# 模型互评实验：课堂场景分类和输出评价

本次实验是为了通过让模型互评对方在课堂场景分类中的结果，来测试它们在分析和评价其他模型输出方面的效果和合理性。

## 实验步骤：

Step1生成分类结果：我将构建一个统一的课堂场景分类prompt，用于指导模型理解其需要执行的分类任务。然后，将课堂场景分类prompt+待分析文本分别送入GPT4和GLM4模型中，以获取各自的输出（GPT4\_result和GLM4\_result）。

**具体实现：**

**课堂场景分类Prompt：**给定一段文本，首要任务是从中识别出老师课堂指令。这个文本中可能涉及多个课堂指令。但每个指令只有一个课堂场景类别。如果文本中没有识别出任何指令，就将课堂类别判定为“无”。接着，需要根据识别出的指令内容将其分类到正确的课堂场景类别中，并给出分到该类别的理由和评分。评分标准只有高，低。

课堂场景类别包括：

学生展示：涉及学生在班级前展示或解释学术或课堂材料的场景。

生生互动：描述学生之间通过讨论、合作等方式进行互动的场景。

学生练习：包括学生独立或在小组中练习课堂所学知识和技能的场景。

学生听写：教师口述内容，学生进行书写的场景。

学生齐读：全班学生同步朗读或背诵文本的场景。

输出格式：

如果“课堂文本”中有一个或多个识别出“课堂活动”的指令，每个指令以下述JSON对象的格式输出 ，每个JSON对象包含：

"original\_sentence": 该指令对应的课堂文本中的老师话语。

"analysis":对 original\_sentence的分析过程。

"conclusion": 一个JSON数组，数组中的每个元素对应识别出的该指令对应的一种”课堂场景“类别（"label"）和相应的评分（"score"），例如，{"label": "学生齐读", "score": "高"}表示该元素的课堂场景类别为“学生齐读”，评分为“高”。

整个课堂文本的最终输出是所有指令语句构成的JSON数组。

**待分析文本：**"自学六七自然段默读自学提示，时间，七分钟，小组之间议一议开始"

**GPT4和GLM4的输出结果：**

**GPT4\_result**:

[{"original\_sentence": "默读自学提示，时间，七分钟",

"analysis": "老师指导学生进行默读自学，明确了时间限制，这是独立练习的一种形式。",

"conclusion": [{"label": "学生练习", "score": "高"} ] },

{"original\_sentence": "小组之间议一议开始",

"analysis": "老师指令让学生在小组之间进行讨论，这是学生间互动的一种活动。",

"conclusion": [ {"label": "生生互动", "score": "高"} ] }]

**GLM4\_result**:

[{'original\_sentence': '自学六七自然段默读自学提示',

'analysis': "这句话中，'自学'和'默读'表明这是一个学生独立学习的指令。",

'conclusion': [{'label': '学生练习', 'score': '高'}]},

{'original\_sentence': '小组之间议一议开始',

'analysis': "这里提到'小组之间议一议'，意味着学生需要在小组内进行讨论，属于生生互动的场景。",

'conclusion': [{'label': '生生互动', 'score': '高'}]}]

Step2构建评价prompt：接下来，构建一个评价模型输出的prompt（promt\_comment），用于把一个模型在课堂场景分类任务上的输出放到另一模型中进行评价。

**promt\_comment：  
 “**在以下的课堂场景分类任务中，我已根据文本内容识别出老师的课堂指令，并对这些指令进行了分类分析，归属到特定的课堂类别。现在需要你评价我的分析是否准确。请针对以下每条指令的分析结果，输出‘好’或‘不好’，并给出相应的理由。  
课堂场景分类prompt+待分析文本+人工判断 **”  
  
 注意*：***这里的人工判断其实就是模型的输出结果，*输出结果是由模型生成的，但是做的是模型互相评价任务，不要说这个输出结果是由xx模型生成的，要说这是人工判断的。*

Step3互评模型输出：使用构建的评价prompt结合GPT4的输出结果，送入GLM4进行评价；同时，将评价prompt与GLM4的输出结果结合，送入GPT4进行评价。

**GLM4对GPT4的评价**：

第一条指令判定：

原文：“自学六七自然段默读自学提示，时间，七分钟”

评价：好。

理由：您的分析正确地识别了“自学”和“默读”作为学生独立学习的指示词。这符合“学生练习”这一课堂场景类别的定义，因为学生被指示在一定时间内独立完成学习任务。评分标准为“高”也是恰当的，因为分析准确且指令清晰。

第二条指令判定：

原文：“小组之间议一议开始”

评价：好。

理由：分析中注意到“小组之间议一议”意味着学生之间的互动和讨论，这与“生生互动”类别的定义相符。同时，指令的表述也明确指向了学生之间的合作活动，评分标准为“高”反映了分析的准确性

**GPT4对GLM4的评价：**

对于第一条指令 "自学六七自然段默读自学提示" 的分析："这句话中，'自学'和'默读'表明这是一个学生独立学习的指令。"

评价：好。这个分析合理地捕捉到了文本中的关键词“自学”和“默读”，准确地指出了这是一个学生独立学习的场景。

对于第二条指令 "小组之间议一议开始" 的分析："这里提到'小组之间议一议'，意味着学生需要在小组内进行讨论，属于生生互动的场景。"

评价：好。分析准确识别了文本中的“小组之间议一议”，并正确地将其归类为生生互动的场景，因为它涉及到学生之间的讨论和合作。