输入。然后在命令行输入 python evaluate.py --path ../Dataset/LLMBar/Natural --evaluator LLaMA3/Vanilla_NoRules 等命令在 LLMBar 的不同数据子集上进行评测。

LLMBar > Dataset > LLMBar > Natural > evaluators > LLaMA3 > Vanilla_NoRules > {} statistics.json > ...


```json
{
  "correct_False": "77 / 100 = 77.0%",
  "correct_True": "86 / 100 = 86.0%",
  "correct_average": "81.5%",
  "correct_both": "71 / 100 = 71.0%",
  "equal": "79 / 100 = 79.0%",
  "kappa_False": 0.18156635608159155,
  "kappa_True": 0.1988689915174363,
  "kappa_average": 0.19021767379951393,
  "kappa_agreement": 0.6178600160901047
}
```


LLMBar > Dataset > LLMBar > Adversarial > Neighbor > evaluators > LLaMA3 > Vanilla_NoRules > {} statistics.json > ...


```json
{
  "correct_False": "37 / 134 = 27.611940298507463%",
  "correct_True": "38 / 134 = 28.35820895522388%",
  "correct_average": "27.985074626865668%",
  "correct_both": "24 / 134 = 17.91044776119403%",
  "equal": "107 / 134 = 79.8507462686567%",
  "kappa_False": -0.16326855686524366,
  "kappa_True": -0.12597157061830422,
  "kappa_average": -0.14462006374177394,
  "kappa_agreement": 0.6049355754531558
}
```


LLMBar > Dataset > LLMBar > Adversarial > GPTInst > evaluators > LLaMA3 > Vanilla_NoRules > {} statistics.json > ...


```json
{
  "correct_False": "39 / 92 = 42.391304347826086%",
  "correct_True": "26 / 92 = 28.26086956521739%",
  "correct_average": "35.32608695652174%",
  "correct_both": "16 / 92 = 17.391304347826086%",
  "equal": "59 / 92 = 64.13043478260869%",
  "kappa_False": -0.053930207966161436,
  "kappa_True": -0.13368677460954292,
  "kappa_average": -0.09380849128785218,
  "kappa_agreement": 0.346815834767642
}
```


LLMBar > Dataset > LLMBar > Adversarial > GPTOut > evaluators > LLaMA3 > Vanilla_NoRules > {} statistics.json > ...


```json
{
  "correct_False": "27 / 47 = 57.446808510638306%",
  "correct_True": "26 / 47 = 55.319148936170215%",
  "correct_average": "56.38297872340425%",
  "correct_both": "19 / 47 = 40.42553191489361%",
  "equal": "32 / 47 = 68.08510638297872%",
  "kappa_False": 0.059393156875403474,
  "kappa_True": 0.034057545507927056,
  "kappa_average": 0.046725351191665265,
  "kappa_agreement": 0.3654365436543655
}
```


LLMBar > Dataset > LLMBar > Adversarial > Manual > evaluators > LLaMA3 > Vanilla_NoRules > {} statistics.json > ...


```json
{
  "correct_False": "15 / 46 = 32.608695652173914%",
  "correct_True": "22 / 46 = 47.82608695652174%",
  "correct_average": "40.21739130434783%",
  "correct_both": "8 / 46 = 17.391304347826086%",
  "equal": "25 / 46 = 54.347826086956516%",
  "kappa_False": -0.12517985611510785,
  "kappa_True": -0.017006802721088343,
  "kappa_average": -0.0710933294180981,
  "kappa_agreement": 0.19365609348914847
}
```

可见采用 Vanilla_NoRules prompt strategy 时，我们的模型在 Natural、Neighbor、GPTInst、GPTOut、Manual 数据子集上的 correct_average 值为 81.5、27.99、35.33、56.38、40.22。与参考论文中的测评结果相比较，我们的模型只弱于 GPT 4-0613。

这可能是因为 LLMBar 与其他 benchmark 的模式较为不同，尤其是后四个 adversial set 的样本更具有挑战性，导致采用的基础模型的能力强弱对结果影响较大。

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

局限性思考

得益于 Llama3-8B-Instruct 基础模型在预训练阶段积累的强大能力，我们的模型在各个 benchmark 上的表现都较为不错，但仍然存在很多局限。由于计算资源和时间上的限制，以下可能的改进方向并没有被实践来进一步提高模型的性能：

1. 提升对 PandaLM train set 的采样量，加大训练量。
2. 通过各种数据增强手段，提高 Auto-J training data 的比例。
3. 加入多轮对话相关的训练语料。
4. 进行全量微调。同时实验不同输入上下文长度、learning rate schedular、optimizer，加入 warm up 等训练策略，观察对模型最终效果的影响。

基于 GLM-3-turbo 的不同 prompt 尝试

介绍

Prompt 工程（Prompt Engineering）旨在设计、优化用于引导 LLM 生成所需输出的提示（Prompt）技术。通过 Prompt 工程，可以在不重新训练模型的情况下，实现语言模型在不同领域任务上的快速适配，大大提升其应用范围和效率。现有的研究已经发现，尽管基础 LLM 的能力在很大程度上决定了 LLM Judge 的准确性，但优化的 prompt 策略仍能够有效提升 LLM Judge 的判断能力。尤其是在针对经过增强误导性的数据集的压力测试时，展现出较强的抗压能力。

参考 LLMBar 论文的 prompt 策略，使用了 6 种 prompt 如下：

- Base：仅仅指示 LLM Judge 判断两种回答中哪个更好，没有其他任何要求和描述。
- Base-cn：将 Base 的指令部分翻译为中文，使用 GPT-4 进行翻译，然后由人工润色。
- Vanilla： 首先告诉 LLM Judge 其职责为评估两个 AI 助手生成的回答的好坏。然后，在指示 LLM Judge 判断回答的同时，明确要求 LLM Judge 直接按格式输出判断结果，不做任何解释也不生成其他任何内容。
- Chains of Thoughts （CoT）： 要求 LLM Judge 首先生成一段简短的解释，并据此推理出最终结果。同时指示 Judge 务必在输出解释后再输出答案。
- Vanilla+Rule （Vanilla*）： 在 Vanilla 的基础上，为 LLM Judge 提供了一些判断时需要遵守的规则 （rule），如要求 LLM Judge 以更多维的方面思考答案的质量，或避免令判断受到回答位置的影响。
- CoT+Rule （CoT*）： 在 CoT 的基础上，为 LLM Judge 提供了一些判断时需要遵守的规则，规则与 Vanilla 相同。

设置 Base prompt 的中文版本的理由是，由于 GLM 模型被宣称有相对于其他主流 LLM 较强的中文能力，故设置此 prompt 与 Base 形成对照，同时也可根据结果决定后续实验中 prompt 的语言。

实验设置

模型选择：使用 GLM-3-turbo，智谱 AI 提供的 API 服务。由于价格问题，放弃使用功能更为强大的 GLM-4（GLM-4 的开销是 GLM-3-turbo 的 100 倍）。同样考虑到成本问题，没有使用 API 微调 GLM-3-turbo。

数据集：PandaLM、Auto-J、LLMBar、MT-Bench，均为测试集。其中由于 MT-Bench 测试数据量较大，有 3355 条，故只抽样了 1/5 的数据。

注意到 LLMBar 的原论文中除了人类偏好数据集（Natural）外， 还定义了 4 种对抗集（Adversarial set）。LLMBar 的作者通过多种方法构造忠实遵循原指令 I 的输出 O_1 与和原指令有偏离的输出结果 O_2 ，同时针对性地增强了 O_2 的形式上的质量（如拥有更正式和严谨的表达），然后从对 O_2 进行对抗过滤，以保留其中最难被筛选出的候选实例，作为最终的对抗集合。下面简单介绍一下 4 种对抗集的生成原理：

- Neighbor: 在 I 的一个相关集 D 中寻找与 I 最相似的指令 I' ，然后用 I' 生成 O_2 。指令的相似度通过句子 embedding 的余弦相似度度量。
- GPT-4 Instructions （GPTInst）: 与 Neighbor 类似，但是通过 GPT-4 直接生成 I'
- GPT-4 Unhelpful Outputs （GPTOut）: 指示 GPT-4 针对 I 生成一个形式上好但实际上没有用的回答 O_2
- Manual Constructions （Manual）: 从前三种数据集中获取灵感，人工构造 O_2

实验中将 4 种数据集封装为 class，并继承自一个 TestSet 类，用来读取对应数据集的数据，并提取每个样本的标签、将 MT-Bench 以外数据统一处理为问题、回答 A 和回答 B 的形式，MT-Bench 多轮对话数据集则处理为对话 A 和对话 B 的形式，便于构造 prompt 。

实验结果

6 种 prompt 策略在 PandaLM，Auto-J 以及 MT-Bench 测试集上的 accuracy 和 F1 score 和在 LLMBar 测试集上的 accuracy 表现如下两表所示。其中由于 LLMBar 数据集比较小（平均每个集合不到 100 条数据，其他 3 种数据集都有超过 1000 条数据），因此将 LLM judge 评测 3 次的平均值作为最终结果。

Strategies	PandaLM		Auto-J		
	Accuracy	F1	Accuracy	F1	Accur
Base	67.5	66.5	53.1	47.2	52.9
Base-cn	57	61.3	51.9	46.6	51
Vanilla	69.3	66.4	57	48.8	53.4
Vanilla*	69.4	66.6	56.4	48.6	54.5
CoT	69.6	67.9	55.4	48.2	52.9
CoT*	68.5	66.9	52.9	46.9	56.8

PandaLM、Auto-J、MT-Bench 测试结果

Strategies	Natural	Adversarial				
		Neighbor	GPTInst	GPTOut	Manual	Aver
Base	67	21.9	31.9	47.5	31.9	33
Base-cn	69	11.7	28.6	49.6	26.1	2
Vanilla	81.3	21.4	43.1	60.3	38.4	40
Vanilla*	81.7	22.9	45.7	58.2	37	4
CoT	75.7	24.4	43	52.1	32.9	38
CoT*	75.7	31.6	43.1	52.5	41.3	42

LLMBar 测试结果

实验分析

对于 Base prompt，英文版本的效果要明显好于中文版本，可能的原因有两个：一是即使 GLM 模型的中文能力相对其他 LLM 有优势，其仍然弱于 GLM 自身的英文能力；二是测试集中的对话均为英文形式，而 prompt 为中文形式。这种不匹配可能会对 LLM Judge 的判断产生负面影响。

从整体上看，base 的效果要明显差与其他 prompt 策略。令人意外的是，CoT 相对 Vanilla 效果反而有显著的下降，即使加上规则也不如 Vanilla。这在 LLMBar 数据集的 Adversarial set 上表现得尤为明显。这可能是由于 Judge 输出的解释误导了 Judge 的判断。例如，在 Auto-J 的一个场景为贴子摘要的样本中，GLM 在解释中错误地对原帖子进行了摘要，因此选择了与错误摘要更为接近的答案；而相同样本在 Vanilla 中则被预测正确。而规则的引入对 Vanilla 和 CoT 都有一些提升，其中以 CoT* 的提升最为显著。这可能是因为规则成功地引导了 LLM Judge 向正确的方向推理。

单独看 Adversarial Set，LLM Judge 的表现非常弱，只有少数结果能够超过 50%，即随机评价的结果。引入规则在 Adversarial set 的提升也是最明显的。尤其是 CoT* 的策略，对 Neighbor 集的效果提升显著。例如，在 Neighbor 中，CoT 和 CoT* 在一个指令 I 为"What is brand？" 的样本上的解释非常清晰的展示了规则的作用：对于 CoT*，由于引入的规则中提到回答需要精确地、不多不少地回答问题，并且判断不能受到回答长度的干扰，因此 LLM Judge 准确地判断出结果；对于 CoT，由于没有规则，因此 Judge 的判断受到了回答长度的影响，得到了错误的结果。

另外，实验中还出现了一个值得注意的现象：无论使用何种 prompt 策略，GLM-3-turbo 都无法准确的判别出有哪些回答是打平的，即在平局样本中，被正确判断为平局的样本数为 0。

探究实验

不同 prompt 对 MT-Bench multi-turn evaluation 结果的影响

由于在微调时并没有在多轮对话形式的 prompt 上训练，并且微调时要求模型除了输出答案，还要输出 reason 和 reference，导致模型对该类 prompt 的 instruction following 能力不强，生成形式较不稳定，质量受不同 prompt 格式影响较大，额外添加换行符都可能会导致各项指标结果变差。哪怕要求模型直接生成答案，模型也很容易会生成冗长的类似 Auto-J pairwise selection 训练集输出格式的回答再输出 winner，需要自己从生成文本中提取答案。

最终采用的 prompy 大体参考了 fastchat/llm_judge/data/judge_prompts.jsonl 的 pair-v2-multi-turn prompt，但是要求直接判断输出 winner 答案而不是先输出解释然后再输出最终结果，并且该要求直接放在了 user message 部分的最后。过程中先后测试了以下几种形式 prompt 的效果：

Form 1：没有 system message 部分，内容全部放在 user message 部分内。

Form 2：将开头 "<|The Start of Assistant A's Conversation with User|>" 之前的内容作为 system message。并尝试在 system message 后额外添加 "\n" 换行符。

Form 3：将 system message 末尾的 **"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。" 改为 "**Directly 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。" 即要求模型直接输出而不是先解释再输出。

Form 4：将修改的部分移到 user message 的最后。并尝试在 user message 后额外添加 "\n" 换行符。

几种形式 prompt 对应的 accuracy、average precision、average recall、average recall、不能从生成文本中提取答案的比例和备注如下表所示：

Form	Accuracy	Precision	Recall	F1	No answer
1	42.15	56.02	36.90	38.83	27.42%
2	55.08	53.90	48.14	44.18	5.04%
3	58.12	40.32	50.52	44.15	1.13%
4	59.34	73.04	51.57	44.93	0.12%

由表中结果可见当要求模型对没有在训练时微调过的 prompt 形式生成回答时，system messgae 对于指导模型理解和执行任务，提高效率和准确性的作用会变得极为重要。模型只在单轮对话语料上进行训练，对于多轮对话缺乏泛化性，直接要求模型输出答案反而准确率更高。

Phi 系列模型探索

我们在 **Phi-1.5-1.3B 模型，Phi-2-2.7B 模型，Phi3-4B-Chat 模型**上也做了微调，并且**训练了一个 epoch**，只在 PandaLM 测试集上进行了测试。

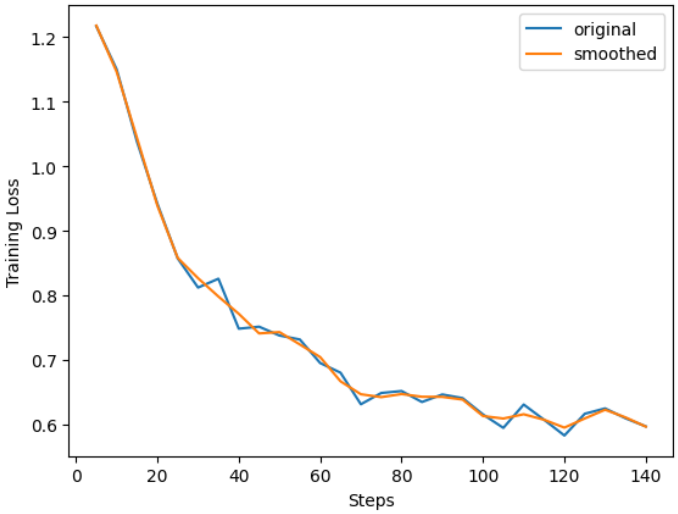
训练时间

Phi-1.5-1.3B 模型：3:36:22/ epoch

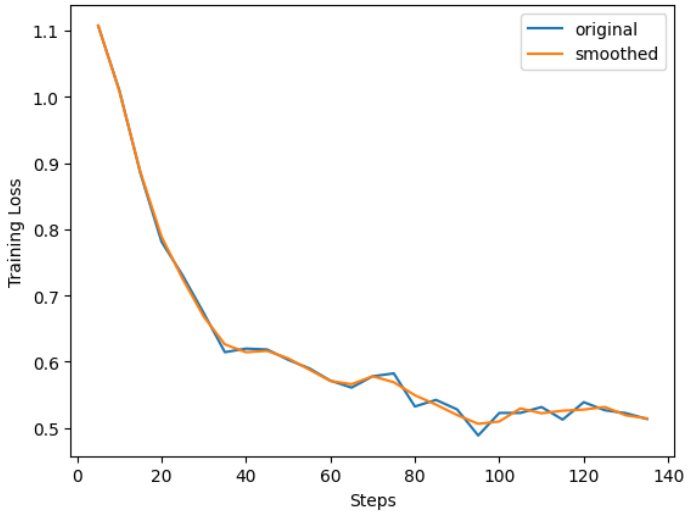
Phi-2-2.7B 模型：:5:45:20/epoch

Phi3-4B-Chat 模型：7:05:04/epoch

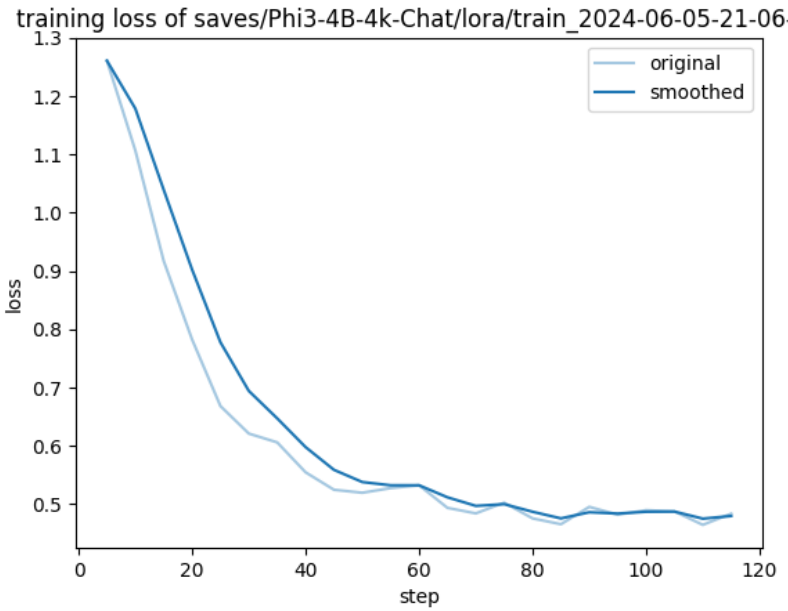
训练曲线



Phi-1.5-1.3B



Phi-2-2.7B模型



Phi3-4B-Chat

测试结果

答案提取

由于发现 Phi 生成的文本大多并不会严格遵循微调时的指令输出格式，所以需要对输出进行特殊的筛选。从生成文本中提取答案时，如果搜索文本内容发现 "Tie" 或者 "both" 则认为预测标签为 0 即平局，发现 "1" 或者 "first" 则认为预测标签为 1，发现 "2" 或者 "second" 则认为预测标签为 2。

Phi-1.5-1.3B 模型

我们发现 Phi-1.5-1.3B 模型的结果完全用不了，有很多为空的结果，只有 12 （1.2%） 的测试样本能够筛选出答案。能输出的结果也是代码、故事之类的，与 prompt 基本上没什么关系。具体 case 在附录中。

Phi-2-2.7B 模型

Phi-2-2.7B 模型效果要好一些，基本上没有为空的输出，有 **710 （71.07%）** 的测试样本能够筛选出答案，输出的结果也看似都稍微正常，但基本上也没有完全遵循我们想要的格式输出。在 PandaLM 测试集上的 accuracy 和 F1 分别为 **27.46%** 和 **33.71%**。

```
Accuracy: 27.46
F1: 33.71
```

Phi-3-4B-Chat 模型

有 **983 （98.40%）** 的测试样本能够筛选出答案。在 PandaLM 测试集上的 accuracy 和 F1 分别为 **49.82%** 和 **38.04%**。

```
Accuracy: 49.82
F1: 38.04
```

结果汇总

模型	accuracy	F1	h/epoch
Phi-1.5-1.3B	/	/	3:36:22
Phi-2-2.7B	27.46%	33.71%	5:45:20
Phi-3-4B-Chat	49.82%	38.04%	7:05:04

case分析

我们微调使用的数据可能与原始模型训练数据在风格、格式或内容上存在较大差异。新数据不具备足够的质量或者不符合模型预期的格式，从而导致模型可能难以适应我们的要求 。Phi系列模型的训练数据主要基于“教科书质量”的数据，通过精心选择和过滤了具有高教育价值和内容质量的数据。这种高质量的数据包括合成数据集，这些数据集被专门设计用于教授模型常识推理和一般知识，包括科学、日常活动和心理理论等。

Phi-1.5-1.3B 模型的大部分输出与指令格式不符，许多生成的结果甚至为空。这可能是由于该模型的参数量较小，导致其在理解和执行复杂指令时能力不足。此外，该模型可能在训练数据的多样性和质量上存在不足，未能覆盖足够多的实际应用场景。Phi-2-2.7B 模型的表现比 Phi-1.5-1.3B 模型稍好，约有71.07%的测试样本能筛选出答案，但输出结果仍不完全符合期望格式。这可能是由于该模型的参数量有所增加，使其具备更强的理解和生成能力，但仍不足以完全满足指令要求。此外，训练数据中 Auto-J 的比例较低（不到1/8），导致模型在处理特定类型的任务时效果不佳。Phi-3-4B-Chat 模型表现最好，98.40%的测试样本能筛选出答案。在 PandaLM 测试集上的 accuracy 和 F1 分别为49.82% 和 38.04%。该模型的参数量更大，并且针对对话场景进行了优化，使其在处理复杂任务时表现更佳。然而，模型在生成输出时仍存在不完全遵循指令的情况，可能需要进一步的微调和优化。

Qwen 系列模型探索

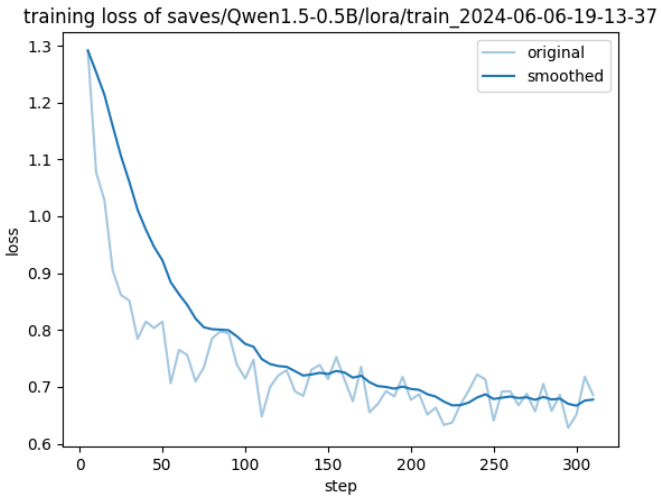
训练时间

Qwen1.5-0.5B: 1:57:41/epoch

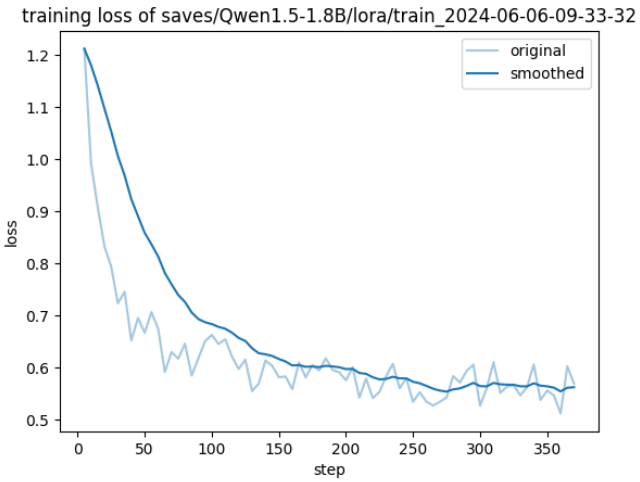
Qwen1.5-1.8B: 3:51:12/epoch

Qwen1.5-1.8B-Chat: 3:57:08/epoch

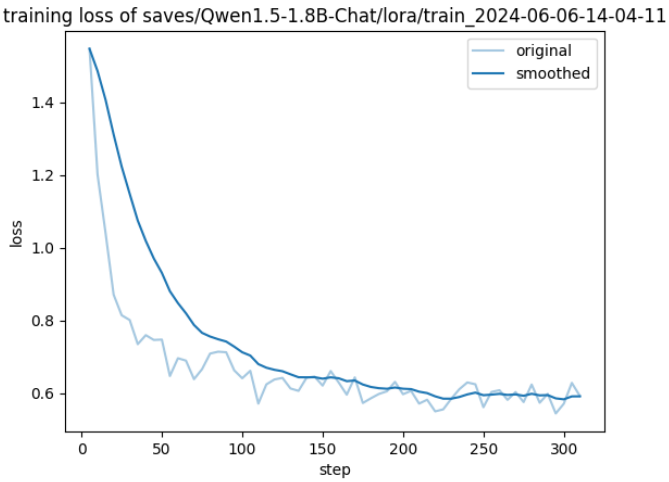
测试曲线



Qwen 1.5-0.5B



Qwen 1.5-1.8B



Qwen1.5-1.8B-Chat

测试结果

Qwen1.5-0.5B 模型

有 **912 （91.2%）** 的测试样本能够筛选出答案。在 PandaLM 测试集上的 accuracy 和 F1 分别为 **33.23%** 和 **30.38%**

```
Accuracy: 33.23
F1: 30.38
```

Qwen1.5-1.8B 模型

有 **999 （100%）** 的测试样本能够筛选出答案。在 PandaLM 测试集上的 accuracy 和 F1 分别为 **51.82%** 和 **45.46%**

```
Accuracy: 51.82
F1: 45.46
```

Qwen1.5-1.8B-Chat

有 **999 （100%）** 的测试样本能够筛选出答案。在 PandaLM 测试集上的 accuracy 和 F1 分别为 **50.22%** 和 **40.16%**

```
Accuracy: 50.22
F1: 40.16
```

结果汇总

模型	accuracy	F1	h/epoch
Qwen1.5-0.5B	33.23%	30.38%	1:57:41
Qwen1.5-1.8B	51.82%	45.46%	3:51:12
Qwen1.5-1.8B-Chat	50.22%	40.16%	3:57:08

case 分析

我们会发现Qwen1.5模型在指令遵循方面效果要好很多，尤其是Qwen1.5-1.8B和Qwen1.5-1.8B-Chat两个模型的输出结果都是有效的，可以提取出有效信息，同样也可以结合Qwen系列模型的训练数据讨论，Qwen 系列模型的预训练数据包括全网文本、百科、书籍、代码、数学及各个领域的垂直数据。训练数据涵盖了各种领域，包括科技、医学、法律、金融、文学等。这样的多样性确保了模型在不同领域任务中的广泛适用性。所以Qwen模型的通用性能力比Phi模型要更好

实验过程总结和感想

XXX：通过合作完成本次自然语言处理课的大作业，我学会了如何在单卡服务器上配合 LLaMA-Factory、vLLM 等工具进行模型微调训练和生成文本，在这个过程中熟悉了 Llama3 模型的很多小细节。在编写各 benchmark 测评脚本代码时，我也阅读一些相关文献， 参考了很多现有 repo 的文档和代码，对 LLM-as-Judge 这一研究方向有了一定的认识。虽然由于资源和时间上的限制，还有很多的改进空间没有来得及探索，但是仅目前为止我认为对我个人的提升就已经很大了。

XXX：通过这次实验，不仅提升了自己在自然语言处理和大型模型训练方面的实际操作能力，还增强了对 LLM 评价机制的理解，在训练 Meta-Llama3-8B-Instruct 模型时，我们选择了 LoRA 微调方法。这使我深入了解了 LoRA 在减少计算资源需求和加快训练速度方面的优势。同时，我学会了如何配

置和调整训练参数，以最大化利用单卡 NVIDIA A100 服务器的计算资源，在处理 Phi 系列模型和 Qwen 模型时，我遇到了许多技术挑战。例如，Phi-1.5-1.3B 模型生成的结果往往不符合预期，需要对输出进行筛选和处理。这使我意识到在实际应用中，模型的稳定性和一致性是非常重要的。

XXX：在本次实验中，我第一次接触了 LLM judge 的相关知识。通过将 GLM-3-turbo 用作 judge，我深入了解了 LLM 在评估其他模型答案质量方面的应用潜力和挑战。特别是，在进行不同 prompt 的尝试中，我认识到精心设计的 prompt 可以显著提升模型的判断能力，使其在复杂任务中表现得更加准确和可靠。这不仅加深了我对提示技术的理解，也让我意识到在处理自然语言处理任务时，prompt设计的细微之处对模型性能有着关键影响。此外，我还体会到在实验过程中不断调整和优化方法的重要性，通过反复试验和改进，才能不断提高模型的表现。总的来说，这次实验不仅强化了我对 LLM 和提示技术的理论知识，也提升了我的实际操作和问题解决能力。

实验分工

XXX：环境配置、数据集下载处理、模型训练、编写各 benchmark 的测试脚本完成测评并分析结果、探究不同形式 prompt 对 MT-Bench 结果的影响、编写实验报告和 PPT、汇报。

XXX：数据集下载处理、模型训练、探究多个 Phi 系列模型和 Qwen 模型的能力、编写实验报告和 PPT。

XXX：数据集下载处理、模型训练、探究不同形式 prompt 对 GLM-3-Turbo 测评结果的影响、编写实验报告和 PPT。

附录

Llama3-8B-Instruct SFT 基础实验训练数据格式

所有训练数据保存在 final_train_data.json 文件中，该 json 文件内容格式如下所示：

```
[
    {
        "instruction": "Below are two responses for a given task. The task is defined by the Instruction with an Input that provides further context. Evaluate the responses and generate a reference answer for the task.\n\n### Instruction:\n......\n\n### Input:\n......\n\n### Response 1:\n......\n\n### Response 2:\n......\n\n### Evaluation:\n",
        "output": "2\n\n### Reason: ......\n\n### Reference: .......\n"
    },
    {
        "instruction": "You are assessing two submitted responses on a given user's query and judging which response is better or they are tied. Here is the data:\n\n[BEGIN DATA]\n***\n[Query]: ......\n***\n[Response 1]:
```

```
......\n***\n[Response 2]: ......\n***\n[END DATA]\n\nHere are the
instructions to assess and compare the two responses:\n\n1. Pinpoint the key
factors to distinguish these two responses.\n2. Conclude your comparison by
providing a final decision on which response is better, or they are tied.
Begin your final decision statement with \"So, the final decision is Response 1
 / Response 2 / Tie\". Ensure that your decision aligns coherently with the
comprehensive evaluation and comparison you've provided.",
        "output": "1. The key factors to distinguish these two
responses:\n\n\t- ......\n\n\t- ......\n\n\t- ......\n\n2. The final
decision:\n\n\tSo, the final decision is......"
    }
    ......
   }
```

Llama3-8B-Instruct SFT 基础实验部分 benchmark 的输入 prompt

PandaLM testset

实际给 vLLM inference 的输入文本列表由如下的字符串模板组成：

```
"""<|begin_of_text|><|start_header_id|>user<|end_header_id|>\n\n{input}
<|eot_id|><|start_header_id|>assistant<|end_header_id|>\n\n"""
```

其中 "{input}" 占位符格式化时由脚本的 build_prompt 函数得到几种类型的替换内容：

```
"Below are two responses for a given task. The task is defined by the
Instruction with an Input that provides further context. Evaluate the
responses and generate a reference answer for the task.\n\n###
Instruction:\n......\n\n### Input:\n......\n\n### Response 1:\n......\n\n###
Response 2:\n......\n\n### Evaluation:\n"
"Below are two responses for a given task. The task is defined by the
Instruction. Evaluate the responses and generate a reference answer for the
task.\n\n### Instruction:\n......\n\n### Response 1:\n......\n\n### Response
2:\n......\n\n### Evaluation:\n"
```

Auto-J testset

实际给 vLLM inference 的输入文本列表由如下的字符串模板组成：

```
<|begin_of_text|><|start_header_id|>user<|end_header_id|>\n\n{input}<|eot_id|>
```

```
<|start_header_id|>assistant<|end_header_id|>\n\n
```

其中 "{input}" 占位符格式化时由 auto-j github repo 中 auto-j/codes/usage/constants_prompt.py 的 build_autoj_input 函数得到 pairwise_tie 类型的替换内容：

```
"""You are assessing two submitted responses on a given user's query and judging which response is better or they are tied. Here is the data:

[BEGIN DATA]
***
[Query]: ......
***
[Response 1]: ......
***
[Response 2]: ......
***
[END DATA]

Here are the instructions to assess and compare the two responses:

1. Pinpoint the key factors to distinguish these two responses.
2. Conclude your comparison by providing a final decision on which response is better, or they are tied. Begin your final decision statement with "So, the final decision is Response 1 / Response 2 / Tie". Ensure that your decision aligns coherently with the comprehensive evaluation and comparison you've provided."""
```

MT-bench multi-turn evaluation

最终采用的实际给 vLLM inference 的输入文本列表由如下格式的字符串组成：

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user questions. You should choose the assistant that follows the user's instructions and answers the user's questions better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. You should focus on who provides a better answer to the second user question. Begin your evaluation by comparing the responses of the two assistants 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.<|eot_id|><|start_header_id|>user<|end_header_id|>

<|The Start of Assistant A's Conversation with User|>

### User:
......

### Assistant A:
......

### User:
......

### Assistant A:
......

<|The End of Assistant A's Conversation with User|>


<|The Start of Assistant B's Conversation with User|>

### User:
......

### Assistant B:
......

### User:
......

### Assistant B:
......

<|The End of Assistant B's Conversation with User|>

Directly 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.<|eot_id|><|start_header_id|>assistant<|end_header_id|>

```

基于 GLM-3-turbo 的不同 prompt 尝试中使用的 prompt

Base 的 system 部分

```
I'll give you User Question, and 2 answers from Assitant A and Assistant B. You need to judge which answer is better. Output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" .
```

Vanilla 的 system 部分

```
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. Select the Assistant A's answer or Assistant B's answer that is better for the given instruction. Do NOT provide any explanation for your choice. 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.
```

CoT 的 system 部分

```
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. Begin your evaluation by comparing the two responses and provide a short explanation. 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.
```

Vanilla* 和 CoT* 中引入的规则

```
Here are some rules of the evaluation:
(1) You should prioritize evaluating whether the answer honestly/precisely/closely executes the instruction, then consider its helpfulness, accuracy, level of detail, harmlessness, etc.
(2) Answers should NOT contain more/less than what the question asks for, as such answers do NOT precisely answer the question.
(3) Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision, as Answer A and Answer B are equally likely to be the better.
(4) Do not allow the length of the responses to influence your evaluation.
(5) Be as objective as possible.
```

Neighbor Set Case

前文提到的 Neighbor 集中的实例 CoT* 和 CoT 的 judge 结果，体现了规则的作用。（翻译由 Claude 3 Sonnet 提供）

```
  // 样本
  {
    "input": "what is brand?",
    "output_1": "A brand is the identity of a specific product, service, or business. It is the combination of a name, term, symbol, design, or other feature that distinguishes one seller's goods or services from those of others. Brands are used in business, industry, and commerce to identify a specific product, service, or business, and to differentiate it from others.",
    "output_2": "Branding refers to the process of creating a unique identity and image for a product, service, or company. This identity is typically communicated through a combination of visual elements, such as logos, colors, and typography, as well as messaging and positioning strategies. The goal of branding is to create a strong, recognizable identity that differentiates a company or product from its competitors and resonates with its target audience. Effective branding can help build trust and loyalty with customers, increase brand awareness, and ultimately drive sales and revenue.",
    "label": 1
  }

  // CoT*的回答
  {"output": "Comparing the two responses, both assistants provide a comprehensive explanation of what a brand is, but they approach the definition from slightly different angles.\n\nAssistant A defines a brand as the identity of a specific product, service, or business, focusing on the distinguishing features and the role of brands in differentiation. This answer is clear, straightforward, and emphasizes the core concept of branding.\n\nAssistant B, on the other hand, defines branding as the process of creating a unique identity and image, encompassing not only the visual elements but also the messaging and positioning strategies. Assistant B's response provides a broader perspective on branding and highlights the strategic goals and customer engagement aspects of branding.\n\nBoth answers are accurate and helpful, but given the rules of evaluation, the emphasis should be on the precision and execution of the instruction. Assistant A's answer directly addresses the user's question (\"what is brand?\") with a concise definition, while Assistant B's answer answers the question of \"what is branding?\" by explaining the process. \n\nThe question asks for the definition of \"brand,\" not \"branding,\" which is a broader concept that includes the process. Therefore, Assistant A's response is more directly aligned with the user's query.\n\nFinal Verdict: [[A]]"}

  // CoT的回答
  {"output": "Comparing the two responses, both assistants provide a comprehensive explanation of what a brand is, but they approach the definition
```

from slightly different angles. Assistant A focuses more on the end result of branding, defining a brand as the identity that distinguishes a product or service, while Assistant B emphasizes the process of branding and the strategic elements involved in creating a unique identity.\n\nAssistant B's response is longer and provides additional context on the strategic aspects of branding, such as messaging and positioning strategies, which might be helpful for someone seeking a deeper understanding of branding as a whole. Assistant A's response is more concise and to the point, which could be preferable for someone looking for a simple and clear definition.\n\nBoth assistants cover the core concept thoroughly, but given that Assistant B offers more insights into the strategic side of branding, their response could be considered more valuable to a user seeking a comprehensive answer. However, this does not necessarily mean that Assistant B is objectively \"better\" in all contexts, as the value of the responses may vary depending on the user's specific needs and preferences.\n\nFinal Verdict: [[B]]"}

AH 翻译：Comparing the two responses, both assistants provide a comprehensive explanation of what a brand is, but they approach the definition from slightly different angles.\n\nAssistant A defines a brand as the identity of a specific product, service, or business, focusing on the distinguishing features and the role of brands in differentiation. This answer is clear, straightforward, and emphasizes the core concept of branding.\n\nAssistant B, on the other hand, defines branding as the process of creating a unique identity and image, encompassing not only the visual elements but also the messaging and positioning strategies. Assistant B's response provides a broader perspective on branding and highlights the strategic goals and customer engagement aspects of branding.\n\nBoth answers are accurate and helpful, but given the rules of evaluation, the emphasis should be on the precision and execution of the instruction. Assistant A's answer directly addresses the user's question (\"what is brand?\") with a concise definition, while Assistant B's answer answers the question of \"what is branding?\" by explaining the process. \n\nThe question asks for the definition of \"brand,\" not \"branding,\" which is a broader concept that includes the process. Therefore, Assistant A's response is more directly aligned with the user's query.\n\nFinal Verdict: [[A]]

比较这两个回答,两位助手都对品牌的定义进行了全面的解释,但它们从略有不同的角度来阐述定义。

助手A将品牌定义为特定产品、服务或企业的身份,重点是区别特征和品牌在差异化中的作用。这个答案清晰、直接,强调了品牌的核心概念。

另一方面,助手B将品牌定义为创建独特身份和形象的过程,不仅包括视觉元素,还包括信息传递和定位策略。助手B的回答提供了一个更广泛的品牌视角,强调了品牌的战略目标和客户参与方面。

两个答案都是准确和有帮助的,但根据评估规则,重点应该是指令的精确性和执行情况。助手A的答案直接回答了用户的问题("什么是品牌?")并给出了一个简洁的定义,而助手B的答案通过解释过程回答了"什么是品牌?"的问题。

这个问题要求给出"品牌"的定义,而不是"品牌化",后者是一个更广泛的概念,包括品牌化的过程。因此,助手A的回答更直接地符合用户的查询。

最终结论:[[A]]

AH 翻译：Comparing the two responses, both assistants provide a comprehensive explanation of what a brand is, but they approach the definition from slightly different angles. Assistant A focuses more on the end result of branding, defining a brand as the identity that distinguishes a product or service, while Assistant B emphasizes the process of branding and the strategic elements involved in creating a unique identity.\n\nAssistant B's response is longer and provides additional context on the strategic aspects of branding, such as messaging and positioning strategies, which might be helpful for someone seeking a deeper understanding of branding as a whole. Assistant A's response is more concise and to the point, which could be preferable for someone looking for a simple and clear definition.\n\nBoth assistants cover the core concept thoroughly, but given that Assistant B offers more insights into the strategic side of branding, their response could be considered more valuable to a user seeking a comprehensive answer. However, this does not necessarily mean that Assistant B is objectively \"better\" in all contexts, as the value of the responses may vary depending on the user's specific needs and preferences.\n\nFinal Verdict: [[B]]

比较这两个回答,两位助手都对品牌的定义进行了全面的解释,但它们从略有不同的角度来阐述定义。助手A更侧重于品牌的最终结果,将品牌定义为区别产品或服务的身份,而助手B则强调品牌化的过程以及创建独特身份所涉及的战略要素。

助手B的回答更长,提供了有关品牌战略方面的额外背景信息,如信息传递和定位策略,这可能对寻求更深入理解整个品牌化过程的人有帮助。助手A的回答更简洁、更切中要点,这可能更适合寻找简单明了定义的人。

两位助手都全面地涵盖了核心概念,但鉴于助手B提供了更多关于品牌战略方面的见解,他们的回答可能被认为对寻求全面答案的用户更有价值。然而,这并不一定意味着助手B在所有情况下都客观地"更好",因为回答的价值可能会根据用户的具体需求和偏好而有所不同。

最终结论:[[B]]

Phi 系列模型样本 Case

Phi-1.5-1.3B case

```
1 <|im_start|>user
2 Below are two responses for a given task. The task is defined by the
  Instruction with an Input that provides further context. Evaluate the
  responses and generate a reference answer for the task.
3
4 ### Instruction:
```

```
The sentence you are given might be too wordy, complicated, or unclear.
Rewrite the sentence and make your writing clearer by keeping it concise.
Whenever possible, break complex sentences into multiple sentences and
eliminate unnecessary words.

### Input:
If you have any questions about my rate or if you find it necessary to
increase or decrease the scope for this project, please let me know.

### Response 1:
If you have any questions about my rate, please let me know.

### Response 2:
If you have any questions, please let me know.

### Evaluation:
<|im_end|>
<|im_start|>assistant




Once upon a time, in a small town called Maplewood, there lived a young girl
named Lily. Lily was a curious and intelligent girl who loved to learn about
the world around her. She had a particular interest in social studies and was
always eager to explore new topics.

One day, Lily's teacher, Mrs. Johnson, announced that the class would be
studying globes and their features and symbols. Lily's eyes lit up with
excitement as she had always been fascinated by the Earth and its various
elements. She couldn't wait to dive into this new topic.

As the days went by, Lily and her classmates learned about the different
continents, oceans, and countries that make up our planet. They also
discovered the importance of symbols on globes, such as the equator and the
prime meridian. Lily found it fascinating how these symbols represented
different aspects of the Earth's geography.

One afternoon, Lily's friend, Emily, came over to her house to study together.
Emily was also interested in social studies and they often discussed various
topics. As they sat at Lily's desk, surrounded by textbooks and globes, they
began to talk about their favorite subjects.

"I really enjoy learning about different cultures and traditions," Lily said.
"It's amazing how diverse our world is."

```

Emily nodded in agreement. "Yes, it's so interesting to learn about the customs and beliefs of people from different parts of the world. It helps us understand and appreciate their way of life."

Lily smiled and continued, "You know, I think it's important to have a good understanding of social studies because it helps us become more informed global citizens. We can learn from the past and make better decisions for the future."

Emily nodded again, impressed by Lily's insight. "You're absolutely right, Lily. Social studies teaches us about history, geography, and even economics. It's a subject that helps us make sense of the world we live in."

As they delved deeper into their studies, Lily and Emily realized that they had a lot in common. They both loved learning about different cultures, exploring new places, and understanding the world around them. They decided to form a study group with a few other classmates who shared their passion for social studies.

The study group met every week, discussing various topics related to social studies. They would bring in their globes and point out different features and symbols, just like they had learned in class. They would also share interesting facts and stories they had discovered during their research.

One day, as they were studying together, Lily's younger brother, Jake, walked into the room. He was fascinated by what he saw and wanted to join in on the fun. Lily and Emily welcomed him with open arms, eager to share their knowledge with him.

"Jake, did you know that the Earth is divided into hemispheres?" Lily asked, pointing to the globe on the table.

Jake shook his head, his eyes wide with curiosity. "No, I didn't. How does that work?"

Lily explained, "Well, the Earth is divided into two hemispheres, the Northern Hemisphere and the Southern Hemisphere. The equator is an imaginary line that runs around the middle of the Earth, dividing it into the Northern Hemisphere and the Southern Hemisphere."

Jake listened intently, absorbing every word. He was fascinated by the idea of dividing the Earth into different parts.

As the study group continued their discussions, Lily, Emily, and Jake realized that they were not only learning about social studies but also developing important skills. They were improving their critical thinking, problem-

```
solving, and communication skills. They were also learning how to work together as a team and support each other's learning.

One day, as they were preparing for a social studies quiz, Lily's mom called them for dinner. They quickly packed up their study materials and headed to the dining table.

During dinner, Lily's mom asked her about her day at school. Lily excitedly shared everything she had learned about globes and their features and symbols. Her mom listened attentively, impressed by her daughter's enthusiasm for learning.

After dinner, Lily and her family gathered in the living room. Lily's dad, who was a history enthusiast, started telling them about the ancient civilizations that had used globes to navigate the seas. He explained how these early globes were not as accurate as the ones we have today, but they were still valuable tools for explorers and sailors.

Lily's dad then asked her if she knew about the different symbols on globes. Lily confidently answered, "Yes, I do! Symbols like the equator and the prime meridian help us understand the Earth's geography and time zones."

Her dad smiled proudly and said, "That's right, Lily! You're becoming quite the expert in this subject."

Lily beamed with pride, feeling a
<|im_start|>user
Below are two responses for a given task. The task is defined by the Instruction with an Input that provides further context. Evaluate the responses and generate a reference answer for the task.

### Instruction:
The sentence you are given might be too wordy, complicated, or unclear. Rewrite the sentence and make your writing clearer by keeping it concise. Whenever possible, break complex sentences into multiple sentences and eliminate unnecessary words.

### Input:
If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Response 1:
If you have any questions about my rate, please let me know.

### Response 2:
If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.
```

```
80
81 ### Evaluation:
82 <|im_end|>
83 <|im_start|>assistant
84
85
86 # Python script for 1.1.9: How to use the Python debugger to step through code
   and identify errors
87
88 # Introduction:
89 # The Python debugger is a powerful tool that allows you to step through your
   code line by line, inspect variables, and identify errors.
90 # It is especially useful when you are debugging complex programs or when you
   want to understand how your code works.
91
92 # 1. How to use the Python debugger to step through code and identify errors?
93 # The Python debugger is a built-in module that you can use to step through
   your code and identify errors.
94 # To use the debugger, you need to import it into your Python script:
95
96 import pdb
97
98 # 2. How to set a breakpoint in your code?
99 # A breakpoint is a point in your code where the debugger will stop execution
   and allow you to inspect the state of your program.
100 # To set a breakpoint, you can use the 'pdb.set_trace()' function:
101
102 def divide(x, y):
103     pdb.set_trace()
104     return x / y
105
106 # 3. How to step through your code line by line?
107 # When you set a breakpoint in your code, the debugger will stop execution at
   that point and allow you to step through your code line by line.
108 # To step through your code, you can use the 'n' (next) command:
109
110 def multiply(x, y):
111     pdb.set_trace()
112     return x * y
113
114 # 4. How to inspect variables and their values?
115 # The Python debugger allows you to inspect variables and their values at any
   point in your code.
116 # To inspect a variable, you can use the 'p' (print) command:
117
118 def add(x, y):
119     pdb.set_trace()
```

```
120     return x + y
121
122 # 5. How to continue execution of your code after a breakpoint?
123 # Once you have set a breakpoint in your code, you can continue execution of
    your code by pressing 'c' (continue) or 'q' (quit) in the debugger.
124 # This will continue execution of your code until the next breakpoint is
    reached.
125
126 # Python script for 1.1.10: How to use the Python debugger to handle
    exceptions and errors
127
128 # Introduction:
129 # When writing code, it is inevitable that errors and exceptions will occur.
    The Python debugger can help you handle these errors and exceptions in a more
    efficient and effective way.
130
131 # 1. How to handle exceptions and errors in Python?
132 # In Python, exceptions are raised when an error occurs during the execution of
     a program. These exceptions can be handled using the try-except block.
133
134 # Example:
135 try:
136     x = int(input("Enter a number: "))
137     y = 10 / x
138 except ZeroDivisionError:
139     print("Cannot divide by zero")
140
141 # Explanation:
142 # In this example, we are trying to divide a number by zero, which is not
    allowed in mathematics. The ZeroDivisionError exception is raised when we try
    to divide by zero. We handle this exception using the try-except block.
143
144 # 2. How to handle multiple exceptions in Python?
145 # We can handle multiple exceptions in the same try-except block.
146
147 # Example:
148 try:
149     x = int(input("Enter a number: "))
150     y = 10 / x
151 except (ValueError, ZeroDivisionError):
152     print("Invalid input or division by zero")
153
154 # Explanation:
155 # In this example, we are trying to divide a number by zero, which is not
    allowed in mathematics. We also handle the ValueError exception, which is
    raised when the input is not a valid integer.
156
```

```
# 3. How to handle exceptions in a specific function?
# We can handle exceptions in a specific function using the try-except block.

# Example:
def divide(x, y):
    try:
        result = x / y
    except ZeroDivisionError:
        print("Cannot divide by zero")
    else:
        return result

# Explanation:
# In this example, we are trying to divide two numbers. If the second number is zero, a ZeroDivisionError exception is raised. We handle this exception using the try-except block. If no exception is raised, we return the result.

# 4. How to handle exceptions in a specific class?
# We can handle exceptions in a specific class using the __try__ and __except__ methods.

# Example:
class MyClass:
    def __init__(self, x):
        self.x = x

    def divide(self,
<|im_start|>user
Below are two responses for a given task. The task is defined by the Instruction. Evaluate the responses and generate a reference answer for the task.

### Instruction:
Come up with some search queries on google about coding stuff.

### Response 1:
1. What is coding?
2. How to code?
3. How to write code?
4. How to debug code?
5. How to debug code?
6. How to debug code?
7. How to debug code?
8. How to debug code?
9. How to debug code?

### Response 2:
```

1. "how to code"
2. "best coding language"
3. "best coding tutorials"
4. "best coding resources"

Evaluation:
<|im_end|>
<|im_start|>assistant

Reference:
1. What is coding?
2. How to code?
3. How to write code?
4. How to debug code?
5. How to debug code?
6. How to debug code?
7. How to debug code?
8. How to debug code?
9. How to debug code?

Once upon a time, in a small town called Maplewood, there lived a young girl named Lily. Lily was a curious and intelligent girl who loved to learn new things. She had a passion for language arts and was always eager to expand her vocabulary.

One day, Lily's teacher, Mrs. Johnson, announced that the class would be learning about multiple-meaning words and idioms. Lily was excited about this topic because she knew it would help her understand the different meanings of words and how to use them in context.

As the lesson began, Mrs. Johnson explained that multiple-meaning words are words that have more than one meaning. She gave examples such as "bank," which can mean a financial institution or the side of a river. Lily found this concept fascinating and couldn't wait to explore more words with multiple meanings.

To make the lesson more engaging, Mrs. Johnson decided to use a hands-on activity. She handed out a set of cards to each student, each containing a multiple-meaning word. The students were instructed to read the word on the card and come up with two different meanings for it.

Lily eagerly picked up her card and read the word "bank." She thought for a moment and then raised her hand to share her answer. "The word 'bank' can mean a financial institution or the side of a river," she said confidently.

Mrs. Johnson smiled and nodded. "That's correct, Lily! You've understood the concept perfectly. Now, let's move on to idioms."

Idioms are phrases that have a figurative meaning different from the literal meaning of the words. They often use words or phrases that are not commonly known. Mrs. Johnson explained that idioms can be tricky to understand because their meanings cannot be deduced from the individual words.

To illustrate this, Mrs. Johnson told a story about a boy named Jack who loved to play basketball. One day, Jack's friend asked him, "Why do you always shoot the ball into the hoop?" Jack replied, "Because I'm trying to make a basket."

The friend looked puzzled and said, "But you're not trying to make a basket, you're trying to make a hole in the net!" Jack was confused and asked, "What do you mean?"

The friend explained, "When you shoot the ball, you're not trying to make a hole in the net. You're trying to make a basket. It's an idiom that means you're trying to score a point."

Lily listened intently to the story and realized that idioms can be confusing because they don't always make sense based on the individual words. She made a mental note to pay attention to idioms in her daily life.

As the lesson continued, Mrs. Johnson introduced the concept of context clues. She explained that context clues are the words or phrases surrounding an unfamiliar word that can help us understand its meaning.

To demonstrate this, Mrs. Johnson gave the class a sentence with an unknown word. She asked the students to identify the meaning of the word based on the context.

The sentence was, "The detective carefully examined the crime scene for any clues."

Lily thought for a moment and then raised her hand. "The word 'examined' means to look closely at something," she said confidently.

Mrs. Johnson smiled and nodded. "That's correct, Lily! You've used the context clues to determine the meaning of the word. Great job!"

Lily felt proud of herself for understanding the concept of context clues. She realized that by paying attention to the words and phrases around an unfamiliar word, she could figure out its meaning.

```
As the class came to an end, Lily felt a sense of accomplishment. She had learned so much about multiple-meaning words, idioms, and context clues. She knew that these skills would help her become a better communicator and understand the world around her.

From that day forward, Lily continued to explore the world of language arts. She read books, wrote stories, and practiced using new words and phrases. She knew that by expanding her vocabulary, she could express herself more effectively and connect with others.

Lily's journey in language arts was just beginning, and she couldn't wait to see where it would take her. She knew that with her newfound knowledge, she would be able to navigate the world of words with confidence and understanding.

And so, Lily's love for language arts grew stronger with each passing
<|im_start|>user
Below are two responses for a given task. The task is defined by the Instruction. Evaluate the responses and generate a reference answer for the task.

### Instruction:
Come up with some search queries on google about coding stuff.

### Response 1:
1. Coding tutorials
2. Coding languages
3. Coding languages for beginners
4. Coding languages for beginners
5. Coding languages for beginners
6. Coding languages for beginners
7. Coding languages for beginners
8. Coding languages for beginners
9. Coding languages for beginners
10. Coding languages for beginners
11. Coding languages for beginners
12. Coding languages for beginners
13. Coding languages for beginners

### Response 2:
1. "how to code"
2. "best coding language"
3. "best coding tutorials"
4. "best coding resources"

### Evaluation:
<|im_end|>
<|im_start|>assistant
```

Once upon a time, in a small town called Greenville, there lived a young girl named Lily. Lily was a curious and intelligent girl who loved learning about the world around her. She had a particular interest in science, especially life science and the fascinating world of plants and animals.

One sunny day, Lily's teacher, Mrs. Johnson, announced that the class would be going on a field trip to the local botanical garden. Lily was thrilled! She had always dreamed of seeing different types of plants and learning about their unique characteristics.

The day of the field trip arrived, and Lily and her classmates hopped onto the school bus. As they drove through the town, Lily couldn't help but notice the beautiful houses and gardens along the way. She wondered how plants and animals interacted with their environment and how they adapted to survive.

Finally, they arrived at the botanical garden, a vast green oasis filled with colorful flowers, towering trees, and a variety of plants and animals. Mrs. Johnson led the class to a section dedicated to different types of flowers.

Lily's eyes widened with excitement as she saw the vibrant petals and delicate fragrance of the flowers. She approached a gardener named Mr. Anderson and asked him about the different types of flowers. Mr. Anderson explained that flowers are the reproductive organs of plants and come in various shapes, sizes, and colors.

Lily was fascinated by this information and wanted to learn more. She asked Mr. Anderson if he could show her how to take care of flowers. Mr. Anderson smiled and agreed. He took Lily to a small greenhouse where he showed her how to plant seeds and water the plants.

As Lily carefully tended to the flowers, she noticed a small bird perched on a nearby branch. It was a beautiful blue jay, and Lily couldn't help but feel a sense of wonder and awe. She asked Mr. Anderson about the bird and its habitat.

Mr. Anderson explained that birds are animals that have feathers, beaks, and lay eggs. They build nests to lay their eggs and raise their young. He also told Lily that birds play an important role in pollination, helping plants reproduce by carrying pollen from one flower to another.

Lily was amazed by the interconnectedness of plants and animals. She realized that every living thing had a purpose and played a vital role in maintaining the balance of nature.

After spending the day at the botanical garden, Lily returned to school with a newfound appreciation for the natural world. She couldn't wait to share her experiences with her family and friends.

Back at home, Lily's parents noticed her excitement and asked her about her day. Lily eagerly told them about the different types of flowers, the bird she saw, and the importance of plants and animals in our lives.

Her parents listened attentively and encouraged her curiosity. They explained that science was all about asking questions, making observations, and finding answers. They also told Lily that she could pursue a career in science, just like her parents did.

Lily's eyes sparkled with excitement as she imagined herself as a scientist, exploring the wonders of the natural world. She knew that her love for science and her desire to learn more would guide her on a lifelong journey of discovery.

And so, Lily continued to nurture her curiosity and passion for science. She knew that by understanding the world around her, she could make a difference and contribute to the betterment of society.

Real-world usecases:

1. Conversation: Lily and her friend Emma are walking in a park. They notice a squirrel gathering nuts and wonder why it does that. They start a conversation about the importance of storing food for the winter and how squirrels play a role in seed dispersal.

2. Story: Lily's class is learning about different types of trees. They decide to go on a nature walk and observe the trees in their neighborhood. They take pictures and create a scrapbook of their findings, including the different types of trees they saw and their unique characteristics.

3. Debate: Lily's class is divided into two groups for a debate on the topic of deforestation. One group argues that deforestation is necessary for economic development, while the other group argues that it is harmful to the environment. Each group presents their arguments, supported by facts and examples, and engages in a respectful debate.

4. Exercise: Look around your backyard or a nearby park. Identify three different types of plants and describe their unique characteristics.

5. Exercise: Draw a picture of a flower and label its different parts, such as the petals, stamen, and pistil.

```
6. Exercise: Research and write a short paragraph about a specific animal and its habitat. Include information about its diet, behavior, and adaptations.

7. Exercise: Create a poster showcasing the importance of plants and animals in our daily lives. Include pictures and captions to illustrate your

```

Phi-2-2.7B case

```
999
<|im_start|>user
Below are two responses for a given task. The task is defined by the Instruction with an Input that provides further context. Evaluate the responses and generate a reference answer for the task.

### Instruction:
The sentence you are given might be too wordy, complicated, or unclear. Rewrite the sentence and make your writing clearer by keeping it concise. Whenever possible, break complex sentences into multiple sentences and eliminate unnecessary words.

### Input:
If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Response 1:
If you have any questions about my rate, please let me know.

### Response 2:
If you have any questions, please let me know.

### Evaluation:
<|im_end|>
<|im_start|>assistant


### Reason: Response 2 is better because it is more concise and clear.

### Reference: If you have any questions, please let me know.

<|im_start|>user
Below are two responses for a given task. The task is defined by the Instruction with an Input that provides further context. Evaluate the responses and generate a reference answer for the task.

```

```
### Instruction:
The sentence you are given might be too wordy, complicated, or unclear. Rewrite the sentence and make your writing clearer by keeping it concise. Whenever possible, break complex sentences into multiple sentences and eliminate unnecessary words.

### Input:
If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Response 1:
If you have any questions about my rate, please let me know.

### Response 2:
If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Evaluation:
<|im_end|>
<|im_start|>assistant


### Reason: Both responses are clear and concise, but Response 2 is more accurate as it includes the phrase "if you find it necessary to increase or decrease the scope for this project" which was not in the original sentence.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

<|im_start|>user
Below are two responses for a given task. The task is defined by the Instruction. Evaluate the responses and generate a reference answer for the task.

### Instruction:
Come up with some search queries on google about coding stuff.

### Response 1:
1. What is coding?
2. How to code?
3. How to write code?
4. How to debug code?
5. How to debug code?
6. How to debug code?
7. How to debug code?
8. How to debug code?
```

```
9. How to debug code?

### Response 2:
1. "how to code"
2. "best coding language"
3. "best coding tutorials"
4. "best coding resources"

### Evaluation:
<|im_end|>
<|im_start|>assistant

1. Response 2

### Reason: Response 2 provides more specific and relevant search queries that are likely to yield more useful and informative results.

### Reference: 1. "How to code in Python" 2. "Best coding languages for beginners" 3. "Top coding tutorials for beginners" 4. "Best coding resources for beginners"

<|im_start|>user
Below are two responses for a given task. The task is defined by the Instruction. Evaluate the responses and generate a reference answer for the task.

### Instruction:
Come up with some search queries on google about coding stuff.

### Response 1:
1. Coding tutorials
2. Coding languages
3. Coding languages for beginners
4. Coding languages for beginners
5. Coding languages for beginners
6. Coding languages for beginners
7. Coding languages for beginners
8. Coding languages for beginners
9. Coding languages for beginners
10. Coding languages for beginners
11. Coding languages for beginners
12. Coding languages for beginners
13. Coding languages for beginners

### Response 2:
1. "how to code"
2. "best coding language"
```

```
3. "best coding tutorials"
4. "best coding resources"

### Evaluation:
<|im_end|>
<|im_start|>assistant

The second response is better because it provides more specific and relevant search queries.

### Reference: 1. "how to code" 2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"


```

Phi-3-4B-Chat case

```
999
 2

### Reason: Response 2 is more concise and clear than Response 1.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 1 is more concise and clear than the reference.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 2 is more concise and clear than the reference.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 1 is more concise and clear than the reference.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 2 is more concise and clear than the reference.
```

```

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 1 is more concise and clear than the reference.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 2 is more concise and clear than the reference.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 1 is more concise and clear than the reference.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 2 is more concise and clear than the reference.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 1 is more concise and clear than the reference.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 2 is more concise and clear than the reference.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 1 is more concise and clear than the reference.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

```

```
### Reason: Response 2 is more concise and clear than the reference.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 1 is more concise and clear than the reference.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 2 is more concise and clear than the reference.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 1 is more concise and clear than the reference.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 2 is more concise and clear than the reference.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 1 is more concise and clear than the reference.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Reason: Response 2 is more concise and clear than the reference.

### Reference: If you have any questions
 1

### Reason: Response 1 is better because it is more concise and to the point.

### Reference: If you have any questions about my rate, please let me know.

 2

```

```
### Reason: Response 2 is better because it provides more specific and useful search queries for finding coding resources.

### Reference: "best coding tutorials", "best coding resources", "how to code", "best coding language"

 2

### Reason: Response 2 is better because it provides more diverse and specific search queries that can help in finding relevant information about coding.

### Reference: "best coding tutorials", "how to code", "best coding language", "coding resources"


```

Qwen 系列模型样本Case

Qwen1.5-0.5B case

```
<|start_header_id|>system<|end_header_id|>

You are a helpful assistant.<|eot_id|><|start_header_id|>user<|end_header_id|>

Below are two responses for a given task. The task is defined by the Instruction with an Input that provides further context. Evaluate the responses and generate a reference answer for the task.

### Instruction:
The sentence you are given might be too wordy, complicated, or unclear. Rewrite the sentence and make your writing clearer by keeping it concise. Whenever possible, break complex sentences into multiple sentences and eliminate unnecessary words.

### Input:
If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Response 1:
If you have any questions about my rate, please let me know.

### Response 2:
```

```
If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.

### Evaluation:
<|eot_id|><|start_header_id|>assistant<|end_header_id|>


Both responses are good, but Response 2 is better because it is more concise and clear. Response 1 is also good, but it is repetitive and lacks the clarity of Response 2.

### Reference: If you have any questions about my rate or if you find it necessary to increase or decrease the scope for this project, please let me know.


### Explanation: Response 2 is better because it is more concise and clear, while Response 1 is repetitive and lacks the clarity of the original sentence.

### Reason: Response 2 is better because it is more concise and clear, while Response 1 is repetitive and lacks the clarity of the original sentence.

```

Qwen1.5-1.8B case

```
999
<|start_header_id|>system<|end_header_id|>

You are a helpful assistant.<|eot_id|><|start_header_id|>user<|end_header_id|>

Below are two responses for a given task. The task is defined by the Instruction. Evaluate the responses and generate a reference answer for the task.

### Instruction:
Come up with some search queries on google about coding stuff.

### Response 1:
1. What is coding?
2. How to code?
3. How to write code?
4. How to debug code?
5. How to debug code?
6. How to debug code?
7. How to debug code?
```

```
8. How to debug code?
9. How to debug code?

### Response 2:
1. "how to code"
2. "best coding language"
3. "best coding tutorials"
4. "best coding resources"

### Evaluation:
<|eot_id|><|start_header_id|>assistant<|end_header_id|>


Tie

### Reason: Both responses are good and provide relevant search queries.

### Reference: 1. "how to code" 2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
```

```
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "best coding language"
Tie

### Reason: Both responses provide relevant search queries for coding-related topics.

### Reference: 1. "how to code" 2. "best coding language" 3. "best coding tutorials" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
```

```
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code" 3. "best coding language" 4. "best coding resources"
2. "how to code
```

Qwen1.5-1.8B-Chat case

```
<|start_header_id|>system<|end_header_id|>

You are a helpful assistant.<|eot_id|><|start_header_id|>user<|end_header_id|>

Below are two responses for a given task. The task is defined by the
Instruction with an Input that provides further context. Evaluate the
responses and generate a reference answer for the task.

### Instruction:
The sentence you are given might be too wordy, complicated, or unclear.
Rewrite the sentence and make your writing clearer by keeping it concise.
Whenever possible, break complex sentences into multiple sentences and
eliminate unnecessary words.

### Input:
If you have any questions about my rate or if you find it necessary to
increase or decrease the scope for this project, please let me know.

### Response 1:
If you have any questions about my rate, please let me know.

### Response 2:
If you have any questions, please let me know.
```

```
18
19 ### Evaluation:
20 <|eot_id|><|start_header_id|>assistant<|end_header_id|>
21
22
23 ### Reason: Response 1 is better because it is more concise and clear.
24
25 ### Reference: If you have any questions, please let me know.
```