Multi-SWE-bench标注问卷

Multi-SWE-bench标注说明

👋 您好！

我们拥有来自多个开源 **JavaScript语言**仓库的 GitHub issue数据集，每个数据集都有一个成功解决该issue的 PR（Pull Request），每个 PR 包含两部分：

1. 解决issue的代码(Gold Patch)；
2. 测试文件相关的代码(Test Patch)，以验证该issue是否已解决。

我们希望将数据集中的样本作为**AI Agent**的编码能力的基准测试：对于每个样本，我们会将**问题描述（Issue Description）**提供给**工程师（AI Agent）**，并要求他们编写代码解决该issue（不提供原始 PR 的解决方案）。然后，我们会使用原始 PR 的测试文件来运行测试，以检查该**工程师（AI Agent）**的解决方案是否正确。

我们假定：

- 问题描述足够详细明确，能够让人明白问题是什么，以及正确的解决方案应该是什么样的。
- 测试的范围与问题相符，即它们能准确地按照问题描述中所阐述的那样对解决方案进行测试，而不会测试任何其他不相关的功能。

在这项任务中，您将帮忙检查这些假定，并确定哪些问题及测试样本适合用于我们的基准测试。

注意事项：

- 对于所有未获得满分的评分，必须提供相应的理由说明；
- 在选择**1 存在其他严重问题**作为2.1部分的评估时，请在详细说明理由后不填写其他维度。

您现在正在检查来自**JavaScript语言Kong/insomnia** 仓库的issue。

- **instance_id**: Kong-insomnia-8027
- **解决的issue的URL**:https://github.com/Kong/insomnia/issues/8020
- **PR的URL**: https://github.com/Kong/insomnia/pull/8027

请在继续标注之前，先在浏览器中打开上述相关链接，花点时间熟悉一下代码库的相关部分、issue和对应的PR。

第 1 部分 - 总结

问题1.1 请根据您的理解简要描述该issue的具体内容，并说明其属于哪一类型的 Issue（例如 Bug、New Feature、线程安全问题等，或更细分的类型）。

[自由文本，至少50个字符]

第 2 部分 - 是否存在严重问题？

问题 2.1 您认为该样本是否存在该问卷未涵盖的其他严重问题，即是否存在导致该样本不适合用作评估编码能力的其他问题？

- **0**: 不存在其他严重问题。
- **1**: 存在其他严重问题。

问题 2.2

如果您选择“是”，请解释原因：

[自由文本，最少50个字符]

第 3 部分 - 问题描述评估

请花几分钟阅读以下问题描述。如果需要，您可以从 PR 中导航到原始问题描述。

注意：

请不要点击任何外部链接，也不要阅读 GitHub issue 讨论区的评论。我们的设置只提供下面显示的主要问题文本，因此请仅根据这些信息回答问题。

工程师（AI Agent）可看见的Issue Description如下：

- Issue Description URL:https://github.com/Kong/insomnia/issues/8020

问题描述的详细程度评估

想象一下，你是一位经验丰富的软件工程师，接到指令要创建一个能成功解决上述 GitHub 问题的拉取请求（PR）。你可以完全访问代码库，并且能看到如上所述的问题描述（Issue Description）。但你只能完全依据这些信息来工作。

问题 3.1 问题描述（Issue Description）是否足够明确，以便尝试提出一个有意义的解决方案？

- **0**: 没有进一步信息，几乎不可能理解问题需要解决什么。
- **1**: 问题描述模糊，存在歧义，不清楚成功解决方案的样式。
- **2**: 有些细节需要推测，但大致可以解释出解决方案的要求。
- **3**: 问题描述明确，对于成功解决方案的要求清晰。

问题 3.2

请解释您的选择，并引用具体的文件名、函数/类名或代码行号（如果需要）：

[自由文本，最少50个字符]

第 4 部分 - 测试补丁评估

当提交了一个新的候选解决方案(新PR)，将会使用测试补丁(Test Patch)来检查该候选解决方案是否正确解决了问题；接下来您将评估Test Patch是否足以检查问题解决情况。

我们使用“黄金补丁”（Gold Patch）作为原始 PR 的解决方案，并用“测试补丁”（Test Patch）作为FAIL_TO_PASS测试以验证问题是否成功被该解决方案解决。请仔细研究以下测试补丁。如果需要，可以在 GitHub 差异查看器(Diff-Viewer)中查看原始 PR。

- Gold Patch与Test Patch的URL: https://github.com/Kong/insomnia/pull/8027.patch

假阴性检查：

在这个部分，重点检查即使提供了有效的解决方案，测试补丁是否也可能导致该有效方案不通过，从而导致假阴性：

如果有新的候选解决方案，我们打算使用测试补丁来检查该解决方案是否正确地解决了问题。然而，请记住，这些测试是针对特定的解决方案(黄金补丁)编写的，因此它们可能不适合评估其他有效的解决方案。我们希望知道这些测试是否正确地确定了问题的所有合理解决方案的范围，或者测试是否依赖于狭窄的细节，这将不公平地惩罚其他正确的新解决方案。 换句话说，只有当测试不依赖于问题描述中不存在的任何细节时，我们的假定才有效。我们发现，最常见的问题发生在测试与问题文本之间存在细微差异的情况下，例如测试依赖于在黄金补丁中引入的新函数、变量名或错误消息，但未提及或与问题描述不同。请仔细检查此类差异，并记住尝试解决此问题的工程师（AI agent）将无法访问原始PR或测试。

注意:只要有必要，您可以回到前面的部分重新熟悉这个问题。

问题 4.1 测试是否覆盖所有合理的解决方案？

- **0**: 测试范围过窄或过宽，或者关注的问题与描述的需求不同。
- **1**: 测试有效，但可能漏掉一些完全合理的解决方案。
- **2**: 测试覆盖大多数正确解决方案，但可能忽略一些不寻常的方案。
- **3**: 测试完美覆盖所有可能的解决方案。

问题 4.2

请解释您的选择，并引用具体的文件名、函数/类名或代码行号（如果需要）：

[自由文本，最少50个字符]

第 5 部分 - 其他问题难度评估

请重新查看 GitHub 问题，并回答以下问题：

难度评估：

问题 5.1 假设经验丰富的软件工程师有几小时熟悉代码库，理解问题、提出方案和编写代码大概需要多长时间？

- **<15 分钟**：只需要简单修改，例如为函数添加一些断言。
- **15 分钟 - 1 小时**：需要一些思考的小型改动。
- **1-4 小时**：需要对函数或多个文件进行较大修改。
- **4 小时以上**：需要大量研究的复杂问题，涉及修改 100 行以上代码。

问题 5.2

请解释您选择的难度等级：

[自由文本，最少50个字符]

第 6 部分 - 您的信心等级

您对所提供评估答案的信心有多高？

1 - 5 分（1 = 最低信心，5 = 最高信心）

- **1分**: 对自己的评估答案非常不确定，感觉在很多方面都缺乏足够的依据和理解，可能是因为对相关技术细节、问题背景或评估标准的把握不够准确，存在很多模糊和疑惑的地方，所以给出的答案很可能存在较大偏差。
- **2分**: 对评估答案有一定的疑虑，虽然对部分内容有一定的了解，但整体上仍感觉自己的判断可能受到一些不确定因素的影响，例如对某些代码片段的理解不够深入，或者对问题描述的解读可能存在多种可能性，导致对答案的准确性没有十足的把握。
- **3分**: 对大部分评估内容有一定的信心，基于现有的信息和自己的经验，能够做出相对合理的判断，但仍存在一些小的不确定因素，可能需要进一步的信息或他人的反馈来确认答案的准确性，不过总体上认为自己的评估方向是正确的。
- **4分**: 对自己的评估答案比较有信心，经过仔细的分析和思考，结合对问题的理解以及相关技术知识，认为答案在很大程度上是准确可靠的。虽然可能还存在一些细微的地方可以进一步完善，但整体上对自己的判断有较强的信心。
- **5分**: 对所提供的评估答案充满信心，确信自己已经充分理解了问题的各个方面，包括问题描述、测试补丁、代码库等相关内容，并且依据明确的标准和扎实的知识做出了准确无误的判断，对答案的正确性深信不疑。

附录说明

1. 标注人员需要假设问题描述中的嵌入图像是可见的。
2. 可以随时返回前面的部分查看问题描述，补丁代码以及相关代码上下文。