

Industrial RAG: how to measure answer quality

Evaluate answers from engineering manuals with source relevance, device context and appropriate abstention.

In short

- Retrieval success does not prove answer correctness, so each technical claim in a generated answer should be linked to a supporting passage.
- When no source covers a question, the right behavior is to name the evidence gap and ask for context, not to invent a parameter.
- A pilot acceptance record should freeze the evaluation set before changes and include definitions, comparisons, troubleshooting and out-of-scope requests.

01 / Define the engineering question

A request such as “Why does the drive stop?” lacks the product family, fault code and operating conditions. Ask for those details before proposing settings. A matching code from another device family can produce a plausible but technically wrong answer. Include the model, document version, symptoms and checks already performed in your NQ question.

02 / Evaluate retrieval and generation separately

Keep a test table with the question, expected document, supporting section and acceptable answer. First check whether the correct document appears in the first five results. Then link each technical claim in the generated answer to a supporting passage. Successful retrieval does not establish answer correctness. An obsolete manual may look relevant while being inapplicable.

03 / Include unanswerable questions

Test an absent product, conflicting revisions and a question missing its fault code. The expected behavior is to describe the evidence gap and request the missing context. Inventing an exact parameter from general model knowledge is not a successful answer. A retrieval score is a ranking signal, not a calibrated probability of correctness.

04 / Keep a pilot acceptance record

For a proposed pilot, include definitions, comparisons, troubleshooting and out-of-scope requests. Freeze the evaluation set before changes. Record the code version, collection version, latency and source links. Have two engineers review disputed answers. This is a suggested engineering workflow; it does not constitute a measured production accuracy claim.

Try it

Ask NQ the same question with and without a device model and revision. Compare which claims can actually be traced to a reference.

References and online guide

[Open the online guide](#)

[NIST AI Risk Management Framework](#)

Proposed pilot steps do not claim field accuracy. Synthetic examples are not measurements from a production site.