LIONBRIDGE

AI 数据分析
用于模型评估

评估角度

语言生成需要从多个维度进行衡量，因此，要评估大型语言模型 (LLM) 的性能就需要采用系统性的合理方法。我们选择从以下这些角度进行评估，以便全面地衡量模型产出内容的质量、可靠性和用户相关性。这种多角度方法确保评估绝不停留于字面是否准确，还可衡量流畅度、完整性、行业专业术语和文化相关性等。这让组织可以自信地评估模型输出的内容是否符合用户的具体期望、业务目标和伦理道德。

准确性

评估回复内容是否符合事实，没有错误。

例如：在回复法律问题时能够引用正确的法规条款。

完整性

检查对问题的回答是否全面、到位。

例如：在用户要求总结时列出所有要点。

相关性

评估是否会答非所问，以及回答是否直截了当。

例如：在描述产品时避免出现无关的信息。

一致性

评估回复的内容是否符合逻辑，不会自相矛盾。

例如：在较长篇幅的回复中不会自相矛盾。

流畅度

评估语法是否正确，语言是否流畅。

例如：句子结构和标点使用正确。

幻觉

评估模型是否会无中生有，捏造事实。
（幻觉越少，评分越高）

例如：不会捏造产品功能

术语

评估能否正确使用该领域的专业术语。

例如：在答复医疗保健问题时使用准确的医学术语。

可读性

评估产出的文字对于用户来说是否易于阅读理解。

例如：回答一般性问题时用词简洁明了。

文化相关性

评估回复中是否存在敏感内容，是否符合用户的文化习俗。

例如：在回复中避免出现不恰当的用语或俗语。

通过系统地从这些角度进行评估（采用李克特五点量表法），客户可以深入了解模型在哪些方面表现出色、哪里可能需要微调、是否还需多加训练或人工监督，并据此采取措施予以改善。这样可以确保人工智能 (AI) 部署在各种应用中（无论是客户支持、内容生成、产品推荐或其他领域）都提供一致的优质体验。

实例分析 | 挖掘更深层次的见解

Lionbridge 的数据服务分析可深入洞察 AI 训练的质量、一致性和效率，让您获取重要见解。通过这些见解，您可以了解数据多样性、标注一致性、模型优缺点以及输出趋势方面的情况，确保 AI 系统可靠、公平、性能优异。利用这些分析，组织可以根据数据做出决策，优化模型训练、改进标注工作流程，并自信地加速 AI 部署。

准确性

平均得分
4.3
1 — 5

标准差
0.6
0 — 1

通过率
（4 分以上通过）

排名前 25%
的数据占比

相关性

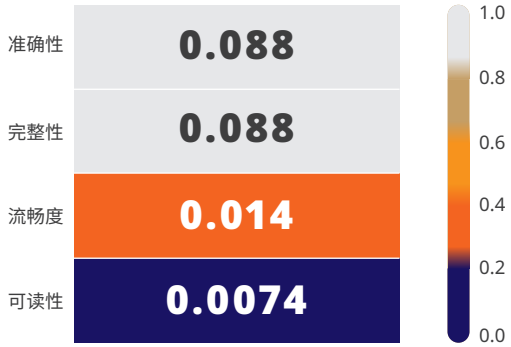
平均得分
4.2
1 — 5

标准差
0.5
0 — 1

通过率
（4 分以上通过）

排名前 25%
的数据占比

评分者间一致性（Fleiss' Kappa 系数）



FLEISS KAPPA 系数

评估角度之间的相关性

更多详情，请访问
LIONBRIDGE.COM