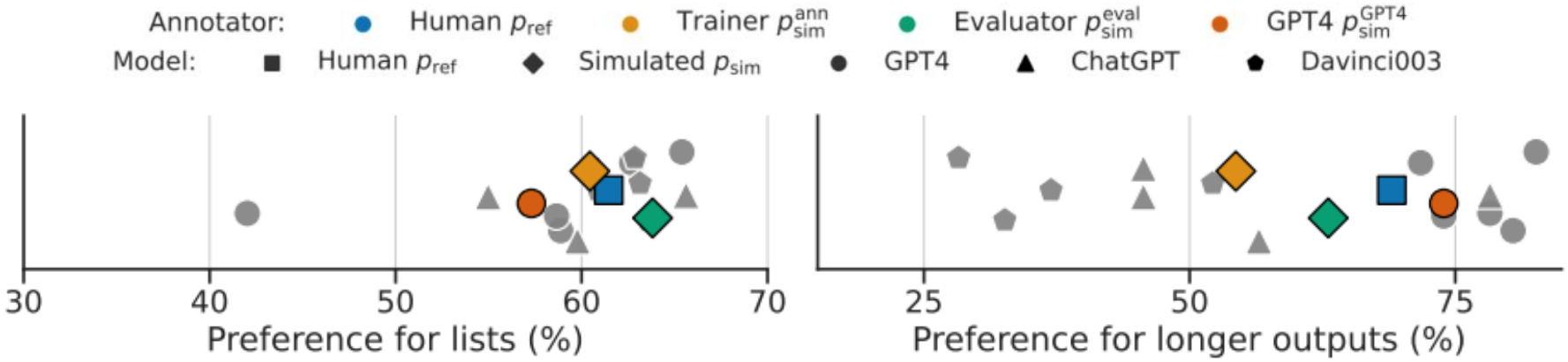
大模型自动评价的局限性

- 评价存在偏好
  - 大模型存在自我偏好，对于自己生成的文本内容会给予更好的评价[1]。
  - 存在排序偏差，交换两个待评价文本的前后顺序会导致评价出现偏差[1]。
  - 在AlpacaFarm的实验中，大模型偏好包含列表，文本更长的输出。



[1] LLM Evaluators Recognize and Favor Their Own Generations