How to evaluate an LLM application: evals, hallucination tests and guardrails

26 October 2026 · 6 min read · blogs.iictp.org/how-to-evaluate-an-llm-application

You cannot improve what you cannot measure, and an AI feature that “seems fine” in a demo can fail quietly in production. An evaluation set is to an AI application what a test suite is to ordinary code.

TL;DR The short version

- An evaluation set is to an AI feature what a test suite is to ordinary code.
- Build it from real and awkward inputs, write down what a good answer must and must never contain, and include questions the system should refuse.
- Add guardrails around the model, not only in the prompt: validate output, filter unsafe content, limit tool permissions and log for review.

Read online, listen to it, or change the text size: https://blogs.iictp.org/how-to-evaluate-an-llm-application

Build the eval set from real use

1. Collect real or realistic inputs, including the awkward ones.
2. Write down what a good answer must contain and what it must never contain.
3. Add cases for each failure you find, so it cannot return unnoticed.

Ways to score

- **Exact checks** for structured output: valid format, required fields.
- **Reference answers** compared by a person or by another model.
- **Rubrics** for qualities such as tone, safety and use of sources.
- **Human review** of a sample, to keep the automated scores honest.

Hallucination tests

Include questions the system should refuse or say it cannot answer, and questions with a false premise. Check that answers stay within the supplied sources.

Guardrails

Add checks around the model, not only inside the prompt: validate output format, filter unsafe content, limit tool permissions, and log inputs and outputs for review. Run the eval set again whenever the model, the prompt or the data changes.

> Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. OpenAI. Evals guide. platform.openai.com/docs/guides/evals.

Published by IICTP, an independent technical credential body, on 26 October 2026. The current version of this post is at https://blogs.iictp.org/how-to-evaluate-an-llm-application. More guides: blogs.iictp.org.