The AI Pilot Report Template

The 6-field, 3-page report that turns an AI pilot on your sample data into a business decision — accuracy and cost per use, logged from day one.

A polished demo on clean sample data proves nothing: the pilot lands on your real documents — inconsistent formatting, scanned PDFs, three naming conventions for the same client — and accuracy collapses while the API bill climbs. This template closes that gap. It is the report structure used for the two highest-payback pilot shapes — permissioned **document search** and **inbox triage** — run on a representative sample over 2–4 weeks. Two numbers decide everything: **how often the system is right on your data**, and **what each correct answer costs**.

> **One-task discipline:** one task, one dataset, one scoreboard. Run one pilot, not two — two pilots at once split attention and produce two inconclusive reports instead of one decision.

Section A — Pilot one-pager (fill in before kickoff, both sides sign)

Task (one sentence)	.....
Dataset (few hundred documents / few thousand messages; access model cleaned first)	.....
Users who will judge it	.....
Accuracy target (e.g. 85%+ cited accuracy)	.....
Cost ceiling (per resolved query or per week)	.....
Window (2–4 weeks) and **decision that follows**	.....

Anything outside this page is a different pilot, priced separately. If neither pain is sharp, say so — a pilot without a sharp pain measures nothing worth paying for.

Section B — Grading set (agreed before the pilot starts)

A fixed set of **50–100 real queries with known-good answers**. Every system answer is graded against it: **right with citation**, **right without citation**, **wrong**, or **refused**. Citations are the load-bearing column — an answer a human can verify in ten seconds is a feature; an answer without one is a rumour with good grammar. Spot-check at least **20% of grades by hand**, especially confidently-wrong answers.

#	Query	Known-good answer / passage	System answer	Grade
1	.....	.....		
2	.....	.....		

#	Query	Known-good answer / passage	System answer	Grade
3	.....	.....		
…	continue to 50–100 rows; attach the full set as an appendix			

Section C — Accuracy table (log weekly, not just at the end)

Grade	Count	%	Notes / trend vs last week
Right with citation			
Right without citation			
Wrong			
Refused			
Human spot-check rate (target ≥20%)	—		

> **Design targets we set with clients** (not a published industry benchmark): **85%+ cited accuracy** on a bounded task is a green light; **below 70% after two tuning rounds** means the task, the data or the model choice is wrong — the report should say which.

Section D — Cost table (from day one)

Measure	Week 1	Week 2	Total
API spend (per-token + retrieval costs)			
Queries resolved			
Cost per resolved query			
Tending minutes (checking, correcting, maintaining)			
Corrected cost = API spend + human cleanup cost			
Current staff cost of the same task (baseline)			

Failure economics: wrong answers cost double — API spend plus human cleanup. A system that is cheap per query but often wrong is more expensive than doing nothing. **Project to production volume before deciding** and state the scaling assumption in writing — token costs and tending minutes rarely scale linearly.

Section E — Kill criteria (agreed in writing before the pilot begins)

- Cited accuracy below the agreed floor after two tuning rounds
- Corrected cost above the current staff cost of the task
- Tending time that never falls

- Sample proves unusable (access rules too tangled, documents too inconsistent) — fix the data or stop honestly
- The team stops checking outputs — trust without verification is how confident errors reach customers

Section F — Recommendation & conditions

Verdict	What it means
Scale	Accuracy at or above target, corrected cost comfortably below staff cost, and a named owner for tending
Adjust	One specific change with a re-measurement window — different chunking, tighter scope, better source documents
Stop	The numbers say no, with the reasoning preserved so nobody re-runs the same pilot in a year

Conditions: state what would change the answer (volume thresholds, data quality, model changes).

Appendix — attach the raw logs

- The full graded query set with known-good answers
- Weekly accuracy and cost tables
- The list of confidently-wrong answers and what caused each

Circulate the report to everyone who will live with the decision — especially the sceptics. A pilot report that only the enthusiasts read is marketing; one the sceptics accept is an asset. Their sign-off is the real deliverable.

> **A successful pilot is one you can decide from** — a written go/no-go with the numbers attached is a success even when the answer is no. A small fixed fee that buys a documented no beats a large wrong build.

Want this run on your documents with a fixed price? Start with a short discovery: **www.anyplacehub.com/contact/** or **contact@anyplacehub.com**. Related reading: “What a Real AI Pilot Report Looks Like” — www.anyplacehub.com/insights/real-ai-pilot-report-cost-accuracy/