请您以公正的评判者身份，评估 AI 助手在执行指定的[文本即时回复]任务时所提供答案的质量。 在评估过程中，请从以下四个维度进行考量：

1. \*\*自然性\*\*: 回复是否流畅自然，符合人类的交流习惯，避免机械或不真诚的表达。

2. \*\*清晰准确\*\*: 在情感交流中保持内容准确无误，传达明确信息，避免歧义。

3. \*\*语境相关\*\*: 确保回复紧密贴合对话上下文，情感适宜，紧扣话题。

4. \*\*尊重礼貌\*\*: 使用友好礼貌的语言，体现对用户的尊重，维护良好的交流氛围。

我们会为您提供文本即时回复任务的对话上下文、一个8分左右的参考答案，以及需要您评估的AI助手生成的答案。当您开始评估时，请按照以下流程操作：

1. 仔细阅读对话上下文，理解对话背景和用户意图。

2. 将AI助手生成的答案与参考答案进行比较，指出AI助手生成的答案有哪些不足，并进一步解释。

3. 从不同维度对AI助手生成的答案进行评价，在每个维度的评价之后，给每一个维度一个1～10的分数。

4. 最后，综合每个维度的评估，对AI助手生成的答案给出一个1～10的综合分数。

5. 你的打分需要尽可能严格，并且要遵守下面的评分规则：总的来说，AI助手的答案质量越高，则分数越高。

- 当AI助手的答案存在严重错误，例如完全不符合对话上下文、答非所问等情况时，总分必须是1到2分；

- 当AI助手的答案存在较多问题，例如表达生硬、语义不清等情况时，总分为3到4分；

- 当AI助手的答案基本满足要求，但在某些方面表现较差，例如流畅度不足、语境相关性不强等，总分可以得5到6分；

- 当AI助手的答案质量与参考答案相近，在所有维度上表现良好，总分得7到8分；

- 只有当AI助手的答案质量显著超过参考答案，能够全面准确地理解对话内容并做出高质量的回复，并且在所有维度上都接近满分的情况下，才能得9到10分。作为示例，参考答案可以得到8分。

请记住，你必须在你打分前进行评价和解释。在你对每个维度的解释之后，需要加上对该维度的打分。之后，在你回答的末尾，按照以下字典格式（包括括号）返回你所有的打分结果，并确保你的打分结果是整数：{'自然性': 打分, '清晰准确': 打分, '语境相关': 打分, '尊重礼貌': 打分, '综合得分': 打分}，例如：{'自然性': 9, '清晰准确': 7, '语境相关': 8, '尊重礼貌': 6, '综合得分': 7}。

[任务开始]

specific\_task

[任务结束]

[参考答案开始]

refer\_answer

[参考答案结束]

[AI助手答案开始]

ai\_answer

[AI助手答案结束]