THE AI ACCOUNTANT
Your Practice, powered by AI

FREE RESOURCE

The Vendor Test Pack

Three operational tests every CAS firm needs before the next AI vendor breaks something quietly.

Drift detection. Portability evaluation. Version trail. An afternoon to set up. Five minutes a week to run.

Peter McCarroll, CPA — **The AI Accountant**

theaiaccountant.ai

The companion piece to *Three tests every firm needs before the next AI vendor breaks something quietly*. The article makes the case. This pack gives you the prompts.

Inside:

1. **Prompt 1 — The rubric generator** — paste your workflow in, get a scoring rubric out
2. **Prompt 2 — The cross-LLM scorer** — score new outputs against the rubric using a *different* AI model than the one being tested
3. **The version-trail line template** — one line in your workpaper, full accountability
4. **A worked example** — monthly close commentary for a recurring SaaS client, all three tests applied

Each prompt is plain text — paste directly into Claude, ChatGPT, Gemini, Copilot, or any frontier LLM. The version-trail template plugs into your existing workpaper or workflow log without any tooling change.

Prompt 1 — The rubric generator

Use this prompt to produce a scoring rubric for one workflow your firm runs repeatedly with AI. Run it once per workflow; save the rubric.

```
You are an experienced CAS practice quality reviewer. I'm going to give you a workflow my
firm runs repeatedly with AI assistance. Your job is to produce a scoring rubric I can use
to evaluate whether the AI's output is good.

Here's how this rubric will be used: once a week, I'll run the workflow against a stable
input. Then a different LLM will score the new output against this rubric. The rubric is the
trust mechanism — it has to be specific enough that scores are stable when the model is
stable, and sensitive enough to drop visibly when the model degrades.

==========================
WORKFLOW INPUTS (I fill these in)
==========================

WORKFLOW: [Describe in one sentence — e.g., "Monthly close commentary for a recurring SaaS
client"]
TYPICAL INPUT: [What you feed the AI — e.g., "TB export, prior-month commentary, and the
client's KPI list"]
TYPICAL OUTPUT: [What the AI produces — e.g., "3-paragraph variance commentary covering
biggest swings, cash position, and operating-margin movement"]
KNOWN-GOOD EXAMPLE: [Paste a partner-approved past output that represents the quality bar]

==========================
PRODUCE THE RUBRIC
==========================

Build the rubric with these properties:

- 4 to 6 dimensions, each scored 0–10
- Dimensions specific to this workflow, not generic. "Did the variance commentary explain
the three biggest swings?" — not "Is the writing clear?"
- Each dimension has a one-sentence definition
- Each dimension has a scoring guide (0 = absent or wrong; 5 = present but weak; 10 = sharp
and complete)
- A weighting rationale — which dimension matters most, and why
- A target overall score for this workflow (e.g., "below 35 out of 50 needs a partner re-
pass")

Output the rubric in markdown. After the rubric, briefly explain your weighting choices.
```

How to use it: Run this prompt once for each workflow you want to monitor. Save each rubric in a master rubric file alongside your prompts. Most firms start with one workflow; add rubrics over time.

Prompt 2 — The cross-LLM scorer

Use this prompt to score the AI's new output against the rubric. **Run this in a different LLM than the one that produced the output.** That cross-LLM step is what makes the score trustworthy. If you score

the model with itself and the model has degraded, the scoring degrades with it.

```
You are an independent reviewer. I'm going to give you (1) a scoring rubric and (2) an
output produced by a different AI model that needs to be scored.

Important: you are NOT the AI model that produced this output. You are a critical reviewer.
Score the output strictly against the rubric. Don't be charitable — small drops in quality
are exactly what we're trying to catch.

==========================
INPUTS (I paste these in)
==========================

RUBRIC: [Paste the rubric produced by Prompt 1]
OUTPUT TO SCORE: [Paste the new output you want to score]

==========================
SCORING TASK
==========================

For each dimension in the rubric:

1. Quote the specific evidence from the output that maps to this dimension. (If the
dimension's evidence is missing, quote that fact — "the output does not mention X.")
2. Score the dimension 0–10 based on the rubric's scoring guide.
3. Note any qualitative concern that the score doesn't fully capture.

After scoring all dimensions, produce:

- TOTAL SCORE — the sum of dimension scores
- STRONGEST DIMENSION — one sentence on the output's best dimension and why
- WEAKEST DIMENSION — one sentence on the output's worst dimension and why
- THRESHOLD FLAG — does the total fall below the workflow's target? If yes, say so plainly
and recommend a re-pass

Be specific. Do not hedge. If the output is bad, say it's bad with the score that reflects
it.
```

How to use it: Once a week, run your stable input through the workflow's standard AI tool. Take the new output and run it through this scorer in a *different* LLM. Save the score with the date and model version. After three to four weeks, you have a baseline. After eight weeks, you can spot a drop the moment it happens.

Cross-LLM pairings that work well:

- Workflow runs on Claude Cowork (Opus 4.7) → score with GPT-5.5
- Workflow runs on ChatGPT (GPT-5.5) → score with Claude Sonnet 4.6
- Workflow runs on Gemini → score with Claude or GPT
- Workflow runs on Copilot → score with Claude or Gemini

The pairing matters less than the principle: a different model, a different vendor, a different harness. The point is to break out of the testing-self trap.

The version-trail line template

One line in your workpaper, your workflow log, or your file metadata. The difference between defensible AI use and "we used AI somewhere."

Template:

```
AI assistance: [Tool] on [Model], prompt v[N], run [Date]
```

Examples:

- `AI assistance: Cowork on Opus 4.7, prompt v3, run April 25, 2026`
- `AI assistance: ChatGPT on GPT-5.5, prompt v1, run May 1, 2026`
- `AI assistance: Claude API on Sonnet 4.6, prompt v2 (drift-tested), run April 30, 2026`

Field notes:

- **[Tool]** — the application surface (Cowork, ChatGPT, Gemini Enterprise, Copilot, custom API, etc.)
- **[Model]** — the specific model version (Opus 4.7, GPT-5.5, Sonnet 4.6, etc.). When the model changes mid-engagement, the line changes.
- **[Prompt vN]** — your internal version number for the prompt or workflow instruction. Track these in a master prompt file with a one-line changelog so v1, v2, v3 mean something. If the prompt changed mid-engagement, the version number changes.
- **[Date]** — the date the AI portion was run, not the date the partner reviewed.
- **(drift-tested)** — optional tag indicating the workflow was scored against the rubric within the last seven days. The tag tells future-you that the model was confirmed working at the time.

Where to put the line:

- Workpaper cover sheets or footers
- Workflow logs in your practice management system (Karbon, Financial Cents, custom)
- File metadata or naming conventions for AI-touched deliverables
- Engagement files, in a dedicated "AI assistance log" tab

Why this matters: When the model changes — and it will change every six weeks now — your sign-off doesn't. The accountability stays with the partner. The version trail is what lets future-you (or a regulator, or a litigator) reconstruct what the model was when the work was done. Without it, you've taken on a risk you didn't price into the engagement.

A worked example — Monthly close commentary for a SaaS client

This is what all three tests look like applied to one workflow. Adapt the specifics to your firm.

The workflow

A recurring SaaS client (subscription business, ~$8M ARR, monthly close on the 7th business day). Every month, the prep manager produces a 3-paragraph variance commentary for the client's CFO covering the top swings, cash position, and operating-margin movement.

Input to the AI:

- Trial balance export (current month, prior month)
- Last month's commentary (for tone and structure reference)
- The client's top 5 KPIs (MRR, gross retention, CAC, payback period, gross margin)
- A short instruction: "draft this month's commentary in the same structure as last month, focusing on variances over $25,000 or 10%"

Output from the AI: A 3-paragraph commentary the manager reviews, edits, and sends.

The rubric (Prompt 1 output)

Five dimensions, scored 0–10 each. Target ≥ 38/50 for partner acceptance without re-pass.

#	Dimension	Definition	Weight
1	Variance accuracy	Does the commentary correctly identify the three biggest swings by dollar value, with correct numbers?	High — wrong numbers are an immediate fail
2	Cash position narrative	Does the commentary state the closing cash position and explain the change since prior	High

#	Dimension	Definition	Weight
		month using actual numbers?	
3	Operating-margin story	Does the commentary land the operating-margin movement in fewer than four sentences with a named cause?	Medium
4	Tone match	Does the commentary read in the same voice as last month — same sentence shape, same level of formality, same use of client-specific shorthand?	Medium
5	Anomaly flagging	Does the commentary flag any line item that moved >25% without explanation? (Even if the cause is unknown, the line should be called out.)	High — missed flags become client questions

Weighting rationale: Variance accuracy and cash position are partner-fail criteria — anything below 7/10 on either gets a re-pass regardless of total. Anomaly flagging is the canary; missed flags are how the client catches you, not the other way around.

Target threshold: ≥ 38/50 overall AND ≥ 7/10 on dimensions 1, 2, and 5.

A scored output (Prompt 2 output, abbreviated)

Run the workflow on the standard input. Paste the rubric and the new output into a different LLM. Get something like:

```
DIMENSION 1 — Variance accuracy: 9/10
Evidence: "MRR rose $48,200 to $677,400" — matches the TB ($48,217 actual).
"Gross retention dipped to 91.2% from 94.1%" — matches.
"Sales & marketing spend up $112,000 vs prior month" — matches.
Concern: rounding to $48,200 is fine; tone is consistent.

DIMENSION 2 — Cash position narrative: 8/10
Evidence: "Closing cash $1.84M, up $112K from prior month."
Concern: doesn't explicitly tie the cash movement to the operating-margin story; one beat
short.

DIMENSION 3 — Operating-margin story: 7/10
Evidence: "Operating margin compressed 180 bps to 14.2% on the S&M timing."
Concern: "S&M timing" is generic — last month's commentary said "Q2 sales hire phasing"
which is the actual cause. Voice drift.

DIMENSION 4 — Tone match: 8/10
Evidence: Sentence shapes match. "MRR rose" / "GM dipped" / "Spend up" — same staccato.
Concern: missing "as expected" or "ahead of plan" tags that the client's CFO uses internally
— this is in last month's reference.

DIMENSION 5 — Anomaly flagging: 6/10
Evidence: Flagged the S&M variance.
Concern: did not flag the $14K spike in "Software subscriptions" line, which is 32% above
prior month.

TOTAL: 38/50
THRESHOLD FLAG: At threshold but dimension 5 is at 6/10 (below 7/10 minimum). RECOMMEND
PARTNER RE-PASS.
```

This is what a healthy week looks like — output is good, scorer catches a real concern, partner re-passes for the flagged line. The score becomes the baseline.

What drift looks like

Two scenarios where the rubric would catch a problem the partner might not:

- **Drift in dimension 1 (variance accuracy):** scores were running 9/10 for six weeks, then drop to 6/10 in one week. The model is suddenly hallucinating numbers slightly. The rubric catches it; the partner can pull the prep before it goes to the client.
- **Drift in dimension 4 (tone match):** scores were running 8/10, then drop to 5/10. The model has stopped reading the prior month for tone — it's producing generic SaaS commentary instead of this client's voice. The partner pulls the prep, fixes the prompt, restores quality. Without the rubric, this gets noticed by the client three months in.

The version-trail entry for this workflow

Each month's commentary closes with this line in the workpaper:

```
AI assistance: Cowork on Opus 4.7, prompt v3, run May 5, 2026 (drift-tested May 4 — score
38/50, accepted with re-pass on dimension 5)
```

Three pieces of information that survive any model change: which model produced the work, which prompt version, what the drift test said within seven days of the run. If a year from now a question comes up about this commentary, the partner has the trail.

What to do this week

You don't need all three tests running by Friday. You need one workflow with one rubric and one weekly scoring run. The other two tests follow naturally:

1. **Today:** pick the workflow. The most repeatable, highest-volume one your firm runs with AI is the right starting point.
2. **This afternoon:** run Prompt 1 in your standard AI tool. Save the rubric.
3. **Friday:** run the workflow. Take the output and run it through Prompt 2 in a different LLM. Save the score.
4. **Next Friday:** repeat. After three or four weekly scores, you have a baseline.
5. **Add the version-trail line to the workpaper for any AI-touched deliverable starting this week.**

Once that's running, build the second rubric. Then the third. Within a quarter, you have a quality system that survives any model the vendors ship next — and you have an instrument that catches degradation before your clients do.

A note on this pack

The prompts above are starting points. The rubric prompt is the most likely to need adjustment for your firm — not all CAS workflows look like SaaS variance commentary, and a rubric for a tax-position memo will have different dimensions than a rubric for a payroll reconciliation. Run the prompt, read the rubric it produces, and edit the dimensions until they reflect what *your* partners would catch. The prompt doesn't need to be perfect on the first try; the rubric does.

If you find a better way to structure any of this, send it. peter@theaiaccountant.ai. The next version of this pack will incorporate what's working in the field.

— Peter McCarroll The AI Accountant