AI PILOT AND EVALUATION DESIGN / ILLUSTRATIVE CASE

What evidence would justify expanding an AI pilot?

The management decision

Decide which mistakes matter most and who can recognise them. Agree a separate test for questions a tool should refer to an expert, especially when the source is incomplete.

A demonstration can make an AI tool look reliable while hiding mistakes on important questions. If staff use an outdated answer or accept a reference they cannot verify, any time saving can turn into extra checking or rework. A focused pilot helps decide what the technology can support and where a person still needs to make the judgement.

Worked example

A 40-question trial includes 30 questions the tool should answer and ten it should refer. It answers 27 of the 30 correctly and refers eight of the ten appropriately. Review the three missed answers and two missed referrals separately.

Example figure	What it represents
40	Test questions
27/30	Correct answers
8/10	Correct referrals

What changes the answer

Four plausible answers might point to an outdated document. Score those separately from answers with a verified source and note the consequence if staff followed an outdated instruction.

AI pilot and evaluation design

What to commission

Design a trial for one task using approved reference material, representative questions and questions that require referral.

We define the intended users and the boundaries of the task. We design examples that a tool should answer, questions it should pass to a person and situations where an answer based on an old document would fail.

How to judge the result

The sponsor can separately count correct answers, proper referrals, unsupported answers and outdated references before expanding use.

Estimate expert review time alongside time users recover. A pilot that handles routine questions well can still need restrictions on high-consequence decisions.

Decision worksheet

Project detail	Starting point for your team
Decision owner	AI pilot sponsor and subject expert. Name the person who will approve the next step.
Records to gather	A specific task, approved information and examples of answers that would be unacceptable.
Agreed output	A pilot brief, scored test cases and practical rules for human review.
Decision date	Record the deadline, any unresolved assumption and who will confirm it.

Discuss this case

Tell us what task you would like to improve and who would use the answers. Give one example of a helpful response and one error you could not accept. Mention who can review answers for accuracy and whether there is a decision date for the pilot.

Email info@oxetc.co.uk and mention IT-15. All figures in the worked example are illustrative.