

INTERNAL COACHING DISCUSSION FORUM

Discussion Paper 02

Can AI Assess Coaching Better Than Humans?

A note on perspective: this analysis draws on evidence from building one of the tools discussed. That stake is disclosed where it appears.

Tom Battye

EMCC Global Master Practitioner | Accredited Supervisor | EMCC Global Supervision Award Winner

June 2026

tombattye.co.uk

In many ways, yes.

In one study using the Ovida platform, coaches asked to estimate how much of a session they spent talking under-estimated by 27 per cent on average. Not a small margin: a quarter of the talk-time picture, missing, from the person who was in the room. The machine does not estimate. Provided the transcript and classification rules are sound, it counts.

AI is already better than most humans at the parts of coaching assessment that matter for fairness and consistency. It has limits at the parts that depend on relationship, context and the being of the coach, limits that need to be honestly scoped rather than hidden. Within those limits, the combination of AI and human expertise, whether applied per-session or built into the architecture, is better than either working alone.

What AI does better

It catches everything. A human assessor listening to a 47-minute session cannot extract, transcribe and classify every question without fatigue or drift. The machine applies the same rule to question 3 and question 58 with exactly the same attention. No recency bias. No Friday-afternoon fade.

It sees patterns easy to miss in the flow of conversation. In one session, the coach was warm, supportive, mostly non-directive, low talk-time. A good session by any casual measure. But the full question distribution told a different story. Evaluative praise, hedging before every challenge, repeated thanking, checking questions that were grammatically closed. Each unremarkable on its own. Together, a coherent deferential stance: comfortable but not accountable. An experienced assessor can spot the pattern. The machine's advantage is assembling the evidence quantitatively: every instance counted, every example timestamped, the whole distribution laid out.

It is the same assessor every time. Two human assessors disagree. The same assessor disagrees with themselves across sessions. AI applies identical criteria to every session it reads. For fairness across a cohort, that consistency is an advantage humans cannot match.

It makes feedback accessible. Until recently, coaches received structured feedback on their practice in two contexts: submitting a recording for reaccreditation or sharing one with a supervisor. Feedback was tied to formal processes and limited by assessor availability. AI tools are changing this at two levels. For evaluators, platforms like Ovida and RaeNotes report saving roughly two hours per session review, making supervision faster and more consistent. The larger shift is direct access: tools that deliver feedback to the coach without requiring an evaluator in the workflow for every session.

What stays hard

AI has no relationship with the person receiving the feedback. Michael Hudson's observation at the 2026 Digital and AI Coaches' Conference captures it: AI has no skin in the game. A human supervisor can pace a difficult finding, notice how it lands, and moderate for the person in front of them. The machine cannot. It delivers the same feedback regardless of who is reading it.

It reads the *doing*, not the *being*. Two coaches can produce near-identical transcripts, one from unconscious mastery and one from well-drilled technique. The page cannot tell them apart. ICF's own framework distinguishes ACC, PCC and MCC partly by saying the ACC coach works on what the client needs to do, the PCC coach works on who the client is becoming, and the MCC coach makes that distinction disappear. A transcript holds the first two with decreasing reliability. It does not hold the third.

This is a structural limit, not a technology gap. The strongest AI assessment systems scope their output honestly: cap the formal level determination at Senior Practitioner / PCC, and acknowledge that anything above that boundary depends on dimensions a transcript cannot hold. The same limit applies to contextual fit. Whether a particular intervention was the right move for that client, at that moment, depends partly on the relationship and the context, which the transcript only partially reveals.

The market now

Ovida and RaeNotes are the two established specialist platforms. Both have made a deliberate choice to keep the evaluative judgment with the human. RaeNotes describes its AI as an assistant, not a substitute, and says it will not auto-generate a review. Ovida calls its AI feedback descriptive, not evaluative, a 'reflective mirror, not a judge,' explicitly not for credentialing decisions. That restraint may reflect an ethical position as much as a technical one, and it deserves respect.

CoachInsights is ours, and it takes a different position. Rather than surfacing evidence for a human to interpret, it produces a full evaluative report delivered directly to the coach. The feedback is direct by design. The machine has no social stake in the conversation, so it does not soften a finding to preserve rapport. That directness comes at a cost: the machine cannot pace a difficult finding or notice how it lands, and a level-of-practice determination is exactly the kind of feedback where relational sensitivity matters. For coaches who want unfiltered feedback on their practice, that trade-off is worth making.

What the coach receives is a structured report covering each competency dimension in turn: contracting, session structure, non-directive approach, listening, holding insight and emotional shift, questioning, support and challenge, and relational presence. Each section cites specific moments from the session with timestamps and verbatim quotes, identifies patterns across those moments, and offers developmental recommendations grounded in the evidence. The report closes with a question distribution (the percentage of open, closed, leading and other question types) and an advisory level of practice.

Specific design choices address specific known risks. The counting is verified in a separate stage before the narrative is written, so the numbers that shape the report are audited, not estimated. Contextual evaluation is built into the report generation: a directive intervention that served the client is credited as a strength, not flagged as a deficit. The level determination is capped at Senior Practitioner / PCC, an honest scope that reflects what a transcript can and cannot support. The design encodes the human expertise rather than requiring a human to apply it per-session.

When CoachInsights and RaeNotes analysed the same session independently, both converged on the same primary developmental finding. The evidentiary detail differs: CoachInsights produces quantified, timestamped, verbatim analysis; RaeNotes produces qualitative, paraphrased summary. The developmental reading was the same.

What this means for coaches in organisations

Coaching development has a feedback problem. Hawkins and Schwenk (2006), surveying 120 organisations and 530 coaches for the CIPD, found that 88% of coaching organisers and 86% of coaches believed coaches should have regular supervision, but only 44% were receiving it and only 23% of organisations provided it. Among those who did receive supervision, 58% had only started within the previous two years. Even where supervision exists, it is typically monthly or quarterly. The gap between practice and feedback is long, the feedback is filtered through memory, and the opportunity for rapid, targeted improvement is limited. Most internal coaching functions have no established mechanism for regular feedback on actual practice.

AI assessment closes that gap. A coach can receive detailed feedback within minutes of a session ending, review it while the session is still fresh, identify specific adjustments, and apply them in the next session. Musicians practise with recordings. Athletes review footage between sessions. Coaching has lacked an equivalent mechanism until now. The tight feedback loop, coach, receive feedback, adjust, coach again, is well established in other domains as a powerful accelerator of skill development. Huthwaite's research into the SPIN selling model found that without follow-up coaching, 87% of the training benefit was lost (Rackham, 1988, cited in Hawkins, 2012). The same principle applies to coach development: skills acquired in training decay without timely, practice-connected feedback.

This is not a replacement for supervision. Supervision serves a broader and deeper purpose: navigating relational dynamics, processing emotional responses, working through ethical dilemmas, developing the capacity for the kind of presence AI cannot assess. When the technical feedback is handled before the supervision session, the supervisor has more space for that deeper work. AI and supervision serve different levels of the coaching competency spectrum. Both are needed.

Know what the system can and cannot do. If your organisation is considering an AI assessment tool, Susan Caesar's five questions apply: What is it? What can it not do? How does it make its decisions? What evidence supports its claims? How is data protected? The answers matter more than the marketing. AI assessment tools need the same quality assurance as human assessors: moderation, verification, consistency checks and continuous improvement. The mechanism is different, but the principle is the same.

The future

Accountability, contextual judgment at the edges, the being of the coach: these are structurally outside what a machine reading a transcript can do. Richer data (voice analysis, non-verbal signals, eventually micro-expressions) will make the surfacing layer better. It will not give the machine skin in the game.

The competitive pressure on specialist assessment tools comes less from each other than from the general-purpose models that power them. Any coach can already upload a transcript to ChatGPT, Gemini or Claude and ask for feedback. The underlying models will keep improving, and that improvement benefits everyone. What differentiates a specialist tool is not the engine but the instrument built around it: a secure file system for recordings, structured workflows for consent and data protection, and the capability to track the same coach's development across sessions and clients over time. Longitudinal analysis, not just single-session feedback, is where the platform earns its place over the generic model.

The line

Coaching has had a structural feedback gap for as long as the profession has existed. For the first time, there is a practical mechanism to close it.

Questions for discussion

1. How do the coaches in your organisation currently receive feedback on their practice?
How timely and specific is it?
2. If an AI tool identified a pattern in a coach's practice (say, a deferential stance built from evaluative praise and closed checking questions), how would that coach likely respond? How would it differ from hearing the same finding from a supervisor?
3. Where would you draw the line between what AI should and should not assess in coaching? Are there dimensions where AI assessment would feel inappropriate?
4. How might AI assessment change the role of supervision in your organisation? Would it free supervision to focus on deeper, relational work?
5. What would your coaches need to see before they trusted AI feedback on their practice?

References

- Hawkins, P. (2012) Creating a Coaching Culture. Maidenhead: Open University Press.
- Hawkins, P. and Schwenk, G. (2006) Coaching Supervision: Maximising the Potential of Coaching. London: CIPD.
- Rackham, N. (1988) SPIN Selling. New York: McGraw-Hill.
-

About the author

Tom Battye is an EMCC Global Master Practitioner, Accredited Supervisor, and EMCC Global Supervision Award winner. He has spent 25 years training internal coaches across the public and private sectors. He is the founder of Tom Battye Coaching and the creator of CoachInsights, an AI-powered coaching assessment platform. His programmes include the Certificate in Executive Coaching (EMCC Practitioner, AC accredited) and the Leader as Coach programme (EMCC Foundation, AC accredited).

tombattye.co.uk | tom@tombattye.com

About this series

This paper is part of the Internal Coaching Discussion Forum, a six-session discussion series for L&D, OD, and talent leaders running from May to November 2026. Each session explores a question about what it takes to create a coaching culture, drawing on published research, named case studies, and 25 years of practical experience. Format: 20 minutes of content, 40 minutes of discussion, via Zoom. Details and registration at tombattye.co.uk/forum.