一、比赛评分项说明

1、评测指标定义：

诊断疾病的准确性：

准确率（Accuracy）： 模型正确诊断的病例数占总病例数的比例。

召回率（Recall）： 模型正确诊断的病例数占实际为该疾病的病例数的比例。

诊断依据的质量：

BLEU（Bilingual Evaluation Understudy）： 用于评估生成文本与参考文本之间的相似度，主要用于衡量生成文本的精确度。

ROUGE-L（Recall-Oriented Understudy for Gisting Evaluation）： 用于评估生成文本与参考文本之间的召回率，主要用于衡量生成文本的完整性。

2、评测流程：

数据准备：

提供一组包含病例信息的测试集，确保数据的多样性和代表性。

每个病例应包含完整的临床信息，如主诉、现病史、体格检查、辅助检查结果等。

模型提交：

参赛选手需提交模型对每个病例的诊断结果，包括诊断疾病和诊断依据。

诊断依据应为模型生成的文本，解释其诊断结果的理由。

评测过程：

诊断疾病评测：

计算模型诊断结果的准确率和召回率。

准确率 = 正确诊断的病例数 / 总病例数。

召回率 = 正确诊断的病例数 / 实际为该疾病的病例数。

诊断依据评测：

使用 BLEU 和 ROUGE-L 指标评估模型生成的诊断依据与参考答案之间的相似度。

BLEU 和 ROUGE-L 的计算可使用现有的评测工具，如 NLTK 库中的相关函数。

二、限定规则

1、模型限定 Qwen2.5-7B-Instruct，可以进行预训练、SFT、prompt Learning 等操作。

2、过程中不可以使用任何收费 API 蒸馏数据（开原模型 api 可以使用，如 deepseek、qwen-72B）。

3、复赛入围选手比赛结束时需要提交代码以及过程中使用到的所有数据（用于保障真实性），蒸馏其他模型的数据也需要保证过程可复现。