

# AI Output Quality Rubric Evaluation Report

Thought leadership article — AI and professional assessment (v0.1 draft)

## Evaluation Parameters

Review type	Full All seven dimensions
Standard	Mid-development Developmental framing; feedback oriented toward next steps
Confidence (stated)	Not confident Structural integrity examined across all dimensions
Output type	Thought leadership article (v0.1 draft) AI and professional assessment — practitioner audience
Rubric version	1.1 Field testing
Date	16 April 2026

## Dimension Scores — Summary

Dimension	Score	Band
Fit to Context	6/10	Adequate
Evidence and Grounding	6/10	Adequate
Analytical Depth	7/10	Capable
Purposeful Structure	6/10	Adequate
Appropriate Register	6/10	Adequate
Critical Integrity	7/10	Capable
Evaluative Judgement	6/10	Adequate
AI Voice (Holistic)	4/10	High AI presence

*Overall pattern: Analytical Depth scores Capable; all remaining dimensions score Adequate. The AI Voice holistic measure scores High AI presence. The output is coherent and substantively strong in its central argument at v0.1, but does not reach Capable in the majority of dimensions, principally because the evidence architecture is shallow, structural consolidation is needed, and AI voice is significantly above the standard for thought leadership work of this kind.*

## Dimensional Evaluations

### Dimension 1: Fit to Context — Score: 6/10 — Adequate

#### Strengths

The piece broadly targets the right audience. The elephant section handles the naysayer respectfully rather than dismissively, which is the correct register for a practitioner-facing thought leadership piece. The decision to frame the central argument around what assessment is for, rather than whether AI belongs in it, is well-calibrated for an audience that will resist technology-first framing.

#### Weaknesses

The skills-based practitioner audience is still underserved. The skills assessment section was identified as a gap late and added rather than woven in. A vocational assessor reading the piece will feel addressed in that section but slightly excluded from the rest of it. The learner experience thread, agreed as woven texture, appears in clusters rather than consistently throughout. Two or three sections pass without any learner-experience grounding, which weakens the human thread the piece needs.

The "Tools That Already Work This Way" section is borderline on the thought leadership positioning. The current framing uses "existence proof" as its justification for inclusion. That is a thin line and in the editing pass it needs to either earn its place more clearly or be reduced to a single sentence with a link.

#### Area for Development

The editing pass needs to audit every section for skills-based practitioner accessibility, not just the dedicated section. The learner experience thread needs redistributing rather than concentrating. The tools section needs a clearer rationale for its presence or a reduced footprint.

---

### Dimension 2: Evidence and Grounding — Score: 6/10 — Adequate

#### Strengths

The Kosmyna MIT study is the piece's best evidential moment. The nuance between passive and active AI use is drawn correctly and does genuine argumentative work. The T Levels, V Levels and apprenticeship reference is grounded in real policy direction rather than assertion. Boud and Sadler are deployed at the right points in the argument.

#### Weaknesses

Several references are name-dropped without their findings doing argumentative work. Hattie is the clearest example. "Hattie's extensive meta-analysis has consistently demonstrated" is followed by a claim about agency and self-regulation that is reasonable but not precisely sourced to anything specific in Hattie's body of work. Luckin is introduced and her position summarised so loosely that the reference adds little beyond a named authority. Sparrow appears in the reference list but is not deployed anywhere in the text.

The gaming counter section is the piece's most assertion-heavy passage. The claim that grading design alongside product inverts the incentive structure is analytically plausible but it is not evidenced. It is presented as if the logic alone is sufficient. For a practitioner audience that will probe implementation, that is not enough.

#### Area for Development

Each named reference needs a specific finding deployed, not just a name and a general characterisation. The gaming counter needs either an empirical anchor or an honest acknowledgement that this is a reasoned proposition rather than an established finding. Sparrow should either appear in the text or be removed from the references.

**Dimension 3: Analytical Depth — Score: 7/10 — Capable****Strengths**

This is the piece's strongest dimension. The central insight — that AI exposed rather than created the proxy problem — is genuinely non-obvious and is well-sustained across the early sections. The inversion of the gaming incentive is analytically sharp. The decision to position the doctoral model as destination rather than starting point shows structural analytical intelligence. The formative to summative bridge is the most developed analytical section in the piece and makes a claim that most practitioners will not have encountered in this form.

**Weaknesses**

The piece loses analytical momentum in its second half. The "What This Makes Possible" section drops into relatively predictable benefits language. "From passive subject to active designer" is a phrase that circulates in education discourse already. The risks section names the risks without interrogating them. The equity concern in particular deserved more analytical weight, both because it is a genuine vulnerability in the model and because the audience will raise it. The conclusion restates the argument rather than landing a final insight the reader has not already encountered.

**Area for Development**

The second half needs the same analytical standard applied to the first. The conclusion in particular needs a closing analytical move rather than a summary restatement. The equity risk deserves a more honest interrogation of whether the model, as currently articulated, has an adequate answer to it.

---

**Dimension 4: Purposeful Structure — Score: 6/10 — Adequate****Strengths**

The macro arc from problem to proposition to objection to solution to benefits to risks to conclusion is coherent and serves the argument. The LinkedIn pointer to long piece relationship is well-designed. The sequencing decision on the doctoral model is structurally intelligent and will pay off in the editing pass.

**Weaknesses**

Ten named sections is too many for a 3,900-word discussion paper. The piece currently has a structural seam where the skills assessment section was inserted, and it shows. It interrupts the proposition-to-objection flow in a way that a reader will feel even if they cannot name it. The "Accepting the Concerns" and "What Assessment Is Actually For" sections do overlapping work and could be consolidated without loss. The "Tools That Already Work This Way" section sits between the risks and conclusions in a way that feels like an afterthought placed before the close rather than a considered structural choice.

**Area for Development**

The editing pass should approach structure as a consolidation exercise. Ten sections to seven or eight, with the skills material integrated into the proposition section rather than standing separately, and the tools section either absorbed into the conclusion or given a clearer structural rationale.

---

**Dimension 5: Appropriate Register — Score: 6/10 — Adequate****Strengths**

The opening sections find a strong and distinctive voice. The elephant section maintains the correct balance throughout. Several individual lines work well, particularly the proposition section's opening ("What if the assessment task itself was the assessment?") and the conclusion's penultimate paragraph. The piece does not sound like a generic education commentary piece, which is the baseline it needed to clear.

**Weaknesses**

Register inconsistencies appear across the piece. The Boud and Sadler passages drift into academic formality. The benefits section reads more flatly than the earlier sections and uses phrasing that has currency in education discourse without adding distinctiveness. The learner experience passages, when they appear, occasionally become slightly reverential toward the referenced source material in a way that shifts the register noticeably.

There is also an AI cadence issue that runs through the piece. See the AI Voice assessment below for detail. It affects register significantly.

### Area for Development

Register needs to be treated as a specific quality target in the editing pass, not an incidental concern. The benefits section in particular needs rewriting with the same voice the opening sections achieve. Every passage where the text sounds like it is summarising a source rather than making an argument should be reviewed.

---

## Dimension 6: Critical Integrity — Score: 7/10 — Capable

### Strengths

The AI correctly challenged the "fit for purpose in an AI world" framing as insufficiently differentiating. That intervention materially improved the piece. The skills assessment gap was named honestly when identified. The qualification framework constraint was not minimised. The gaming objection was engaged with seriously rather than dismissed.

### Weaknesses

The benefits section was produced without equivalent critical challenge to that applied to the risks section. The claimed benefits, particularly around agency, self-direction and practitioner role reframing, are presented with more confidence than the evidence base strictly supports. The AI did not push back on this sufficiently.

The equity concern in the risks section was named but not pressed. A more critically engaged response would have interrogated whether the model, as currently proposed, has anything more than an acknowledgement of the equity problem.

### Area for Development

The editing pass should apply the same critical standard to the benefits section that was applied to the risks section. Claims about learner outcomes, practitioner development and system-level impact need the same evidential grounding as the concerns they are intended to answer.

---

## Dimension 7: Evaluative Judgement — Score: 6/10 — Adequate

### Strengths

The structural decisions were generally sound. The push away from the well-trodden framing was the highest-value single intervention in the conversation and it was made at the right moment. The sense check pause before settling the spine was well-timed and productive. The repositioning of a key structural element was a sharp evaluative call that materially improved the argument's sequencing.

### Weaknesses

Length management was poorly handled. The piece exceeded its target word count before the argument was complete and this was flagged but not resolved before drafting continued. A firmer structural decision at that point — agreeing the extended target before proceeding rather than noting it and moving on — would have produced cleaner drafting.

The key evidential search happened after drafting had begun. The full reference set should have been confirmed before a single section was written. The search added a strong evidential anchor but its late arrival meant it was retrofitted into existing drafts rather than shaping them from the start.

Some of the author's framings were accepted without sufficient interrogation. The central rhetorical premise is a strong position but it is an assertion. The piece does not fully evidence it and this was not challenged with proportionate rigour at the ideation stage.

### Area for Development

Process sequencing: reference confirmation before drafting, structural parameters agreed before word count is exceeded, mid-draft quality checks before proceeding to subsequent sections. These are process habits, not one-off corrections.

## Holistic AI Voice Assessment

### AI Voice (Holistic Assessment) — Score: 4/10 — High AI presence

The current draft carries a detectable AI voice in multiple sections, estimated at 35 to 40 percent of the text. This is significantly above the implicit sub-15 percent standard for thought leadership work of this kind, and it is the highest-priority development concern for the editing pass.

Four markers characterise the AI voice in this output.

#### List-Introduction Structures

"The concerns cluster around three things" is a classic AI framing that signals a list is coming. The piece uses this pattern more than once.

#### Rhythmic Removal Sequences

"Strip the method away. Remove the examination, the observation, the portfolio..." is analytically effective but the rhythm is an AI cadence that an experienced editor will identify.

#### Formulaic Benefit Framing

"The benefits of the model are not abstract. They are observable at the level of the individual learner, the practitioner, and the system" is a structure that appears frequently in AI-generated professional writing.

#### Connective Tissue Passages

Several transitions between sections read as structural glue rather than genuine argument. "There is a further implication of this model that the sector has been circling for some time" is the clearest example.

#### Areas for Development

The humanising pass needs to be systematic, not selective. Every section should be reviewed for AI cadence with the target of reducing detectable AI voice to below 15 percent. The author's voice in the collaborative exchanges is more distinctive than the voice in the drafted sections. The editing pass should move the drafted text toward the conversational register the author used when pushing back and challenging during ideation.

---

## Overall Summary

Three findings define this evaluation.

### Finding 1 — The Central Argument Is the Piece's Most Important Achievement

The insight that AI exposed rather than created the proxy problem is genuinely differentiating, and the analytical development of it through the gaming counter, marking integrity and formative-to-summative sections is the strongest work in the piece. This is the spine the editing pass must protect. At v0.1, this is sufficient to anchor the whole piece.

### Finding 2 — AI Voice Is the Most Urgent Development Priority

At an estimated 35 to 40 percent, the piece does not yet read as the author's thought leadership. It reads as AI-assisted drafting. That is not the standard the piece needs to reach its audience or establish the credibility the argument deserves. Until the AI voice reduction pass is completed, no other editing decision can be fully evaluated.

### Finding 3 — The Evidence Architecture Is Present but Shallow

Names are doing the work that findings should be doing. The editing pass needs to go back into each reference and extract a specific claim that does genuine argumentative work, or reduce the reference to a passing acknowledgement rather than a substantive anchor. The gaming counter in particular needs either an empirical anchor or an honest acknowledgement that it is a reasoned proposition rather than an established finding.

<b>Highest-priority development area</b>	The AI voice reduction pass. Everything else is secondary to establishing the author's voice as the authorial presence in the piece. Until that is done, no other editing decision can be fully evaluated.
--	--

## Evaluation Record

<b>Date</b>	16 April 2026
<b>Output type</b>	Thought leadership article — AI and professional assessment (v0.1 draft)
<b>Review type</b>	Mid-development — Full
<b>Genre</b>	Formal development
<b>Rubric version</b>	1.1
<b>1. Fit to Context</b>	6/10 — Adequate
<b>2. Evidence and Grounding</b>	6/10 — Adequate
<b>3. Analytical Depth</b>	7/10 — Capable
<b>4. Purposeful Structure</b>	6/10 — Adequate
<b>5. Appropriate Register</b>	6/10 — Adequate
<b>6. Critical Integrity</b>	7/10 — Capable
<b>7. Evaluative Judgement</b>	6/10 — Adequate
<b>AI Voice (Holistic)</b>	4/10 — High AI presence
<b>Highest-priority development area</b>	The AI voice reduction pass. Everything else is secondary to establishing the author's voice as the authorial presence in the piece. Until that is done, no other editing decision can be fully evaluated.
<b>Published with approval</b>	Yes — practitioner consent confirmed. No content altered from original evaluation.

*This evaluation was produced using the Dynes Insights AI Output Quality Rubric v1.1, applied in Checker Mode at Full, Mid-development standard. The evaluation has been reviewed by the issuing party. Published with the practitioner's full approval. No content has been altered from the original evaluation.*