请您以公正的评判者身份，评估 AI 助手在执行指定的[文本排版]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. \*\*一致性\*\*：确保文档内的编号、标题级别和列表使用统一的风格。保持编号格式一致，并确保标题层级结构清晰，逻辑关系明确。

2. \*\*可读性和易理解性\*\*：文本应易于阅读，适当分段，使用加粗/斜体强调重点，合理设置标题和子标题，以构建良好的结构，避免过度装饰影响内容表达。

3. \*\*美观度与多样性\*\*：在保证功能性的前提下，通过使用多样的Markdown格式提升文档的视觉吸引力，使其既整洁又不单调。

4. \*\*准确性\*\*：排版后的Markdown代码必须严格反映原始文本的所有内容，不改变任何实质性信息，确保内容的真实性与完整性。保持对原文的忠实度是关键。

我们会为您提供文本排版任务的原始内容、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. 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。

2. 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。

3. 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的综合得分。

4. 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排版错误、信息缺失或内容失真等情况时，总分必须是1到2分；

- 当AI助手的答案没有严重错误但质量较低，未能很好地体现原文内容或排版要求时，总分为3到4分；

- 当AI助手的答案基本满足排版要求，但在某些方面表现较差，质量中等，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，完全满足排版要求，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：{'一致性': 打分, '可读性和易理解性': 打分, '美观度与多样性': 打分, '准确性': 打分, '综合得分': 打分}，例如：{'一致性': 9, '可读性和易理解性': 7, '美观度与多样性': 8, '准确性': 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]