# CCKS-TCMBench中医知识理解与推理能力评测

**会议地点：待定**

**会议时间：9月19日—22日**

# 任务描述

随着大型语言模型（Large Language Model, LLM）如ChatGPT、GPT-4等在自然语言处理（NLP）领域的突破性进展，其展现的类似通用人工智能（AGI）的能力为中医的传承与创新提供了新的可能性。中医作为中国传统医学的重要组成部分，具有深厚的文化底蕴和独特的诊疗特性。与西医循证医学不同，中医在理论基础、诊断方法、治疗手段、防止观念以及整体观念等方面，与西医存在显著差异。同时在中医领域，语言表述往往富含哲理和象征意义，这对语言模型提出了更高的要求。因此直接将已有的西医评估基准用于中医评估无法全面评估语言模型在中医知识方面的潜力和实际效用。然而，在NLP社区中，尚未有一个标准的中医评测基准。

为推动LLM在中医领域的发展和落地，华东师范大学王晓玲教授团队联合复旦大学周雅倩、上海中医药大学周毅萍等专家学者，推出**TCMBench**评测基准。**TCMBench**评测基准依托于中医执业医师资格考试的丰富题库，全面覆盖三大考试范围，包括中医基础理论、中医临床医学，以及西医与临床医学的综合内容以及16个核心考试科目，共计9,788道真题和5,473道练习题。**TCMBench**评测基准旨在深度评估和精准测量LLM模型对中医知识的掌握水平，以及模型在中医情境下的解释和推理能力。作为首个专注于中医医疗场景的LLM评测基准，**TCMBench**不仅为开源社区提供了一个高效的评测工具，也助力业界快速评估自有LLM模型在中医领域的应用潜力。**TCMBench**将作为CCKS-2024的评测任务之一，依托于天池大赛平台进行评测。

为了促进语言模型在中医领域的知识理解和推理能力的提升，我们将对**TCMBench**评测制定两个核心子任务：

1. 中医知识理解评测：基于给定的问题，选择正确的候选答案并给出相应的解析。
2. 中医语义推理评测：基于给定的前提和假设，判断两者之间的关系：a) 蕴含；b) 非蕴含（矛盾或无关）。

同时我们将对TCMBench评测开放两个榜单：

1. **提示学习（Prompt Learning）赛道**，参赛队伍将使用组织方指定的开源大型语言模型（LLM）主干直接进行预测。选手可以通过以下方式来增强模型的表现：
2. 提示模版：通过设计合适的提示模版来引导模型输出。
3. 示例选择（Demonstrations）：从训练集中选择或增加示例，以提供更多的上下文信息（可微调额外的模型作为辅助选择，不超过1.5亿参数）。
4. **高效参数微调赛道**，对于选择使用大型预训练模型作为基座的参赛队伍，在给定的开源大模型主干之上进行参数高效微调。对于第一个任务，选手可以使用任何开源资源构造训练集进行模型参数高效微调，允许采用指令微调（SFT）或持续学习（Continual Learning）策略，需要遵循以下要求：
5. 禁止将验证集和测试集中的数据加入微调数据中。
6. 不得对大模型主干进行微调。
7. 参数高效微调模块参数量不得超过大模型主干参数量的1%。
8. 单一微调模块：整个参赛过程中只能使用一组参数高效微调模块来完成所有任务。

本CCKS-TCMBench评测任务在知识覆盖和任务设置上进行了完备的制定：

1. 知识覆盖面广：CCKS-TCMBench测试集精心挑选了能够充分反映和代表中医理论知识及实践技能水平的题目，覆盖了中医执业医师资格考试的中药学、方剂学、中医基础理论、中医诊断学、中医经典、中医内科学、中医外科学、中医儿科学、针灸学、中医妇科学、诊断学基础、内科学、传染病学、医学伦理学、卫生法规和其他16个中医科目，考察模型对中医知识的全方位理解。
2. 评测方式多样：在第一个任务中，通过给定不同的系统提示，针对中医执业医师资格考试的三种题型进行评测。这不仅考察模型的任务理解能力，还测试其在回答格式、生成解析和上下文处理方面的表现，从而评估模型的泛化性和鲁棒性。在第二个任务中，CCKS-TCMBench采用通过同时考察前提与假设之间的正向推理和逆向推理的逻辑关系，确保评测结果的稳定性。
3. 赛道设置：本次CCKS-TCMBench分为提示学习赛道和高效参数微调赛道，考察提示学习、有监督微调和指令数据质量控制等不同的NLP技术。评测完全依托于开源大模型社区，考察选手对Prompt优化以及模型预训练技术的创新探索和应用能力。

**评测地址**：TBD

## 任务形式：

#### 中医知识理解评测任务：

我们从网络上收集的原始题库中通过人工筛选并结合中医专家的意见，确保TCM-Bench涵盖中医执业医师资格考试的所有问题类型和科目。我们首先将PDF格式的文字版习题进行数据清洗，并基于自制的规则抽取出问题和选项（question）、标准答案（answer）、科目（knowledge\_point）和标准解析信息（analysis），转换为Json结构化的形式，共包含3类题型：

表1. 三类题型组成说明

|  |  |  |
| --- | --- | --- |
| **测试题类型** | **数据结构** | **说明** |
| **单题干单项最佳选择题**（A1/A2） | **问题（题干） + 选项** | 每道试题由一个问题（题干）与A、B、C、D、E 五个备选答案组成，有且仅有一个正确答案。型题采用为A1型题、A2型题2种题型。**单句型最佳选择题（A1型题）**：题干以论述题形式出现，或为叙述式，或为否定式。**病例摘要型最佳选择题（A2型题）**：以一个简要的病例作为题干。 |
| **病例组型最佳选择题**（A3） | **共享题干 + 问题 + 选项** | 题干开始叙述一个以患者为中心的临床情景，然后提出若干个（一个以上）相关问题，每个问题均与开始的临床情景有关，但测试要点不同。**试题考察以临床应用为主。** |
| **标准配伍题**（B1） | **共享选项 + 问题（题干）** | 每道试题由A、B、C、D、E五个备选答案与2个或2个以上的问题（题干）组成，5个备选答案在前，问题在后。答题时，要求为每个问题选择一项作为正确答案。**每个共享的备选答案可以选用任意次。** |

A1/A2类型题型样例如下，模型输入为question的内容，输出的目标为“answer”和“analysis”的内容。

"question": "治疗肩周疼痛，以肩后部为重，疼痛拒按，除肩部穴外，还应选取的主要经穴是（ ）。\nA．手太阳小肠经\nB．手阳明大肠经\nC．手少阳三焦经\nD．足少阳胆经\nE．足太阳膀胱经",

"answer": [

 "A"

 ],

"analysis": "手太阳小肠经起始于手小指外侧的末端，沿着手掌小指边而上行至腕关节部，出于手踝骨中，直行向上沿着前臂外侧后缘到达肘关节内侧尺骨鹰嘴和肱骨内上踝之间，向上沿着上臂内侧后缘到达肩关节部，绕行于肩胛，与诸阳经交会于肩上至大椎穴处，再向前行进入缺盆，络于心，沿食管向下穿过膈肌至腹腔属本腑小肠。此经脉一分支是从缺盆穴处分出，沿颈侧向上达面颊，行至外眼角（目锐眦），折返进入耳中。又一支脉是从面颊部分出，上行至眼眶下方，抵达鼻旁，行至内眼角（目内眦）。",

"index": 263

A3类题型样例如下，每个题干（share\_content）对应一个问题列表（question），由多个问题（subquestion）组成。模型输入的内容为share\_content、subquestion以及上一个subquestion的输出，输出的目标为“answer”和“analysis”的内容：

"share\_content": "患者，女，28岁，已婚，近3月来每逢经前出现小腹灼热胀痛，拒按，经色暗红，质稠有块，平素带下量多色黄，小腹疼痛，经来疼痛加剧，经前低热，小便黄赤，舌紫红，苔黄而腻，脉滑数。",
"question": [
 {
 "sub\_question": "1)．其辨证是（ ）。\nA．肝郁血热证\nB．阳盛血热证\nC．血热互结证\nD．湿热瘀阻证\nE．气滞血瘀证\n",
 "answer": [
 "D"
 ],
 "analysis": "根据患者的临床表现诊断为痛经湿热瘀阻证，主要证候：经前或经期小腹疼痛或胀痛不适，有灼热感，或痛连腰骶，或平时小腹疼痛，经前加剧；经血量多或经期长，色暗红，质稠或夹较多黏液；平素带下量多，色黄质稠有臭味；或伴有低热起伏，小便黄赤；舌质红，苔黄腻，脉滑数或弦数。"
 },
 {
 "sub\_question": "2)．其治法有（ ）。\nA．清热除湿，化瘀止痛\nB．清热解毒，行气止痛\nC．活血化瘀，理气止痛\nD．清热利湿，化瘀散结\nE．清热利湿，解毒排放\n",
 "answer": [
 "A"
 ],
 "analysis": "痛经湿热瘀阻证治法：清热除湿，化瘀止痛。方药：清热调血汤加车前子、薏苡仁、败酱草或银甲丸。"
 },
 {
 "sub\_question": "3)．治疗应首选的方剂是（ ）。\nA．止带方\nB．仙方活命饮\nC．血府逐瘀汤\nD．龙胆泻肝汤\nE．清热调血汤\n",
 "answer": [
 "E"
 ],
 "analysis": "痛经湿热瘀阻证治法：清热除湿，化瘀止痛。方药：清热调血汤加车前子、薏苡仁、败酱草或银甲丸。3"
 }
],
"index": 1

B1类题型样例如下，每个共用备选答案（share\_content）对应一个问题列表（question），由多个问题（subquestion）组成。模型输入的内容为share\_content、subquestion以及上一个subquestion的输出，输出的目标为“answer”和“analysis”的内容：

"share\_content": "（共用备选答案）\nA.清骨散\nB.知柏地黄丸\nC.清营汤\nD.黄连解毒汤\nE.五味消毒饮\n",
"question": [
 {
 "sub\_question": "1)．有清营分之热作用的方剂是（ ）。",
 "answer": [
 "C"
 ],
 "analysis": "清营汤功用为清营解毒，透热养阴。主治热入营分证。症见身热夜甚，神烦少寐，斑疹隐隐，舌绛而干，脉数。"
 },
 {
 "sub\_question": "2)．有清骨蒸潮热作用的方剂是（ ）。",
 "answer": [
 "A"
 ],
 "analysis": "清骨散功用为清虚热，退骨蒸。主治肝肾阴虚，虚热内扰证。症见骨蒸潮热，或低热日久不退，形体消瘦，唇红颧赤，困倦盗汗，或口渴心烦，舌红少苔，脉细数。27"
 }
],
"index": 208

根据问题类别的不同，我们构建了不同的指令模板，包含任务描述，思维链（CoT）和问答格式的限制。其中任务描述是为了明确不同题型的作答，CoT是用于引导LLMs理解和回答做出选项的同时给出相应的解析。

针对A1/A2类型的Prompt为：

请你做一道中医测试中A1或者A2类型的选择题\n请你一步一步思考并将思考过程写在【解析】和<eoe>之间。你将从A，B，C，D，E中选出一个最正确的答案，并写在【答案】和<eoa>之间。\n例如：【答案】: A <eoa>\n完整的题目回答的格式如下：\n【解析】 ... <eoe>\n【答案】 ... <eoa>\n请你严格按照上述格式作答。\n题目如下：**(需要填入question的内容)**

针对A3类型的Prompt为：

请你回答中医测试中A3或者A4类型的题目。每个题目包含一个案例，每个案例下设若干道选择题。\n请根据案例所提供的信息，一步一步思考并将思考过程写在每个选择题的【解析】和<eoe>之间，并从A，B，C，D，E中选出一个最正确的答案，并写在【答案】和<eoa>之间。\n例如：【答案】: A <eoa>\n完整的题目回答的格式如下：1）\n【解析】 ... <eoe>\n【答案】 ... <eoa>\n2）\n【解析】 ... <eoe>\n【答案】 ... <eoa>\n3）\n【解析】 ... <eoe>\n【答案】 ... <eoa>\n请你严格按照上述格式作答。\n案例如下：**(需要填入share\_content的内容，并将上一个问题的答案作为下一个问题（subquestion）的上下文输入)**

针对B1类型的Prompt为：

请你回答中医测试中B1类型的题目。每组题目有若干道选组题，共用列出的A、B、C、D、E五个备选答案，请从中选择一个与问题关系最密切的答案。某一个备选答案可能被选择一次，多次或不被选择。\n请一步一步思考并将思考过程写在每个选择题的【解析】和<eoe>之间，并从A，B，C，D，E中选出正确的答案，并写在【答案】和<eoa>之间。\n例如：【答案】: A <eoa>\n完整的题目回答的格式如下：1）\n【解析】 ... <eoe>\n【答案】 ... <eoa>\n2）\n【解析】 ... <eoe>\n【答案】 ... <eoa>\n3）\n【解析】 ... <eoe>\n【答案】 ... <eoa>\n请你严格按照上述格式作答。\n案例如下：**(需要填入share\_content的内容，并将上一个问题的答案作为下一个问题（subquestion）的上下文输入)**

最后采用统一的规则对模型输出的内容进行结构化并分割答案（answer）和解析（analysis）。

#### 中医语义推理任务

我们共构建了29,497个NLI数据对。受到之前工作的启发，我们认为如果两个句子之间存在稳定的语义相似性，那么应该可以从前提推导出假设，也可以从假设推导出前提。鉴于QWA与标准解析的长度差异较大，我们将TMNLI中一半的数据对的前提和假设进行置换，以消除长度带来的偏差。输入的内容为premise和hypothesis，输出的目标为label。

示例如下：

{"idx":1,

"premise":"具有解表通便功用的方剂是防风通圣散。",

"hypothesis":"“儿科鼻祖”钱乙的《小儿药证直诀》是我国现存最早的儿科专著，其研制的五脏补泻的方剂，成为后世医家临证研究、化裁古方的重要文献。",

"label":2}

其输入的内容为premise和hypothesis，输出的目标为label。

# 评测细则

## 评测数据发布

本次评测数据分为两次发布：

* 5月14日：发布训练集和验证集，A榜测试集(即测试集A)。同时发布测试pipline代码，供选手快速上手。测试集A不包含答案和解析，选手需要在测试集A上进行预测，提交至评测平台上获得得分, 即A榜得分。
* 8月1日：发布B榜测试集，即测试集B。最终评测排名将仅依据B榜结果。B榜评测的开放时间为： 8月1日开始，到8月8日截止。

## 提交文件

TCMBench中将各个任务都已经转化为了根据prompt生成回复的统一格式。

1. 中医知识理解任务：测试样本在test\_A12.json、test\_A3.json和test\_B1.json文件中，对应任务测试的三类题型。根据“keywords”的不同会触发不同的测试pipline。”example”列表即为测试样例数据，其格式在“任务形式”中描述。对于test\_A12.json文件，“question”字段作为模型输入，“answer”和“analysis”字段为空字符串，待选手们填入模型回复。对于test\_A3.json和test\_B1.json文件，”share\_content”和“question”字段列表中的每个“subquestion”字段作为模型输入，每个“subquestion”字段后的“answer”和“analysis”字段为空字符串，待选手们填入模型回复。
2. 中医语义推理任务，测试样本在为test\_NLI.json文件中，“premise“”和“hypothesis”字段作为模型输入，“label”字段为空字符串，待选手们填入模型回复。

**为完成测试集提交和获得评测得分，参与者需要提供四个文件，要求**：

1. 评测参与者需要提交自己训练的模型在测试集上的生成结果，分别命名为test\_A12\_predictions.json、test\_A3\_predictions.json、test\_B1\_predictions.json和test\_NLI\_predictions.json文件。
2. json格式、样本数量和样本顺序要与官方提供的test\_XXX.json文件一致。

## 评价指标

本评测任务有两个测试集，我们采用在各个任务上分别计分的方式进行评测。各个任务上的评测指标如下：

* 对于中医知识理解任务，需要使用两类指标评测。
	+ 采用准确率（Accuracy）比较正确选项和模型回复的选项，用于评估模型对中医知识的掌握程度。
	+ 使用Rouge-1、Rouge-L、BertScore、BartScore、SARI，自动比较模型生成的解析与标准解析之间的语义相似性，用于评估模型解析中医知识的能力。
* 对于中医语义推理任务，采用准确率（Accuracy）分数作为评估指标。

**总体打分：**

我们将对每个任务上的Accuracy或者Rouge-1、Rouge-L、BertScore、BartScore、SARI分数进行平均，得到总体分数，作为榜单排名的及评奖的依据。

# 时间安排

|  |  |
| --- | --- |
| **赛程** | **时间** |
| 评测任务发布 | 2024/04/30 |
| 报名时间 | 2024/05/01 - 2024/08/01  |
| 训练集，验证集发布；baseline模型/代码发布 | 2024/05/14 |
| 测试A榜数据(测试集A)发布 | 2024/05/14 |
| 测试A榜评测截止 | 2024/08/01  |
| 测试B榜数据(测试集B)发布 | 2024/08/08 |
| 测试B榜最终测试结果提交截止 | 2024/08/08 |
| 公布测试结果 | 2024/08/15 |
| CCKS-2024会议 (评测报告及颁奖) | 2024/09/19-2024/09/22 |
| 评测论文提交 | 2024/09/01 |

# 任务提交

## 所有队伍任务提交

本次任务的两个赛道，“提示学习赛道”和“高效参数微调赛道”，都将采取排行榜的方式。

### A榜阶段

A榜测试数据发布后，参赛队伍报名成功后即可通过天池平台下载数据，本地调试LLM模型，在线提交结果。A榜提供第二个任务的训练数据集和两个任务的验证数据集，供参赛选手微调语言模型。同时提供测试数据集A，供参赛选手提交评测结果，参与排名。A榜阶段提交格式在“赛题与数据”页面有详细介绍。

A榜阶段允许参赛队伍每天至多向天池平台提交3次结果，排名实时更新。参赛队伍可在测试A榜截止之前，随时上传测试集A榜的计算结果，管理系统会及时更新各队伍的最新排名情况。测试集A榜不计入比赛最终成绩（即B榜成绩）。

### B榜阶段

B榜阶段会提供B榜测试数据集下载，供参赛选手提交评测结果，参与最终排名与评奖。B榜阶段提交规范和A榜阶段保持一致。同样请注意，验证集和测试集A不可用于模型训练阶段。

B榜阶段允许参赛队伍每天至多向天池平台提交5次实时结果评测，每小时更新排行榜，按照评测指标从高到低排序。排行榜将选择参赛队伍在本阶段的历史最优成绩进行排名展示。

## 获奖队伍材料提交（供资格审查，符合要求后颁发奖项和奖金）

对TCMBench的两个赛道的前五名获奖参赛选手需额外提交相关材料以供资格审查：

1. 测试集结果文件，包含test\_A12\_predictions.json、test\_A3\_predictions.json、test\_B1\_predictions.json和test\_NLI\_predictions.json文件。
2. 由于本次评测允许参赛选手在测试集预测时对prompt进行扩充，获奖队伍需要提交每个测试集样本预测时的完整prompt，供组织方审核验证。
3. 对“预训练赛道”的参赛选手，需要提交：模型实现代码, 所使用的开源模型名称，微调使用的数据集，参数高效微调模块的完整文件或者下载地址供组织方审核验证。对“不微调赛道”的参赛选手，需要提交：模型实现代码, 所使用的开源模型名称。若使用了辅助句子表征模型，则需要提供完整模型文件。
4. 模型实现的方法描述文档。
5. 每个赛道的前三名获奖队伍撰写的评测论文将会由CCKS-2024会议统一安排发表，中稿率高于main track。

### 获奖队伍材料提交方式

将提交材料文件在任务提交截止日期前发送至邮箱wjyue@stu.ecnu.edu.cn。邮件的标题为：“CCKS2024-参赛队名称”，例如“CCKS2024-TCMBench小分队”。代码及其文档需打包成一个文件（tar，zip，gzip，rar等均可），作为邮件附件传送，要求提交所有的程序代码及相关的配置说明，程序应当可以运行，可以供组织者验证。对于大型模型文件或者数据文件，可以在邮件中注明下载链接。

# 赛事交流与通知

微信群二维码：



钉钉群：



# 评测规则

1. TCMBench评测完全依托于开源模型社区，如果参赛选手选择使用大模型，则必须从以下开源大模型中选择一个进行参赛：
2. [Qwen开源20B以下模型](https://huggingface.co/Qwen)
3. [书生开源全系列] (<https://huggingface.co/internlm> )
4. [ChatGLM3-6B](https://huggingface.co/THUDM/chatglm3-6b);
5. [ShenNong-TCM-LLM](https://huggingface.co/michaelwzhu/ShenNong-TCM-LLM)
6. 针对"提示学习赛道"的规则:
	1. 在提示学习赛道，参赛选手需要保持模型在huggingface网站上下载得到的参数，不能对开源大模型进行任何形式的微调，也不能在大模型上添加任何参数。
	2. 选手可以通过扩展提示内容或者从训练集中选择示例等方式来增强模型表现，但不能更改或者删减指令的内容。允许的指令扩充包括：示例(demonstrations),也可以是选手书写的对任务/标注规范的额外说明，也可以是对输出格式的说明或规定。demonstrations中的样例必须来源于TCMBench官方提供的训练集。
	3. 选手可以微调一个参数不超过1.5亿的句子表征模型作为辅助，帮助其针对每个不同的测试样本，在训练集中进行demonstrations选择。
	4. 在A榜，B榜测试集预测时，选手不能将LLM多次生成的结果进行集成(类似于self-consistency的方法)，提交的每个测试样本预测结果必须是LLM单次回复生成的。
7. 针对"高效参数微调赛道"的规则:
	1. 在参数高效微调赛道，参赛选手需要保持大模型主干模型在huggingface网站上下载得到的参数不变；
	2. 选手可以在开源的大模型主干之上添加参数高效微调模块(添加的额外参数量不得高于大模型主干参数量的1%), 且只能微调参数高效微调模块，不得微调大模型主干，且只能使用一组参数高效微调模块完成所有的任务
	3. 选手可以使用任何开源资源进行模型参数高效微调，但需要在材料提交时注明其训练数据来源，同时禁止将验证集和测试集中的数据加入微调数据中。
	4. 选手可以通过扩展提示内容或者从训练集中选择示例等方式来增强模型表现，但不能更改或者删减指令的内容。允许的指令扩充包括：示例(demonstrations),也可以是选手书写的对任务/标注规范的额外说明，也可以是对输出格式的说明或规定。demonstrations中的样例必须来源于TCMBench官方提供的训练集。
	5. 本赛道选手不得借助额外的LLM模型帮助其在训练集中进行demonstrations选择。
	6. 在A榜，B榜测试集预测时，选手不能将LLM多次生成的结果进行集成(类似于self-consistency的方法)，提交的每个测试样本预测结果必须是LLM单次回复生成的。
	7. 对于选择非大模型进行训练的参赛队伍，可以自由选择任意语言模型进行端到端的训练，但模型参数量不得超过1.5亿。
8. 天池平台参赛规则
	1. 参赛选手需在注册时进行实名认证，且需要确保提交信息准确有效，所有的比赛资格及奖金支付均以提交信息为准；
	2. 参赛选手报名后可在“我的团队”页面组队。每支队伍需指定一名队长，队伍成员总数最多不超过4人；报名截止日期之后不允许更改队伍成员名单，请谨慎选择队员组队。每名参赛选手只能参加一支队伍，一支队伍可以既参加TCMBench评测的“不微调赛道”也可以参加“预训练赛道”。 但是在评选奖励时，一只队伍只能选择一个赛道参与评奖。
	3. 队伍名的设定不得违反中国法律法规或社会公序良俗，且参赛队伍命名中不可出现“官方”，“天池”等字样。若命名违规的队伍在收到比赛主办方警告后仍未修改队伍名称，比赛主办方有权解散该队伍 ；
	4. 每名参赛选手只能参加一支队伍，一旦发现某选手以注册多个账号的方式参加多支队伍，将取消相关队伍的参赛资格；
	5. 参赛队伍可在参赛期间随时上传测试集的预测结果，阿里天池平台A榜阶段每天可提交3次、B榜阶段每天可提交5次，系统会实时更新当前最新榜单排名情况，严禁参赛团队注册其它账号多次提交。
	6. 验证集(包括采用验证集进行各种数据扩充操作后得到的数据)不可以用作模型训练。测试集A(包括采用测试集A进行各种数据扩充操作后得到的数据)也不可以参与模型训练。违反此规定将会直接导致参赛团队失去参赛资格和评奖资格。
	7. 要求最终结果排名前3的队伍按照”获奖队伍材料提交”部分的要求进行材料提交，供比赛组织方进行审核验证。参与评奖时，一只队伍只能选择TCMBench的其中一个赛道参与评奖。参赛团队需保证提交作品的合规性，若出现下列或其他重大违规的情况，将取消参赛团队的参赛资格和成绩，获奖团队名单依次递补。重大违规情况如下：（1）使用小号、串通、剽窃他人未公开的代码等涉嫌违规、作弊行为；（2）团队提交的材料内容不完整，或提交任何虚假信息，或者是无法复现预测结果；（3）参赛团队无法就提交材料的疑义进行足够信服的解释说明。

# 评测奖励

1. 本次评测的优胜团队（每个赛道的前三名）可以在CCKS-2024会议上进行技术汇报，与学术界大牛和业界大咖交流。
2. 鼓励优胜团队撰写评测论文，论文将由CCKS-2024组委会统一安排审核，其论文中稿率高于main track。

# 任务组织者和联系人

## 任务组织

王晓玲，华东师范大学，xlwang@cs.ecnu.edu.cn

岳文静, 华东师范大学， wjyue@stu.ecnu.edu.cn

朱威, 华东师范大学， wzhu@stu.ecnu.edu.cn

孙长志, 华东师范大学， czsun.cs@gmail.com

王新宇，华东师范大学，xinyu\_wang@stu.ecnu.edu.cn

周雅倩，复旦大学，yqzhou@cs.fudan.edu.cn

周毅萍，上海中医药大学，sunrising318@163.com

## 学术指导委员会