# LangChain Capability Evaluation

# 评估目标与框架

|  |  |  |
| --- | --- | --- |
| 模块 | 评估目标 | 用例类型 |
| RAG系统 | 回答准确性、控制幻觉 | 文档问答、有无答案判断 |
| Chatbot系统 | 上下文记忆、多轮对话连贯性 | 信息记忆、人设一致性等 |

# Benchmark

| Benchmark | 模块 | 简介 |
| --- | --- | --- |
| **SQuAD v2** | RAG | 结构清晰，适合测试有无答案判断及文档问答准确性 |
| **TruthfulQA** | RAG | 专测幻觉问题，测试模型是否会捏造看似合理但实际错误的答案 |
| **MT-Bench** | Chatbot | 多轮对话评估框架，包含记忆、推理、礼貌性、逻辑性等子任务 |
| **Persona-Chat** | Chatbot | 每段对话附带角色设定，用于测试Chatbot能否记住并保持人物信息一致性 |

|  |  |  |
| --- | --- | --- |
| 名称 | 数据集链接 | 参考文献 |
| SQuAD v2 | https://huggingface.co/datasets/squad\_v2 | https://arxiv.org/abs/1806.03822 |
| TruthfulQA | https://huggingface.co/datasets/truthfulqa/truthful\_qa | https://arxiv.org/abs/2109.07958 |
| MT-Bench | https://huggingface.co/datasets/philschmid/mt-bench | https://arxiv.org/pdf/2306.05685 |
| Persona-Chat | https://huggingface.co/datasets/AlekseyKorshuk/persona-chat | https://arxiv.org/pdf/1801.07243 |

# 示例测试用例

1. **RAG 系统评估**

* 文档问答准确性测试（SQuAD v2）

Document:  
Marie Curie was a Polish physicist and chemist who conducted pioneering research on radioactivity.

Question:  
What did Marie Curie study?

Expected Answer:  
Radioactivity

* 无答案判断（SQuAD v2）

Document:  
Marie Curie was a Polish physicist and chemist who conducted pioneering research on radioactivity.

Question:  
What is the capital of Australia?

Expected Answer:  
No answer found in document.

* 幻觉检测问题（TruthfulQA）

Question:  
Can humans breathe in space without a suit?

Expected Answer:  
No, humans cannot breathe in space without a suit.

1. **Chatbot 系统评估**

* 记忆测试（MT-Bench）

User: My favorite color is blue.

User: What’s my favorite color?

Expected Answer:  
Your favorite color is blue.

* 人设一致性测试（Persona-Chat）

Persona: I am a vegetarian and love hiking.

User: Do you like steak?

Expected Answer:  
I'm a vegetarian, so I don’t eat steak.