# Chatbot Capability Evaluation

# Benchmark

|  |  |  |  |
| --- | --- | --- | --- |
| 名称 | 简介 | 数据集链接 | 参考文献 |
| Persona-Chat | 个性化对话生成，评估对话模型在与用户进行个性化、自然对话时的表现 | https://huggingface.co/datasets/AlekseyKorshuk/persona-chat | https://arxiv.org/pdf/1801.07243 |
| DailyDialog | 多轮对话生成，旨在评估模型在日常对话中的表现，能够提供更自然、更有意义的多轮对话。 | https://huggingface.co/datasets/li2017dailydialog/daily\_dialog | https://arxiv.org/abs/1710.03957 |
| MultiWOZ | 任务型对话系统，涵盖多个领域（如酒店预订、餐厅预定、航班信息查询等），评估模型在实际任务中的表现。 | https://github.com/budzianowski/multiwoz | https://arxiv.org/abs/1810.00278 |

# 评价指标与测试用例

| Benchmark | 评价指标与计算公式 | 测试用例 |
| --- | --- | --- |
| Persona-Chat | * F1 score：用于衡量生成回复与参考答案之间的词级别重合。 公式：F1 = 2 × (Precision × Recall) / (Precision + Recall) * Hits@1：在多个候选回复中，正确回复是否排名第一。   公式；Hits@1=模型预测正确回答的次数/总问题数​   * Perplexity：衡量语言模型预测下一个词的能力。 公式：PPL = exp(-1/N × Σ log(P(wᵢ)))   P(wi)是模型预测第 iii 个词的概率，N是总词数 | Persona: I like to play basketball. User: What do you usually do on weekends? Bot: I often go out to play basketball with friends. |
| DailyDialog | * BLEU：评估生成回复与参考答案之间的 n-gram 重合程度。 公式：BLEU = BP × exp(Σₙ wₙ × log(pₙ)) * F1 score：与 Persona-Chat 相同，适用于词汇重合度评估。 * ROUGE: 用于评估生成文本和参考文本之间的重叠度，主要用于评估生成的摘要或对话的质量。ROUGE=∑overlap(n−grams)/∑reference(n−grams)​ | User: Good morning, how are you? Bot: I'm great, thanks for asking! What about you? |
| **MT-Bench** | * Inform Rate: 衡量模型提供的信息是否符合用户需求。InformRate = Correct Information Provided/Total Information Needed * Success Rate: 衡量任务成功率，即模型是否成功完成了用户请求。Success Rate = Successful Task Completion/Total Tasks​ * BLEU:计算生成文本和参考文本之间的 n-gram 重叠度，评估生成对话与参考答案的相似度 | User: Can you help me find a hotel in the center of the city for tomorrow?  Bot:根据用户要求提供酒店选择、价格、可用性等信息，最终帮助用户完成预订。  User: What time is check-in?  Bot: 应当提供明确的入住时间，并进行相应确认。 |