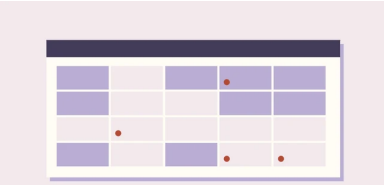
AI output evaluation set



We create 40 representative cases with scoring guidance and failure examples for one task.

Who it helps

Operations and systems owners

The problem to solve

A few successful examples say little about how an AI system behaves with ambiguous or unusual inputs.

Working through the brief

1. **Choose representative and difficult cases**

   Gather the agreed inputs. Approved examples, permitted source material and the named human reviewer.

2. **Write forty evaluated examples**

   Track “Evaluation case” and “Expected quality” together so the reviewer can connect the source information to the result.

3. **Document scores and failure types**

   Use the cases and scoring guidance to compare performance consistently and record failure patterns.

4. **Check the scope at handover**

   Platform, AI usage and connector subscriptions are separate.

Worked example

Include an ambiguous request in the evaluation set and agree what an acceptable request for clarification looks like.

Read the service scope / AUT-017

Sparkle Solutions Group Ltd.

info@sparklesolgroup.com +44 20 3307 6468 sparklesolgroup.com

71-75 Shelton Street, Covent Garden, London, WC2H 9JQ, United Kingdom