WHAT CHANGES IN THE WORK?

Expert judgement needs better-prepared work.

Your finance team already knows how to reconcile, question and decide. Agentic AI could change what reaches their desk: a draft working pack with reconciliations assembled, evidence attached and differences explained or flagged.

Follow one corrected source.

WORKING PAPER / 01

Synthetic GSTR-9C example · Aster Components · FY 2024–25 · All amounts in ₹ lakh

PREPARATION PEOPLE CURRENTLY COORDINATE

Find the schedule. Check the version. Locate the adjustment. Update the working. Reconcile to the return. Revisit the review.

AGREED SOURCE · UB-01

Opening unbilled revenue schedule

Project	v1	v2
A	14	14
B	~~4~~	10
Total	~~18~~	24

v1 retained — v2 current

UB-01 / v2 / total ↗ 5B

Agentic AI
Evidence & mapping

Locate agreed evidence; propose its mapping. Record source, version, change and reviewer questions.

CALCULATION TOOLS — DRAFT / ₹ LAKH

GSTR-9C · Table 5 working

Row / description		Prior	Revised
5B	Opening unbilled revenue	18	24
5P	Adjusted annual turnover	1,231	1,237
5Q	Turnover declared in GSTR-9	1,237	1,237
5R	Unreconciled turnover Q − P	6	0

1,250 (5A · books) + 24 (5B) − 6 (5E) − 31 (5H) = 1,237 (5P)

Programmed calculations update the working. The **+6 source correction** is separate from the unchanged 5E credit-note deduction of 6; 5H remains 31.

YOUR TAX REVIEWER

Reopen the affected review.

Acceptance outstanding

Confirm the corrected schedule, period and adjustment. Your reviewer accepts or challenges the evidence, resolves exceptions and approves treatment; decisions and reasons must be preserved. **Zero difference alone does not establish the right treatment.**

Illustrated method, not an AI run or completed review. Selected rows and assumed treatment; not the complete prescribed form. Other adjustments are unchanged.

Where the effort goes

EY puts **53%** of internal tax personnel's time in routine compliance: data collection, workpapers and reconciliation. [1] This is workload, not time AI can save.

Where judgement stays

FRC found AI use concentrated in extraction, consolidation, reconciliation and validation, with caution around judgement. [2] Reported practice, not causal productivity gains.

1. EY · 2025 Tax and Finance Operations Survey, p. 13. Global workload; not GSTR-9C-specific or an automatable share.
2. FRC / Lancaster · AI technologies in corporate reporting (2026). UK: 39 interviews, 103 survey responses; not a controlled trial.

WHY THIS MATTERS BEYOND ONE RETURN

Test a capability your finance team can reuse.

GSTR-9C is a bounded proving ground: evidence spread across files, repeated preparation, a prescribed output and results a tax professional can inspect.

LOOK FOR THE SAME PATTERN ELSEWHERE



Fragmented evidence
Facts across files

Working / current period v2

Source + version	Calculation	Review note
UB-01 / v2	5B · 24	Reopened
Prior version retained		New input → affected working

Recurring preparation
Periods repeat; inputs change

Defined outputs
A known working to complete

Reviewable results
Numbers, evidence, questions

Plausible output is not enough.

FINANCIAL AUDIT BENCH [3]

>85% Applicable checks passed
On average, for each model

69% Task attempts fully passed
Best model; every applicable check correct

Different denominators: individual checks versus complete task attempts. These are not two scores for the same measure.

FinBalance found plausible balances alongside incorrect source links and inconsistent ledger calculations. [4] Both are controlled, synthetic benchmarks—not BluePi or production evidence. Can a reviewer verify the whole working, including what the system missed?

What a BluePi evaluation should establish

REQUIREMENTS, NOT FEATURE CLAIMS

Trace it.	Can a reviewer reach the exact source and version behind a figure?
Check it.	Do programmed calculations reconcile, with missing evidence and unresolved treatment visible?
Change it.	When agreed inputs change, are affected workings refreshed and prior decisions flagged for review?
Review it.	Does gathering and rechecking fall without more review effort? Measure both against today's process.

The sample illustrates a pack and correction trail, not production automation, measured savings or completed customer review. Success here would not prove another finance process.

The bigger opportunity

If these capabilities hold up, examine close reconciliations or reporting schedules. More frequent reconciliation may help investigate differences earlier. Each application needs its own evidence, controls and business case.

TAKE THIS TO FINANCE AND TECHNOLOGY TOGETHER

Which recurring process would benefit most from a prepared, traceable working pack—and what would convince us it is reliable enough to use?

3. Modus Labs · Financial Audit Bench (Sep 2026). 45 synthetic tasks, 11 models, eight attempts per task; partial checks ≠ strict task completion.
4. FinBalance · Multi-Document Accounting Reconciliation Benchmark (Jun 2026). Synthetic source-linking and ledger tests.