GPT-5 vs Human Markers in Creative Writing Assessment

A Comparative Study of Marking Accuracy and Feedback Quality

Simon Brookes
Associate Dean (Students), Faculty of Creative and Cultural Industries
University of Portsmouth

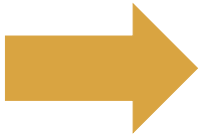
FRAMING

Reframing the question

Not, can we replace academics with AI

Less useful

Can AI mark as well as a human?



More useful

Which functions of marking and feedback is each party suited to?

This study addresses the second question empirically, in a context where standards, judgement, and feedback cannot be reduced to a single score.

Assessment comparison framing: Sadler (2009); Wetzler et al. (2025).

WHY THIS MATTERS

Assessment is a high-stakes trust system

Marking and feedback decisions shape progression, confidence, standards, and workload.

Fairness

Students need decisions that are consistent enough to feel defensible.

Standards

Rubrics encode disciplinary values, not just scoring rules.

Feedback

Written comments matter when they help students see what to do next.

Trust

AI outputs can be fluent and plausible without being correct.

Feedback effectiveness: Hattie & Timperley (2007); Nicol & Macfarlane-Dick (2006); Carless & Boud (2018).

METHOD

Design and corpus

A controlled comparison of human marking practice with two AI conditions.

10

anonymised Level 4 essays

Creative Writing critical essays

3

independent human markers

First-mark judgements; no moderation

2

GPT-5 conditions

Zero-shot and few-shot

5x

runs per essay

Median mark and feedback selected

The human scores are a benchmark (comparison point), not a definitive ground truth. The study compares AI with routine marking variability. So when I talk about AI alignment, I mean alignment with human marking practice, not proof that the AI was objectively correct.

METHOD

What zero-shot and few-shot mean here

The only difference is whether the model saw calibrated examples.

ZERO-SHOT

Inputs

- Module context
- Assessment guidance
- Official rubric
- Student essay

No marked examples are provided.

FEW-SHOT

Same inputs, plus anchors

- Three human-marked exemplars
- Low, mid, and high-performance points
- Mean mark, band, and human feedback
- Leave-one-out rule for anchor essays

PROMPT EXAMPLES

Prompt Structure

Prompt design is an art as much as it is a science

1	**System role**	Expert assessor; use official rubric; stable scoring; JSON only.
2	**Shared assessment context**	Module overview, task guidance, learning outcomes, essay expectations.
3	**Rubric**	Verbatim grade-band descriptors from Fail to First.
4	**Target essay**	Full anonymised essay text to mark.
5	**Output schema**	Percentage mark, mapped band, and feedback comment.

FEW-SHOT ADDS

Anchor exemplars

Three human-marked examples spanning lower, mid, and higher performance points.

Each anchor contains the essay, mean human mark/band, and human feedback. A leave-one-out rule prevents self-calibration.

The prompts were long because the evidence supplied to the model was long, not because the instruction was elaborate.

PROMPT EXAMPLES

Prompt excerpts

Two different prompts

ZERO-SHOT

```
SYSTEM
You are an expert assessor marking a Creative
Writing critical essay...
Use the official rubric provided.
Use deterministic reasoning and consistent
scoring behaviour...
Return ONLY valid JSON.

USER CONTEXT
Essay ID: <E##>
Task: Mark this essay according to the rubric.
Output: percentage mark + mapped band.
Justification: Provide feedback for the student.

[verbatim module context + rubric]
[target essay to mark]

[FULL PROMPT]
```

FEW-SHOT

```
SYSTEM
You are an expert assessor marking a Creative
Writing critical essay...
Use the three anchor exemplars below as
reference points for calibration.
Do NOT reveal or refer to the anchor essays.
Return ONLY valid JSON.

USER CONTEXT
Condition: few_shot
You will see Low, Mid, High anchors...

ANCHOR EXEMPLARS
Low anchor: essay + mean mark + feedback
Mid anchor: essay + mean mark + feedback
High anchor: essay + mean mark + feedback

[FULL PROMPT]
```

The experimental difference is the presence or absence of calibrated anchor examples.

FINDINGS: MARKING

The human benchmark was variable

Represents subjective judgement

11.2

mean range

between the three human marks

5.79

mean SD

within essay, across humans

20%

full band agreement

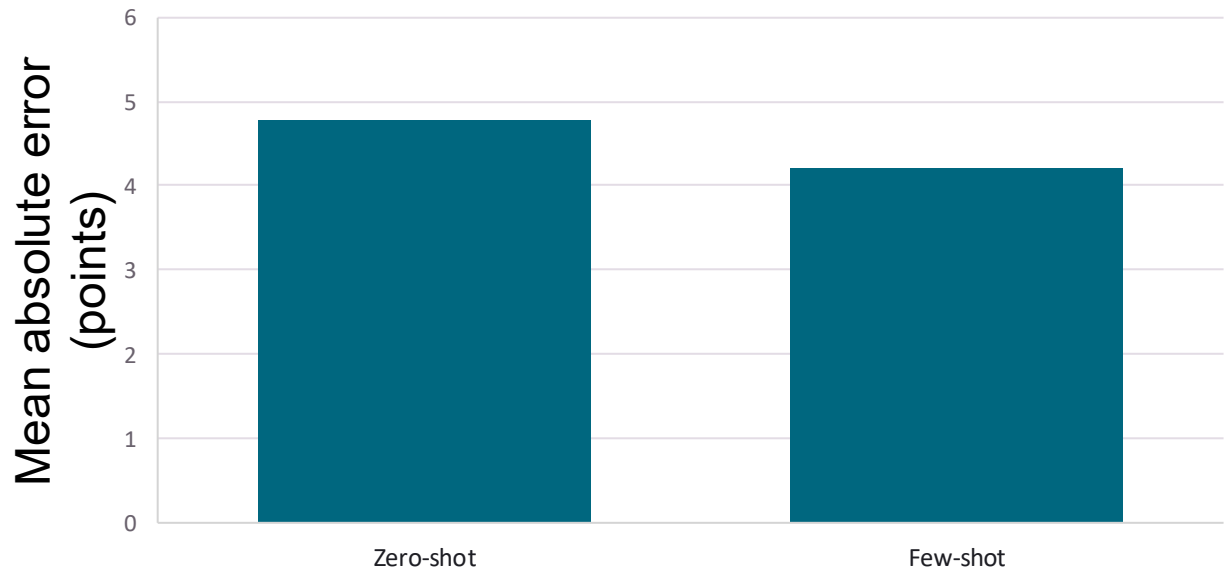
2 of 10 essays

AI-human differences should be read against this baseline. The study evaluates agreement with human practice, not objective correctness.

Marking reliability: Meadows & Billington (2005); Sadler (2009); Wetzler et al. (2025).

AI marks aligned reasonably with the human mean

Few-shot reduced under-marking and slightly reduced average absolute error.



Read carefully

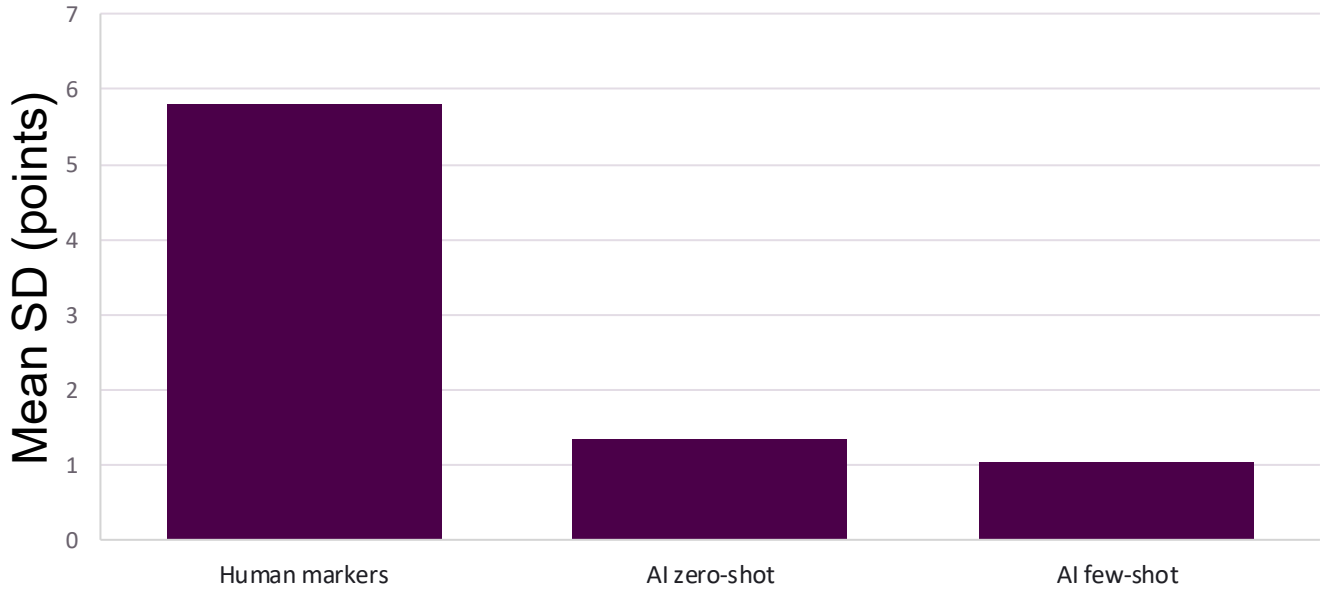
- Few-shot MAE: 4.2
- Zero-shot MAE: 4.77
- Bias: -1.0 vs -2.7
- Correlation: r = .70 vs .68
- Agreement, not objective accuracy

Band agreement with the human mean remained limited: zero-shot 50%, few-shot 40%.

AI-human comparison context: Kaliisa et al. (2025/2026); Chan & Colloton (2024); Karademir Coskun & Alper (2024).

AI marks clustered much more tightly across repeated runs

But this was measured under a prompt that explicitly asked the model to prioritise stability.



Note

This is tight clustering under a stability-prompted setup. It is not, by itself, evidence that the model captured the full construct better than human markers.

Mean ranges: humans 11.2 points; zero-shot 3.0; few-shot 2.4.

Reliability and construct caution: Sadler (2009); Wetzler et al. (2025).

METHOD: FEEDBACK

How feedback quality was coded

Fifty feedback comments were coded on six 1-4 dimensions.

Clarity

Understandable, coherent, readable.

Specificity

Concrete and text-linked rather than generic.

Actionability

Clear, feasible next steps and feed-forward.

Tone

Professional, respectful, constructive stance.

Rubric alignment

Connection to criteria and descriptors.

Recognitive cues

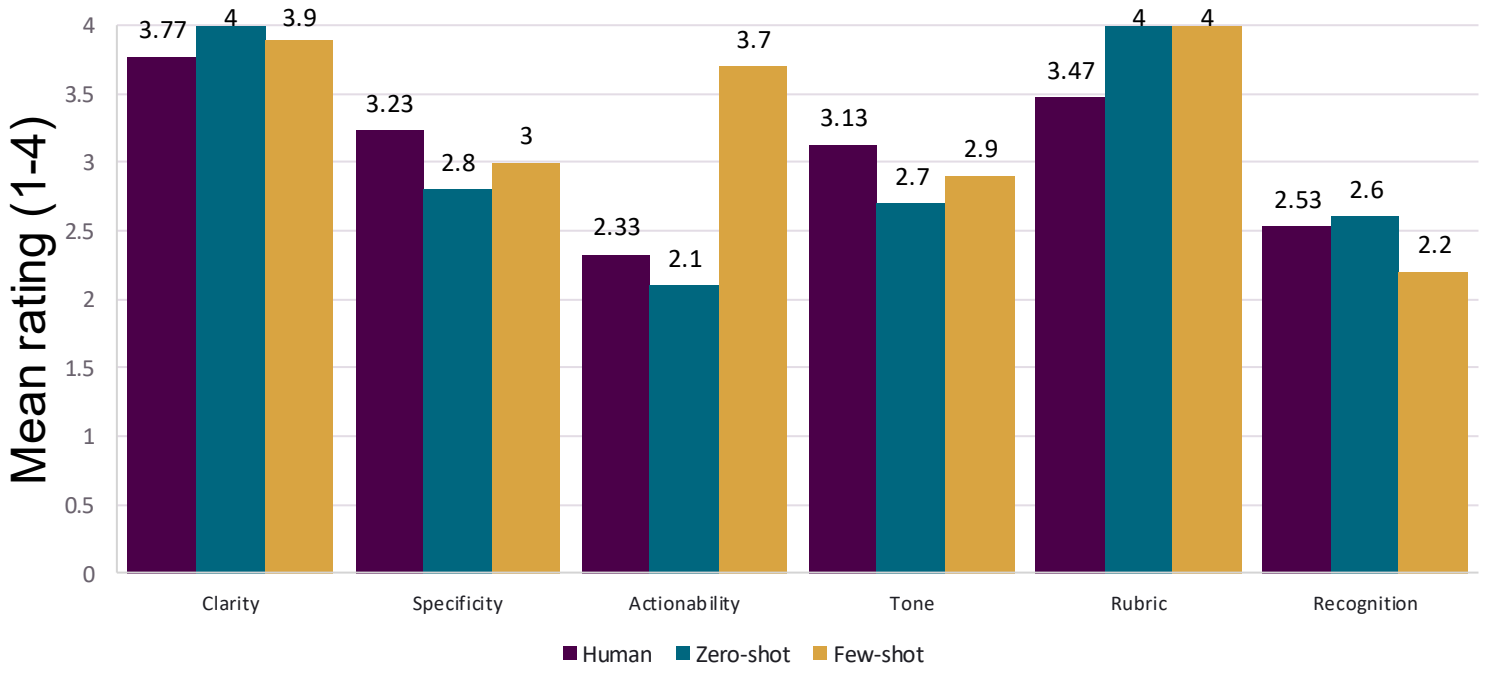
Acknowledges effort, growth, or emerging voice.

Coding was textual: it assessed observable qualities of the feedback, not whether the advice was actually valid for the essay.

After Hattie and Timperley (2007), Nicol and Macfarlane-Dick (2006), Hyland and Hyland (2001), and Corbin et al. (2025)

Feedback quality showed trade-offs

The overall average hides the more interesting pattern.



Pattern

AI: clarity and rubric alignment

Humans: specificity and tone

Few-shot: actionability

Few-shot: lower recognitive cues

Mean feedback length: human 574 chars; zero-shot 767; few-shot 1824.

AI feedback context: Escalante et al. (2023); Steiss et al. (2024); Kasneci et al. (2023).

Actionability was the clearest divergence

Few-shot feedback more often turned critique into explicit next steps.

Human markers

2.33 / 4

Often diagnosed weaknesses clearly, but less often set out an explicit route forward.

“It would have been good to see more primary material engaged with.”

Zero-shot

2.10 / 4

Clear and rubric-aligned, but often relatively general in next-step guidance.

“The work lacks sustained critical analysis of texts, techniques, or the creative process, and does not develop a coherent argument.”

Few-shot

3.70 / 4

More often specified what a stronger performance would require.

“To improve: formulate a precise thesis… select 2-3 short extracts and perform close analysis of style… connect these to recognised craft concepts… ensure each paragraph advances the central argument…”

Feedback uptake and feed-forward: Hattie & Timperley (2007); Carless & Boud (2018).

Recognition is not just a style

Few-shot AI scored lowest on recognitive cues: 2.2, below human feedback at 2.5 and zero-shot at 2.6.

2.2

few-shot recognitive cues

lower than human and zero-shot

Corbin et al. (2025): genuine recognition depends on relational conditions a model cannot supply.

Where recognising language appears in AI feedback, treat it as recognitive styling rather than recognitive practice.

Corbin et al. (2025); Hyland & Hyland (2001).

SYNTHESIS

A complementary division of labour

The findings suggest support and oversight, not substitution.

Suited to GenAI support

- Producing a consistent first draft of feedback
- Mapping comments clearly to rubric criteria (rubric alignment)
- Suggesting concrete next steps for improvement (actionability)
- Keeping feedback clear and well structured (clarity)
- Comparing work against calibrated examples

Retained by human markers

- Making and owning the final judgement
- Checking whether feedback is valid for the actual essay
- Giving text-specific disciplinary guidance (specificity)
- Providing warmth and constructive tone (tone)
- Offering dialogue, encouragement, and recognition (recognitive cues)

The human marker owns the judgement; AI may help draft, structure, compare, or prompt review.

DESIGN IMPLICATIONS

If using AI for feedback, use it cautiously

Design principles suggested by this study.

1 **Use exemplars, not just a rubric**

Few-shot feedback was more actionable when the model saw calibrated examples.

2 **Keep the human in control**

AI output should be checked, edited, and owned by a marker.

3 **Treat AI feedback as a draft**

The model can structure and extend feedback, but cannot validate its own judgement.

4 **Prompt for high-quality feedback**

Include clearly defined feedback specifications to improve feedback efficacy.

5 **Check the interpersonal dimension**

Actionability may increase while recognition, warmth, or student-centredness falls.

6 **Audit accuracy and boundary risk**

Verify that the mark and advice fit the essay, especially near classifications.

Bottom line: AI may support feedback drafting, but the human marker remains responsible for validity, judgement, and the student relationship.

CHALLENGES

What the study does not settle

These are design, governance, and pedagogy questions.

Veracity

Are AI comments accurate and appropriate for the actual essay?

Assessment standards

Does the AI output reflect the qualities the rubric is meant to value? Does it “understand” the discipline?

Student acceptance

Would students trust or welcome AI-assisted assessment in creative disciplines?

Privacy and DPIA

How is student work processed, stored, audited, and disclosed?

Ethical implications

What ethical issues arise if AI is used to process or assess student work?

Prompt steering

Tone, specificity, and recognition may change with different instructions.

Governance and validity concerns: Kapoor & Narayanan (2023); Li & Flanigan (2024); Sadler (2009).

LOCAL DISCUSSION

What this means for Learning and Teaching

An invitation to consider and specify boundaries.

Where is AI useful?

Drafting rubric-mapped feed-forward? Moderation support? Student self-checking?

Where is AI inappropriate?

Final judgement? Relational feedback? Sensitive creative work?

What evidence would we need?

Student trust, feedback uptake, validity checks, governance assurance.

What would we protect?

Human accountability, disciplinary standards, mentoring, and learner/teacher trust.

The technology is clearly capable but what kind of assessment culture do we want to build around it?

Discussion

Questions

What should we do with a tool that can produce plausible, consistent, actionable feedback but cannot own judgement, recognition, or responsibility?

- **Which feedback functions should remain human-owned?**
- What would count as evidence of trustworthy AI assistance?
- Would students experience this as support, surveillance, or substitution?
- What should change in assessment design before tooling changes?

Thank you

Simon Brookes

REFERENCES

Selected references

Brookes, S. (in press). GPT-5 vs human markers in creative writing assessment: A comparative study of marking accuracy and feedback quality. Innovations in Education and Teaching International.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. Assessment & Evaluation in Higher Education, 43(8), 1315-1325. https://doi.org/10.1080/02602938.2018.1463354

Corbin, T., Tai, J., & Flenady, G. (2025). Understanding the place and value of GenAI feedback: A recognition-based framework. Assessment & Evaluation in Higher Education, 50(5), 718-731. https://doi.org/10.1080/02602938.2025.2459641

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. International Journal of Educational Technology in Higher Education, 20(1), 57. https://doi.org/10.1186/s41239-023-00425-2

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81-112. https://doi.org/10.3102/003465430298487

Hyland, F., & Hyland, K. (2001). Sugaring the pill: Praise and criticism in written feedback. Journal of Second Language Writing, 10(3), 185-212. https://doi.org/10.1016/S1060-3743(01)00038-8

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Kruber, S., Kuber, G., Lieber, J. M., Nerdel, C., Peres, F., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. Studies in Higher Education, 31(2), 199-218. https://doi.org/10.1080/03075070600572090

Sadler, D. R. (2009). Grade integrity and the representation of academic achievement. Studies in Higher Education, 34(7), 807-826. https://doi.org/10.1080/03075070802706553