DropTicks AI Article

AI Evaluation Is the Difference Between a Demo and a Product

Evaluation | Mar 10, 2026

How to create practical eval loops for prompts, RAG, agents, tool use, and production regressions.

Article

AI products fail quietly when teams do not measure them. A model response can sound correct while missing a fact, calling the wrong tool, leaking data, or ignoring business rules.

Evaluation should start before launch. Build a small dataset from real expected usage:

1. Normal tasks that should succeed.
2. Ambiguous tasks that should ask a follow-up question.
3. Forbidden tasks that should be refused or routed to a human.
4. Retrieval questions with known supporting evidence.
5. Tool-use tasks with expected arguments and side effects.
6. Regression cases from bugs and customer feedback.

For every AI workflow, decide which failures matter most. A support assistant needs correctness and tone. A sales workflow needs lead qualification accuracy. A coding agent needs tests and diff review. A financial or legal workflow needs strict escalation.

The best evals are not huge at first. They are targeted, repeatable, and tied to product behavior. Over time, production traces become new test cases.

DropTicks should treat evals as part of the delivery package. If the system matters, it needs a way to prove that changes improved it.

Web link

https://www.dropticks.com/articles/ai-evaluation-is-the-difference-between-a-demo-and-a-product

Related links

https://docs.ragas.io/en/stable/
https://arize.com/docs/phoenix
https://openai.github.io/openai-agents-python/