GenAI and the Essay: Evaluating Task Specificity and Rubric Alignment

Considerations When Designing Essay Assessment **...** ***2***

Evaluating Essay Assessments in the Context of GenAI Capabilities **......................** ***4***

Task Audit Instructions:... 5

References **..** ***14***

Considerations When Designing Essay Assessment

When designing assessment criteria for academic essays, the primary consideration is **task specificity**—developing tasks that leverage GenAI's known limitations in reasoning, while avoiding areas that align with its strengths, such as mimicking tone and style through pattern matching and reinforcement learning from human feedback (RLHF) (Lappin 2024, 16). To determine whether your assessment task is specific or generic, and whether it aligns with or resists the current capabilities of GenAI, consider the following guiding question:

1. **Could this assessment task be set at any university?**
   If the task is generic enough to be used across multiple institutions, it is more likely to align with GenAI's pattern-matching strengths. Such tasks may already exist in its training data, or may have been optimised by previous users, increasing the risk that GenAI could generate a competent response. For example, close readings of poetry—even when applied to a novel poem—may still appear persuasive, as the task draws on familiar interpretive patterns.

   - GenAI relies heavily on pattern recognition from vast training datasets and performs well when responding to familiar, formulaic prompts.
   - Tasks common across institutions are likely to reflect prompts already present in GenAI's training data or those it has been optimised to replicate via RLHF.
   - Essays generated under these conditions often appear superficially competent but lack originality or deep analytical insight.

2. **Does the task require engagement with recent, local, or discipline-specific material?**
   Tasks that draw on up-to-date, geographically contextualised, or niche disciplinary content are less likely to overlap with GenAI training data and are more resistant to generic responses.

   - GenAI struggles with content that is highly localised or recent, especially when that material is not included in public datasets or paywalled content.
   - Because it lacks access to many academic or discipline-specific databases, its outputs tend to rely on outdated or general information.
   - GenAI cannot consistently verify the accuracy or relevance of location-specific or cutting-edge developments in a discipline.

3. **Is the task scaffolded around a specific case study, dataset, or real-world scenario?**
   Requiring students to apply theoretical frameworks to concrete, context-specific cases—particularly those tied to specific geographic or temporal settings — also helps resist GenAI-generated responses. These tasks challenge GenAI's limited reasoning capabilities and exposes weaknesses such as redundancy, verbosity,

and a lack of critical thinking or argumentation, which can be directly addressed in the assessment criteria and rubric.

- GenAI lacks robust reasoning about real-world scenarios, especially those that require interpreting the significance of a specific dataset or event.
- AI-generated responses tend to exhibit redundancy, circular logic and shallow analysis when forced to engage with unfamiliar or complex applied contexts
- Tasks demanding synthesis of theory with specific cases require critical thinking and causal reasoning, which LLMs are not capable of replicating reliably

4. **Does the task require methodological or reflective justification?**
   Asking students to explain *how* and *why* they approached the topic a certain way (e.g. methodological rationale, research process, source evaluation) exposes reasoning gaps in AI-generated responses. GenAI doesn't possess the capacity for causal reasoning—it can detect patterns but cannot understand or reason about the underlying causes (Pearl 2018). Without causal models, GenAI is confined to surface-level associations and are incapable of answering questions involving interventions ("What if we do X?") or counterfactuals ("What would have happened if…?").

   - GenAI lacks causal models and cannot explain the rationale behind decisions—it mimics form but not substance (Pearl 2018).
   - It is incapable of reflective thinking or articulating a research process in a logically coherent way beyond surface-level justification.
   - AI-generated writing often exhibit logical inconsistency or abandoned reasoning paths (Wang et al. 2023), especially in multi-step analysis based tasks.

5. **Are students asked to reference and integrate specific academic or scholarly sources?**
   GenAI frequently fabricates or misattributes references. Designing tasks that require the use of precise, discipline-specific scholarly sources—particularly those behind paywalls or less commonly cited—reduces the likelihood of credible AI-generated responses. Incorporating digital and information literacy frameworks into assessment criteria helps counteract GenAI by requiring students to identify and apply authoritative, credible sources. This process is a necessary foundation of a strong academic argument.

   - GenAI is a stochastic parrot (Bender et al. 2021) and cannot verify or trace the accuracy of sources; it often generates hallucinated or sometimes fabricated references**.**
   - It lacks access to subscription-based academic databases and relies on public, often non-scholarly, information.
   - GenAI does not understand what constitutes a credible source in a disciplinary context, and cannot consistently differentiate between academic, trade, and promotional content.

Evaluating Essay Assessments in the Context of GenAI Capabilities

Outcome: This task will help you evaluate your assessment in relation to GenAI's current capabilities. It is designed to ensure that your assessment criteria focus on areas where GenAI is weakest and to identify any misalignments between your rubric and the types of responses GenAI can produce.

Optional Pre-Workshop Activity: Evaluate Your Task with a GenAI Tool

This activity is optional but recommended to complete before the workshop to support more targeted reflection and discussion.

1. Copy the instructions for one of your current written assessment tasks into a commercial chatbot (e.g. ChatGPT, Gemini, Copilot).
2. Review the output:
    - Is the response coherent, persuasive, or superficially competent?
    - Does it resemble what a student might submit?
3. Compare the output to your existing rubric, would this response meet your current assessment criteria?
    - Where does it fall short—if at all?
4. Reflect:
    - How specific is your assessment task?
    - Does it clearly align with your subject's intended learning outcomes (ILOs)?
5. Decide:
    - Could you modify one element of the task to make it more specific to your discipline, content, or context?

Task Audit Instructions:

1. Choose a current essay assessment task from a subject you deliver.
2. For each of the five questions below, select either A or B based on your evaluation of the task.
3. At the end, count how many "A" responses you selected.
4. Discuss your results in pairs or small groups and reflect on your task's design, criteria, and rubric.

Step One: Answer the following questions

1. Could this assessment task be set at any university?

This question invites you to consider how generic or discipline-specific your assessment task is. In other words, could the same task (with minimal or no changes) be used across multiple subjects or institutions — or is it clearly tailored to your course, discipline, or learning outcomes?

A. □ **Yes** → Task may be too generic and easily addressed using GenAI's pattern-matching capabilities.
B. □ **No** → Task is tailored to specific subject content or context, reducing overlap with GenAI's current training and optimisation for certain tasks.

Why this matters:

- GenAI excels at replicating familiar, widely-used essay prompts.
- Generic tasks often align with its training data or previous optimised uses.
- Responses may appear coherent but lack depth, originality, and analytical rigour.

2. Does the task require engagement with recent, local, or discipline-specific material?

A. □ **Yes** → Incorporates material GenAI is less likely to have access to or replicate accurately.
B. □ **No** → May produce responses drawn from outdated or non-specialised sources.

Why this matters:

- GenAI's knowledge is limited to public data and often lacks access to paywalled or up-to-date academic sources.
- It struggles to assess the credibility or relevance of geographically or topically specific content.

- Tasks that demand context-specific insights are more likely to reveal critical thinking.

3. Is the task scaffolded around a specific case study, dataset, or real-world scenario?

A. ☐ **Yes** → Challenges GenAI's reasoning and forces students to synthesise theory and real-world context.
B. ☐ **No** → May allows for generic or overly descriptive responses.

Why this matters:

- GenAI performs poorly with applied reasoning or unfamiliar real-world cases and contexts.
- AI responses in these cases tend to become verbose, repetitive, or shallow.
- Analysing specific scenarios reveals deeper conceptual understanding and original insight.

4. Does the task require methodological or reflective justification?

A. ☐ **Yes** →Highlights GenAI's inability to explain reasoning or articulate decisions authentically.
B. ☐ **No** → May allow for imitation of argument structure without critical awareness.

Why this matters:

- GenAI lacks causal models—it can generate responses but not explain why particular choices were made.
- It cannot genuinely reflect on the research process or methodological rationale.
- Requiring justification exposes limitations in coherence and logic in AI-generated content.

5. Are students required to reference and integrate specific academic or scholarly sources?

A. ☐ **Yes** → Increases the difficulty for GenAI to fabricate or misattribute citations.
B. ☐ **No** → Increases the likelihood of fabricated or misused references in AI responses.

Why this matters:

- GenAI can produce fake or unverifiable citations and cannot reliably access paywalled academic content.
- It does not consistently understand disciplinary standards for authoritative sources.
- Requiring students to apply specific, credible sources fosters scholarly integrity and information literacy.

Step Two: Interpret your results

Count how many **"A"** responses you recorded during the task audit.

If you selected "A" for 2 or more questions :

This suggests that your assessment task may be well-designed to promote student learning. It likely:

- Encourages original, contextualised thinking
- Requires research and reasoning processes that GenAI cannot reliably replicate
- Demands engagement with credible, discipline-specific content
- Minimises reliance on generic optimised essay forms

In short:
Your assessment is unlikely to align with GenAI's current capabilities and instead supports student work that encourages student learning and not just the production of fluent writing.

Next steps:
Review your assessment rubric to ensure it aligns with these strengths by explicitly rewarding:

1. Critical thinking and original argumentation
2. Methodological reasoning and justification
3. Use of authoritative, discipline-appropriate source

Even strong assessment tasks can be weakened by rubrics that overlook these core capabilities.

If you selected "A" for 1 or fewer questions (mostly "Bs"):

This indicates your essay task may benefit from some revision. A low score suggests the task may:

- Be too generic or similar to prompts currently used to optimise GenAI
- Lack opportunities for critical thinking, methodological reflection, or scholarly depth

In short:
Your assessment may allow students to bypass deeper learning allowing them to rely on AI-generated content

Next Steps:

Revise your assessment rubric to better align with learning outcomes that GenAI cannot easily simulate including:

- Analytical depth over description
- Justification of research and source selection
- Integration of credible, specific scholarly materials

Step Three: Revise rubric criteria

Use your responses from **step two** to address your existing assessment rubric. Discuss in pairs or small groups and reflect on your task's criteria and rubric.

1. Begin by identifying at least one criterion that could be improved. If you're unsure where to start, look for criteria that rely on broad or vague language, such as "clarity of written expression," "structure," or "grammar."
2. Select one of these and revise it using the example criteria provided in **Table 1 and** the sample rubrics (Table **2 and Table 3** as a guide).

Ask yourself, Does my rubric clearly reward the kinds of writing that GenAI struggles to replicate?

How to Use This Table (Table 1)

- Review your current rubric against each row.
- **Ask:** Does the rubric already address this GenAI limitation? If not, consider incorporating the recommended example criteria.
- Use the example rubric criteria provided below as a guide to reword or add new descriptors to your assessment tool.
    - Not all criteria will be relevant based on your specific essay assessment task
    - An example essay rubric for a research essay is provided in Table 2 and an example of generic essay criteria is provided Table 3 to assist with this task.

Table 1. GenAI Limitations and example rubric criteria

GenAI Limitation	What GenAI Struggles With	What to Emphasise in Rubric	Example Rubric Criteria
Generic Tasks / Pattern Replication	produces formulaic arguments based on familiar prompts that produces optimised outputs	Originality, specificity, discipline alignment	- Original framing of topic - Relevance and specificity of research focus
Lack of Access to Recent, Local or Discipline-Specific Material	Cannot reliably retrieve up-to-date or paywalled academic content; lacks geographical or contextual nuance	Use of contemporary, localised, or discipline-specific sources	- Use of up-to-date and contextually relevant sources - Engagement with current debates or local case studies
Shallow Reasoning / Poor Application of Theory	Redundant or circular logic; weak synthesis of theory with case data or examples	Analytical depth, application of theory to practice, synthesis	- Integration of theory with case study or real-world example - Clear, logical argument progression - Depth of analysis beyond description
No Methodological Justification / Reflective Capacity	Cannot explain reasoning or process; lacks causal reasoning and reflection	Justification of research decisions; reflective engagement	- Methodological clarity and rationale - Reflection on research approach or limitations - Awareness of interpretive choices
Fabricated or Misused Sources	Hallucinates citations; cannot assess source credibility or relevance based on disciplinary definitions	Scholarly source quality; citation accuracy	- Use of credible, discipline-appropriate sources - Correct citation format and integration - Critical evaluation of source reliability

Table 2. Example Essay Rubric: A sample rubric for a research essay using standard academic essay criteria.

Criteria	H1	H2A	H2B	H3	P	N
Understanding of the Topic *Understanding of the topic and its relevance to coursework and scholarship*	Excellent understanding. Engaged with highly relevant coursework and scholarship.	Very strong understanding. Engaged with relevant coursework and scholarship.	Good understanding. Engaged with some relevant coursework and scholarship.	Mostly understood topic. Some engagement with coursework and scholarship.	Needs improvement. Attempted engagement with relevant material.	Has not understood topic. No relevant engagement with scholarship.
Research Skills *Use of evidence, relevance of sources, and correct citation*	Excellent research skills. Highly relevant sources. Argument well-supported. No citation errors.	Very strong research skills. Persuasive argument with minor citation errors.	Good research. Sources support argument. Minor citation errors.	Reasonable research. Some supporting sources. Some citation errors.	Limited research. Incomplete support. Citation errors.	Poor or inappropriate research. Citation not evident or plagiarised.
Critical Engagement *Analysis and understanding of key issues in research materials*	Excellent engagement. Explored issues to a high standard.	Very strong engagement. Demonstrated clear understanding of critical issues.	Strong engagement. Begun to analyse critical issues.	Some engagement. Greater analysis needed.	Limited analysis. Some understanding evident.	Inadequate engagement. No understanding of critical issues.
Persuasive Argument *Ability to construct an original and well-supported argument*	Highly persuasive and original. Strong independent thinking.	Very persuasive. Strong use of sources. Independent thinking evident.	Persuasive. Relevant sources used. Attempt at independent thinking.	Clear but underdeveloped. Emerging independence.	Attempted argument. Limited support and independence.	No argument. Lacks support and independent thought.
Written Expression *Clarity, grammar, organisation of ideas, and argument staging*	Fluent, precise, error-free. Argument well-staged.	Expressive and error-free. Complex ideas clearly conveyed.	Clear expression. Very few errors. Meaning conveyed well.	Sound expression. Some awkward phrasing or errors.	Weak grammar. Parts unclear.	Incoherent grammar. Difficult to understand.

Table 3. Example Essay Criteria: A sample of commonly used, generic criteria that may be applied when assessing essay assessment.

A holistic rubric evaluates a piece of work as a whole, rather than judging individual components separately. Holistic rubrics are particularly useful when the overall quality of a performance is more important than the evaluation of specific details, or when the elements of a task are interconnected and difficult to isolate. In contrast, an analytic rubric breaks the assessment into distinct criteria, with each aspect scored separately. While this approach is ideal for providing detailed feedback, holistic rubrics allow for a more integrated judgement of a student's work—especially useful in disciplines or tasks, such as essay writing, where creativity, critical thinking, and synthesis are valued as a combined effort (De Boer et al. 2021, 7-8).

Criteria	Description
Understanding of the Topic	*This criterion assesses the student's grasp of the topic within the discipline. You may select one or more of the following descriptors when designing a holistic judgement:* • Clearly defines the topic within the context of the specific discipline. • Demonstrates understanding of the topic's relevance to the subject area. • Explains the topic using appropriate scholarly references and theoretical frameworks. • Provides contextual background that enhances understanding of the topic. • Articulates the significance or implications of the topic within the field. • Uses discipline-specific language accurately and effectively.
Research Skills	*This criterion focuses on the quality and appropriateness of research sources, as well as citation practices. You may select one or more of the following descriptors when designing a holistic judgement:* • Uses relevant and credible sources that align with the topic and academic field. • Selects a sufficient range of sources to support the essay's argument or discussion. • Demonstrates an ability to evaluate and choose sources critically. • Effectively integrates references into the argument or narrative. • Applies a consistent and appropriate referencing style specified in the subject • Avoids overreliance on a single source or non-academic references.
Critical Engagement	*This criterion focuses on how well the student engages with research materials to build an argument. Educators may select one or more of the following descriptors when designing a holistic judgement:* • Goes beyond description to provide analytical insight into ideas and sources.

	• Integrates evidence meaningfully into the overall argument. • Demonstrates critical thinking by questioning assumptions or contrasting perspectives. • Shows understanding of nuance and complexity in the chosen topic. • Connects theory to practice or applies concepts in an original way. • Demonstrates the ability to synthesise ideas from multiple sources.
Persuasive Argument	*This criterion measures the strength and originality of the student's argument. This criterion may distinguish ambitious essays from more generic ones. You may select one or more of the following descriptors when designing a holistic judgement:* • Constructs a clear, logical, and persuasive argument. • Demonstrates independent thinking and intellectual curiosity. • Shows originality in approach, interpretation, or perspective. • Supports claims with well-chosen evidence and critical commentary. • Develops a strong central thesis that is sustained throughout the essay. • Balances multiple viewpoints or anticipates counterarguments effectively.
Written Expression	*This criterion assesses the clarity, coherence, and academic tone of the student's writing. Educators may select one or more of the following descriptors when designing a holistic judgement:* • Writes in a clear, concise, and fluent academic style. • Structures the essay logically, with clear introduction, development, and conclusion. • Uses transitions and signposting to guide the reader through the argument. • Demonstrates correct grammar, spelling, and punctuation. • Uses discipline-appropriate terminology with accuracy and precision. • Communicates complex ideas in an accessible and engaging manner.

Further resources: Assessment rubrics from Learning Environments

References

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜.” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445922.

De Boer, Ivo, Femmie De Vegt, Helma Pluk, and Mieke Latijnhouwers. 2021. Rubrics – a Tool for Feedback and Assessment Viewed from Different Perspectives: Enhancing Learning and Assessment Quality. IAMSE Manuals. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-86848-2.

Lappin, Shalom. 2024. “Assessing the Strengths and Weaknesses of Large Language Models.” *Journal of Logic, Language and Information* 33 (1): 9–20. https://doi.org/10.1007/s10849-023-09409-x.

Pearl, Judea. 2018. “Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution.” arXiv. http://arxiv.org/abs/1801.04016.

Wang, Boshi, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. “Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters.” arXiv. https://doi.org/10.48550/arXiv.2212.10001.