🏠 Guides Evaluation

Evaluation

Building applications with language models involves many moving parts. One of the most critical components is ensuring that the outcomes produced by your models are reliable and useful across a broad array of inputs, and that they work well with your application's other software components. Ensuring reliability usually boils down to some combination of application design, testing & evaluation, and runtime checks.

The guides in this section review the APIs and functionality LangChain provides to help you better evaluate your applications. Evaluation and testing are both critical when thinking about deploying LLM applications, since production environments require repeatable and useful outcomes.

LangChain offers various types of evaluators to help you measure performance and integrity on diverse data, and we hope to encourage the community to create and share other useful evaluators so everyone can improve. These docs will introduce the evaluator types, how to use them, and provide some examples of their use in real-world scenarios.

Each evaluator type in LangChain comes with ready-to-use implementations and an extensible API that allows for customization according to your unique requirements. Here are some of the types of evaluators we offer:

- String Evaluators: These evaluators assess the predicted string for a given input, usually comparing it against a reference string.
- Trajectory Evaluators: These are used to evaluate the entire trajectory of agent actions.
- Comparison Evaluators: These evaluators are designed to compare predictions from two runs on a common input.

These evaluators can be used across various scenarios and can be applied to different chain and LLM implementations in the LangChain library.

We also are working to share guides and cookbooks that demonstrate how to use these evaluators in real-world scenarios, such as:

- Chain Comparisons: This example uses a comparison evaluator to predict the preferred output. It reviews ways to measure confidence intervals to select statistically significant differences in aggregate preference scores across different models or prompts.

Reference Docs

For detailed information on the available evaluators, including how to instantiate, configure, and customize them, check out the reference documentation directly.

🗃️ String Evaluators

4 items

🗃️ Comparison Evaluators

3 items

🗃️ Trajectory Evaluators

2 items

🗃️ Examples

1 items