# CCKS-PromptCBLUE医疗大模型评测

**会议地点：东北大学，沈阳**

**会议时间：8月24日—27日**

# 任务描述

以ChatGPT、GPT-4等为代表的大语言模型（Large Language Model, LLM）掀起了新一轮自然语言处理领域的研究浪潮，展现出了类通用人工智能（AGI）的能力，受到业界广泛关注。在LLM大行其道的背景下，几乎所有的NLP任务都转化为了基于提示的语言生成任务。然而，在中文医学NLP社区中，尚未有一个统一任务形式的评测基准。

为推动LLM在医疗领域的发展和落地，华东师范大学王晓玲教授团队联合阿里云天池推出**PromptCBLUE**评测基准。 **PromptCBLUE**评测基准是对[CBLUE基准](https://tianchi.aliyun.com/dataset/95414)进行二次开发，将16种不同的医疗场景NLP任务全部转化为基于提示的语言生成任务,形成首个中文医疗场景的LLM评测基准，有利于帮助开源社区和业界快速评测公开的或者自有的LLM模型。

考虑到目前的LLM训练可能涉及商业数据，大规模模型开源受到各种外在条件的限制，我们将对PromptCBLUE评测开放两个榜单：

* 通用赛道，接受来自企业，高校，开源社区，各类研究团队或者个人对自研的LLM进行评测。
* 开源赛道，接受各类参赛团队提交评测，但是其必须使用开源的大模型底座，且只能使用开源的或者可以全部提交至比赛组织方审核的数据集进行训练/微调。

**评测地址**：<https://tianchi.aliyun.com/competition/entrance/532084/introduction>

## 任务形式：

在PromptCBLUE任务下，我们采用94个指令微调模板，对CBLUE基准中的各个任务进行改造，统一转化为基于prompt进行回复生成的形式。经过改造后，医疗文本NLP数据集都将转化为如下格式。input字段是模型的输入，target字段是模型的输出，type是原任务类型(不作为模型输入)，answer\_choices字段是选项，只有分类、术语标准化、推理类任务上该字段才会有意义。

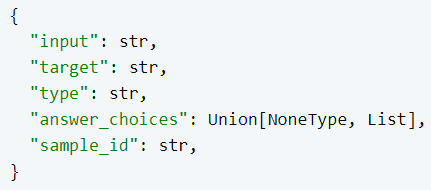


图1. PromptCBLUE中样本的一般格式

为了将CBLUE中的各种不同任务适配为符合LLM的输入输出格式，我们对CBLUE各个数据集进行了相应的改造。以下两个例子是PromptCBLUE中CMeEE任务和CHIP-CDN任务的输入输出形式举例。（更多说明请参考[PromptCBLUE的github库](https://github.com/michael-wzhu/PromptCBLUE" \t "dlt)）

#### CMeEE任务

本任务原本是标准的医学文本NER任务，选手需要给出医学实体mention在待抽取文本中的具体span位置。在PromptCBLUE中，本任务被改造为：根据指定的实体类型，生成实体mention。在评分时，我们只考虑实体mention及其类型标签，不再考虑span位置信息。样例如下:

{

"input": "医学实体识别：\n外周血白细胞计数常明显升高，伴核左移。\n实体选项：疾病，医学检验项目，医院科室，身体部位，微生物类，临床表现，药物\n答：",

"target": "上述句子中的实体包含：\n医学检验项目实体：外周血白细胞计数\n疾病实体：\n医院科室实体：\n药物实体：\n微生物类实体：",

"answer\_choices": ["疾病", "医学检验项目", "医院科室", "身体部位", "微生物类", "临床表现", "药物"],

"task\_type": "ner",

"task\_dataset": "CMeEE-V2",

"sample\_id": "train-134372"

}

#### CHIP-CDN任务

该任务原型是给定一诊断原词，要求给出其对应的诊断标准词，而诊断标准词是从ICD-10这一个4w+的标准词库中选择。由于我们不可能一次性将四万个词输入到LLM中，所以我们将CDN任务改造为：给定原词，从候选的若干个ICD-10诊断标准词中选择出匹配的词(可能有多个, 可能一个都没有)。

{

"input": "主动脉弓缩窄心功能低下\n归一化后的标准词是？\n实体选项：胫前动脉假性动脉瘤，主动脉缩窄，男性性腺功能低下，男性性腺功能低下，垂体功能低下，心功能不全\n说明：从候选的若干个ICD-10诊断标准词中选择出与原诊断描述匹配的词\n答：",

"target": "主动脉缩窄，心功能不全",

"answer\_choices": ["胫前动脉假性动脉瘤", "主动脉缩窄", "男性性腺功能低下", "男性性腺功能低下", "垂体功能低下", "心功能不全"],

"task\_type": "normalization",

"task\_dataset": "CHIP-CDN",

"sample\_id": "train-17932"

}

# 评测细则

## 评测数据发布

本次评测数据分为两次发布：

* 5月10日：发布训练集和验证集，A榜测试集(即测试集A)。同时发布baseline代码和模型，以及评测脚本，供选手本地在dev集上测试评估。测试集A不包含答案，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。
* 7月14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： 7月14日开始，到7月17日17: 59: 59截止。

## 提交文件

PromptCBLUE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，测试样本在test.json文件中，这个文件采用json line格式，每行是一个可json化的字符串，json化后"input"字段作为LLM输入，"target"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，参与者需要提供两个文件：

1. 评测参与者需要提交自己训练的LLM在测试集上的生成结果，命名为test\_predictions.json文件，保持json-line格式。样本数量，样本顺序要与官方提供的test.json文件一致。
2. 参赛者需要提供解析test\_predictions.json文件转化为results.json文件的代码，命名为post\_generate\_process.py 。本代码只限于使用python标准库，并可采用如下命令运行：

python post\_generate\_process.py test\_predictions.json test\_structured.json

results.json文件整体可以采用json.load()方法加载。其内部结构如下：

{

"task\_name": [

{

"sample\_id": str,

"answer": Union[str, dict],

}

]

}

}

results.json文件的更具体的格式说明见[PromptCBLUE的github库](https://github.com/michael-wzhu/PromptCBLUE)。

在参赛过程中，选手需要将test\_predictions.json文件与post\_generate\_process.py文件打包为test\_predictions.zip文件，提交到天池平台的镜像中进行运行，得到results.json文件。注意在zip文件中不要添加文件夹，即两个被打包的文件必须置于zip文件的顶层。比赛组织方将会根据天池镜像中的results.json文件打分。选手手动上传的results.json文件不作为评分依据。

更多对于提交文件格式的要求，以及示例，详见[PromptCBLUE的github库](https://github.com/michael-wzhu/PromptCBLUE)。

## 评价指标

本评测任务只有一个测试集，但是其包含多个任务的测试样本，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CMeEE-V2和IMCS-V2-NER任务，采用基于实体实例层面的严格的(strict)，micro的Precision, Recall, F1分数。这里的实体实例包含mention（即实体名称的所有组成字符）和类型这两个组成字段。这里"严格的"指模型必须在指定的样本sample\_id上，完全正确预测出ground truth中的实体实例的mention和类型，才能算是成功预测出这个实体实例，则true positive (TP) 加1。而如果模型预测的实体实例不在ground truth中，则false positive (FP)加1。如果ground truth中的实体实例未被模型预测到，则false negative(FN)加1。最终根据整个测试集上的TP，FP，FN计算Precision, Recall, F1分数。
* 对于CMeIE任务，采用基于三元组实例层面的严格的(strict)，micro的precision, recall, F1分数。这里的三元组实例包含头实体mention, 尾实体mention，和关系类型字段。
* 对于CHIP-CDEE任务，采用基于临床事件实例层面的严格的(strict)，micro的precision, recall, F1分数。这里的临床事件实例包含主体词，发生状态，描述词和解剖部位字段。
* 对于IMCS-V2-SR和CHIP-MDCFNPC任务，采用基于临床发现或者症状实例层面的严格的(strict)，micro的precision, recall, F1分数。这里的临床发现或者症状实例包含mention和阴阳性判断标签字段。
* 对CHIP-CDN任务，采用基于ICD-10标准词实例层面的严格的(strict)，micro的precision, recall, F1分数。这里的ICD-10标准词实例包含mention和阴阳性判断标签字段。
* 对CHIP-STS， KUAKE-QQR, KUAKE-IR，KUAKE-QTR任务，我们采用Micro的precision, recall, F1分数作为评估指标。对CHIP-CTC，IMCS-V2-DAC，KUAKE-QIC, 采用Macro的precision, recall, F1分数作为评估指标。
* 对于MedDG和IMCS-V2-MRG数据集，我们采用Rouge-1，Rouge-2，Rouge-L分数作为评估指标。为避免分词影响，计算rouge分数前，会将句子中的汉字拆开，用空格分隔。IMCS-V2-MRG任务中，需要将模型生成的诊断报告拆分为主诉, 现病史, 辅助检查, 既往史, 诊断, 建议这六个章节，分别计算rouge得分后取平均分。

上述任务中，F1(micro/macro)或者Rouge-L将作为每个任务的主要指标。

**总体打分：**

我们将对每个任务上的F1(micro/macro)或者Rouge-L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 | 2023/5/4 |
| 报名时间 | 2023/5/4 - 2023/7/13 17:59:59 |
| 训练集，验证集发布；baseline模型/代码发布 | 2023/5/10 |
| 测试A榜数据(测试集A)发布 | 2023/5/10 |
| 测试A榜评测截止 | 2023/7/13 17:59:59 |
| 测试B榜数据(测试集B)发布 | 2023/7/14 |
| 测试B榜最终测试结果提交截止 | 2023/7/17 17:59:59 |
| 公布测试结果 | 2023/7/28 |
| 参与评奖队伍审核材料提交 | 2023/8/5 23:59:59 |
| 评测论文提交 | 2023/8/11 23:59:59 |
| CCKS会议 (评测报告及颁奖) | 2023/8/24-2023/8/27 |

# 任务提交

## 参赛阶段任务提交

本次任务的两个赛道，“通用赛道”和“开源赛道”，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，在线提交结果。A榜提供训练数据集和验证数据集，供参赛选手微调LLM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）

### “开源赛道”提交材料

对PromptCBLUE “开源赛道”的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\_predictions.json文件，post\_generate\_process.py文件的test\_predictions.zip文件。

2. 对“开源赛道”的参赛选手，需要提交：模型实现代码与模型文件/或者下载链接；所有使用过的训练数据集文件/或者数据的公开可下载链接。

3. 模型实现的方法描述文档。

4. 我们鼓励参赛选手撰写评测论文，优秀的评测论文将被推荐至CCKS主会及对应期刊发表，将有机会在CCKS主会上分享研究成果(CCKS会议注册费需自理)。

若没有提交上述材料，将视为自动放弃参与评奖。评奖资格将会自动顺延至排名紧随其后的队伍。选手需要将以上材料在截止日期前(2023/8/5 23:59:59)发送至邮箱wzhu@stu.ecnu.edu.cn。邮件的标题为：“CCKS2023-参赛队名称”，例如“CCKS2023-PromptCBLUE小分队”。代码及其文档需打包成一个zip文件，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。如果方法使用了额外公开的数据资源，要求说明并提供资源文件或地址。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。

### “通用赛道”提交材料

对PromptCBLUE“通用赛道”的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\_predictions.json文件，post\_generate\_process.py文件的test\_predictions.zip文件。

2. 对“通用赛道”的参赛选手，如果条件允许，可以提交：模型实现代码与模型文件/或者下载链接；针对PromptCBLUE评测所采用的所有训练数据集文件/或者数据的公开可下载链接。如果参赛选手采用的大模型底座是商用的或者是受限于机构政策未能开源，参赛选手需要：提供一个可以访问的demo地址与账号(账号使用权限至少具有一天有效期)供组织者抽查验证。

3. 模型实现的方法描述文档。

4. 我们鼓励参赛选手撰写评测论文，优秀的评测论文将被推荐至CCKS主会及对应期刊发表，将有机会在CCKS主会上分享研究成果(CCKS会议注册费需自理)。

若没有提交上述材料，将视为自动放弃参与评奖。评奖资格将会自动顺延至排名紧随其后的队伍。选手需要将以上材料在截止日期前(2023/8/5 23:59:59)发送至邮箱wzhu@stu.ecnu.edu.cn。邮件的标题为：“CCKS2023-参赛队名称”，例如“CCKS2023-PromptCBLUE小分队”。若有代码，则需要代码可运行，可以供组织者验证。如果方法使用了额外公开的数据资源，要求详细说明。若提供demo地址，则需提供demo地址的有效时限，使用账号密码等信息。

# 赛事交流与通知

钉钉群二维码



# 评测规则

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即参加“通用赛道”也可以参加“开源赛道”。 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. PromptCBLUE的目标是评估LLM在不同医疗任务的总体表现，所以评测参与者只能使用一个LLM模型主干来完成整个测试集的评测。对于测试集中的每一个样本，模型输出必须是一个仅用LLM模型主干上连接的语言模型预测头(LM-head)输出的文本序列，LM-head必须是所有任务共享的。选手的最终模型不能在LM-head以外的其他模块产生与任务直接相关的或是(在模型训练过程中)参与损失计算的logits。
7. 如果参与者使用了参数高效微调方法，则其总共使用的参数高效微调模块的总参数量不得超过其LLM模型主干的1%。
8. 所有评测参与团队需要在提交测试集时，对其模型训练/微调方法进行介绍，也需要注明其训练数据来源。评测参与者可以使用任何资源进行LLM训练，包括采用自有的医疗领域(有标注/无标注)数据进行训练。
9. 考虑到目前的LLM训练可能涉及商业数据，大规模模型开源受到各种外在条件的限制，我们将对PromptCBLUE评测开放两个榜单：(1) 通用榜，接受来自企业，高校，开源社区，各类研究团队或者个人进行测试提交评测，不要求其对使用的LLM模型底座开源。但排名靠前的团队需要提供可访问的网页demo地址(最少1天使用权限)供组织者抽查审核。(2) 开源榜，接受各类参赛团队提交评测，但是其必须使用开源的大模型底座，且只能使用开源的或者可以全部提交至比赛组织方审核的数据集进行训练/微调。开源榜排名靠前的团队需要提交模型实现代码至组织者进行审核。
10. 评测参与者不得直接使用GPT-4, ChatGPT，文心一言，ChatGLM等公开模型api进行测试集预测(上述模型的研发机构除外)；上述大模型基座可以作为数据增广的来源。
11. 评测参与者可以采用更多的prompt，包括self-instruct方法得到的数据进行评测。但是在测试集预测时，需要采用测试集现有的prompt进行回复生成。
12. 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。
13. 要求两个赛道最终结果排名前10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。
14. 冠军奖和优胜奖的评选主要依据为B榜成绩。创新奖将由比赛组织方与学术指导委员会根据选手提交的材料评审决定。
15. 我们鼓励评测参与者采用自有的LLM生成预测结果的同时生成其思维过程。LLM的思维/推理过程将会成为创新奖的评选标准之一。

# 评测奖励

本评测总奖励16000元，奖金设置如下：

* 冠军奖：人民币5,000元，2名，奖励每个赛道获得最优平均表现的团队。
* 优胜奖：人民币2,000元，2名，奖励每个赛道排名前二的团队。
* 创新奖：人民币1,000元，2名，奖励在PromptCBLUE评测中采用了最为创新的方法并成绩突出的团队。

# 任务组织者和联系人

## 任务组织

朱威, 华东师范大学， wzhu@stu.ecnu.edu.cn

陈漠沙，阿里巴巴，[Chenmosha.cms@alibaba-inc.com](mailto:Chenmosha.cms@alibaba-inc.com)

王晓玲, 华东师范大学，xlwang@[cs.ecnu.edu.cn](http://cs.ecnu.edu.cn)

## 学术指导委员会

陈亮, 复旦大学附属华山医院

黄萱菁，复旦大学

贺樑，华东师范大学

杨晓春，东北大学

汤步洲, 哈尔滨工业大学（深圳）&鹏城实验室

王昊奋，同济大学