THE UNIVERSITY OF HONG KONG · DEPARTMENT OF GEOGRAPHY

"Assessment Design" series · Faculty Sharing Session

Consistency-Weighted Writing

A Pro-AI Approach to University Writing Assessment

Thursday, 5 November 2026 · 1:00–2:00 pm
Zoom

Prof. Patrick Adler
Assistant Professor, Department of Geography
Faculty of Social Sciences, The University of Hong Kong

adler@hku.hk

Jack of All Trades



Presentation Overview

Consistency-Weighted Writing

A Pro-AI Approach to University Writing Assessment

I Explaining the System

II Motivation and Learning Objectives

III Lessons from a Case Study

IV Vulnerabilities and Questions

I

Explaining the System

The assessment design

Consistency and weighting

Qualitative and automated approaches



In Brief

Reward students who can maintain their analysis, recall and voice in multiple contexts.

Part A: prepared work

Time to research, revise and use the tools authorised for the task.

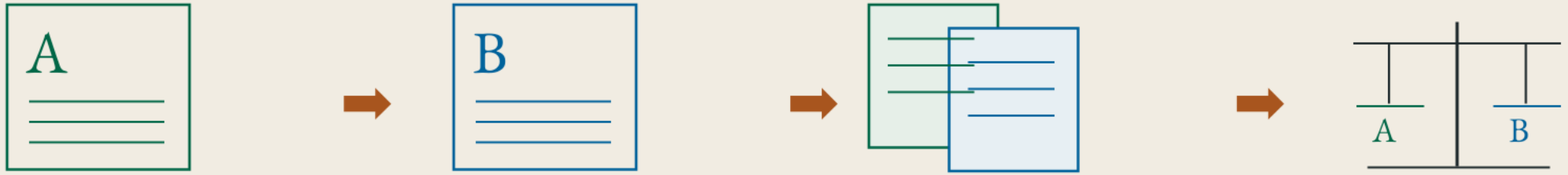
Part B: follow-up

A second context tests what the student can still explain and apply.

The parts can also use interviews, speeches or debates, depending on the learning objectives.

The Weighting Rule

The exact weighting scheme can vary; it should favour Part A when consistency is demonstrated.



Prepared work

Planned communication
Published AI-use rules

Follow-up

Recall and transfer
Specified resources

Style + content

Voice, argument, evidence and methods

Combined mark

Flexible weights
Favour Part A when consistency is shown

Choose and publish the exact weights before students complete the assessment.

Qualitative and Automated Approaches

Both versions ask whether the substance of A remains available in B.

Qualitative judgment

Read A and B side by side.

Use explicit dimensions and paired excerpts to justify the consistency judgment.

Automated matching

When assignments are equivalent, ask a blinded LLM to match A texts to B texts.

Test whether the pairings are reliable.

Vibes-Based Grading: Dimensions of Consistency

"Vibes-based" judgment becomes more useful when the reader states what carries across.

Overall persuasiveness	Does the writer retain the thesis and the logic supporting it?
Overall fluency	Does a recognisable voice persist across writing conditions?
Factual specificity and accuracy	Are examples and factual claims recalled precisely and correctly?
Literature specificity and accuracy	Do named authors, positions and disagreements remain available?
Research-design specificity and accuracy	Can the writer explain variables, comparisons, mechanisms and limits?

Validating Automated Matching

Shuffle and hide

Remove names, IDs and file order; mask topic cues. Keep the true pairs in a hidden key. Masking may leave residual cues.

↓

Match and score

Score all 18 × 18 A–B combinations and choose a one-to-one pairing. Compare with the hidden key: random per-student accuracy is 1/18.

↓

Check against people and controls

Compare with human readers, repeated in-class pieces and AI-generated controls. Test the expected differences rather than assume them.

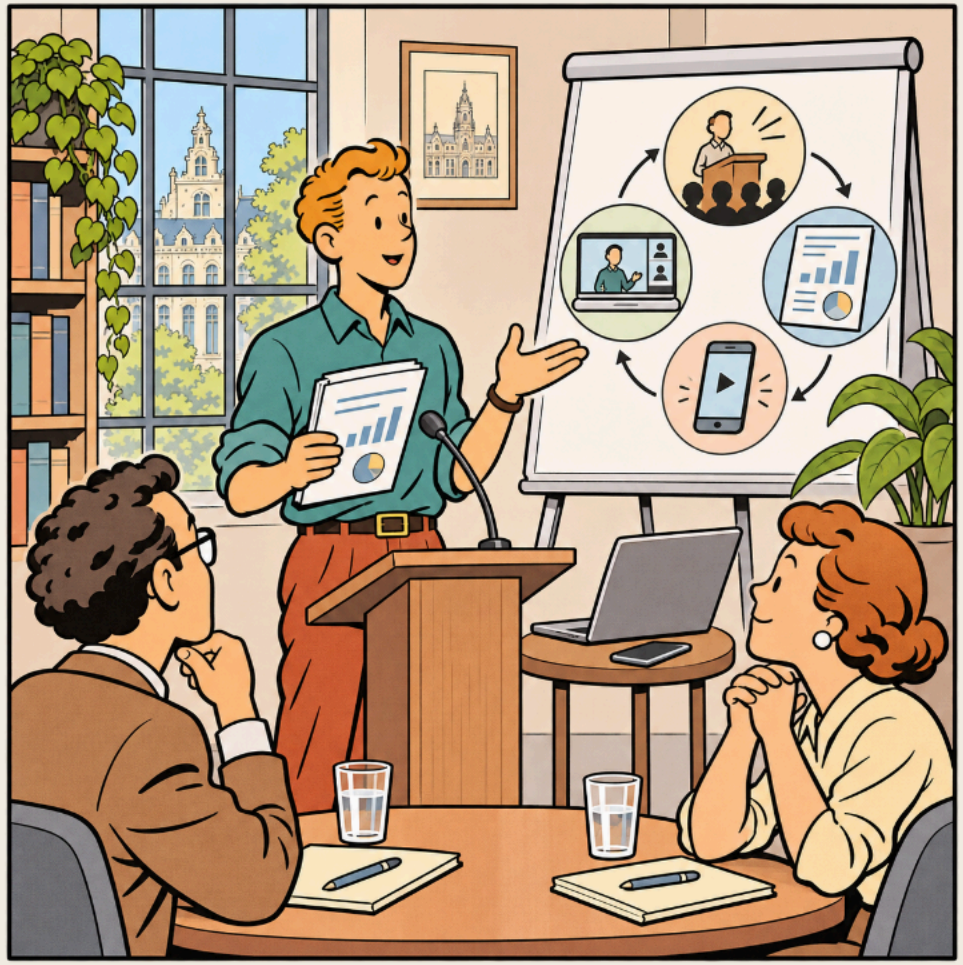
II

Motivation and Learning Objectives

Objective 1: Teach and reward consistency

Objective 2: Legitimise AI-assisted essays

Objective 3: Discourage low-quality AI use

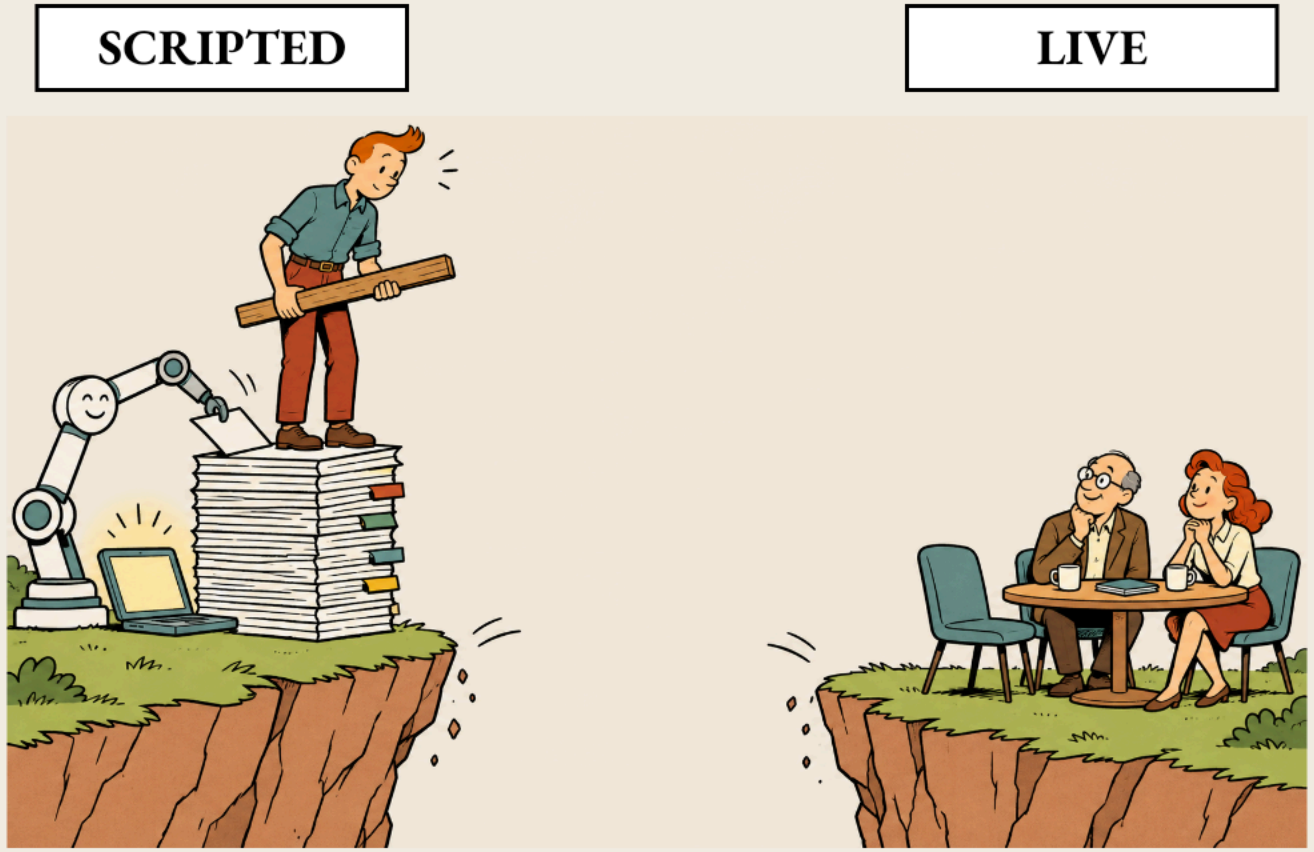


AI widens quality gaps between scripted and live communication

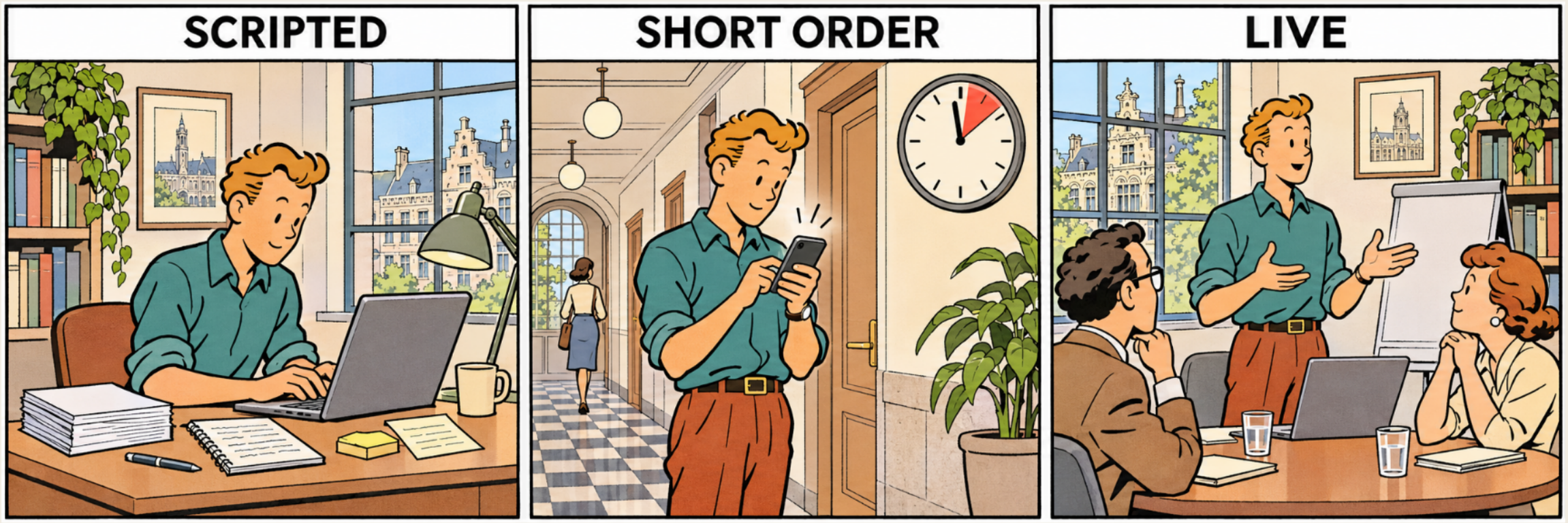
Work already moves between formats: a report, a hallway question, a meeting.

AI makes the written version much better and leaves the live version where it was.

That gap creates doubt, even about honest AI-assisted work. Consistency becomes worth teaching.



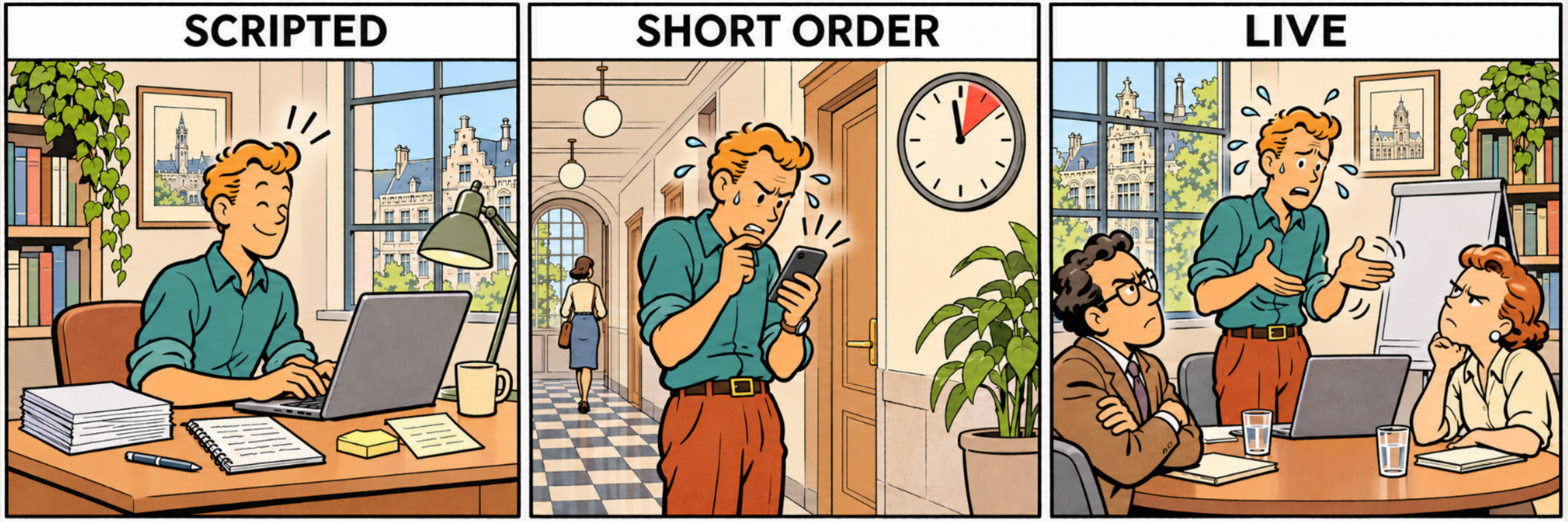
Jack of All Trades



Effective communication carries across scripted, short-order and live settings.

Suspected Becket

Polished on the page, less convincing in the moment.



A gap can cast doubt on scripted work, even when that suspicion is unfair.

We should train consistent communication because it is in demand and because it increasingly signals legitimacy, credibility, and trustworthiness.

Objective 1: Teach and Reward Consistency

Cultivate consistent communicators.

Practise carrying a thesis and its evidence into shorter, less scripted responses. Consistency protects the weight of a stronger take-home product.

Course assessment traditionally blends assessment types (implicit consistency) without explicitly promoting consistency.

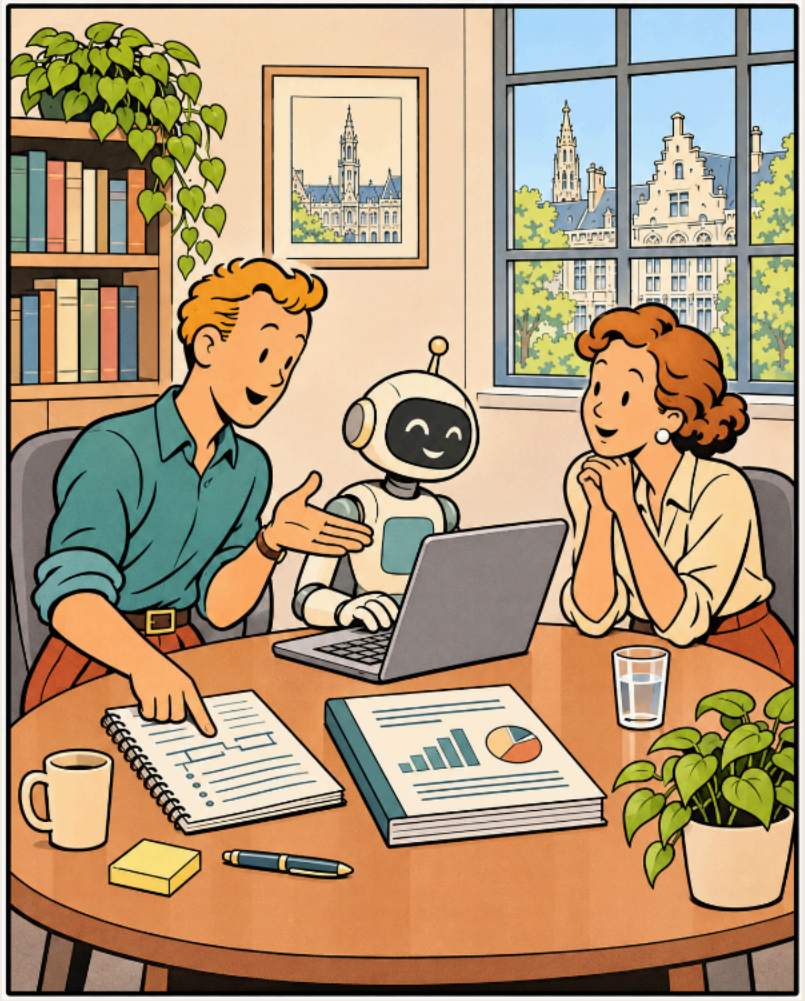


Objective 2: Legitimise AI-Assisted Essays

Co-produced essays need a credible place in assessment.

Currently, take-home assignments have lower status than they should. Suspicion of AI abuse leads instructors to overlook talented writing and co-production.

The move away from term papers reflects this trend.



Objective 3: Discourage Low-Quality AI Use

Assess the performance gap without having to decide the writer's intent.

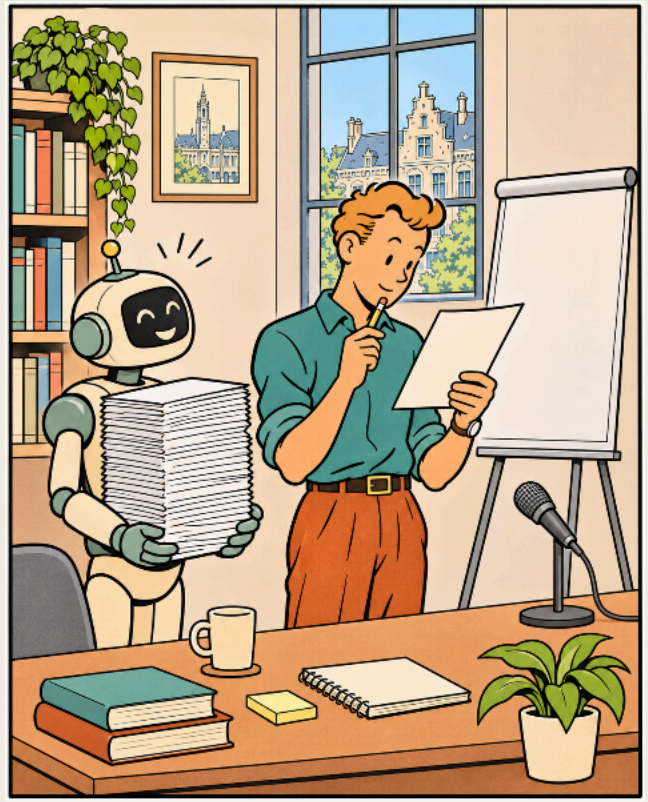
Approach 1: Honor system

Disclosure alone may disadvantage compliant students when undisclosed assistance remains hidden.

Approach 2: Automated detection

False positives and evasion complicate detection. Acceptable assistance can also attract suspicion.

An impressive A earns less weight when the relevant command does not carry into B. Misconduct procedures require their own evidence.



III

Lessons from a Case Study

The assignment and scoring method

Patterns and student examples

Limitations and teaching extensions



The Paired Assignment

20 students · 18 complete A/B pairs · 2 confirmed Part B absences

Take-home essay (A)

A take-home essay on a controversy, a personal project and research that could clarify the dispute.

In-class follow-up (B)

A closed-book, 110-minute response that recaps the controversy and describes a different hypothetical study designed to end it.

Scores are provisional, LLM-assisted appraisals of adherence, fluency, accuracy and sophistication.

Part A restricted substantive AI authorship. The design is meant to extend to openly permitted co-production.

Scope and Limitations of the Pilot

What we observed

- We compared 18 paired submissions.
- We applied an illustrative weighting rule and curve to provisional, LLM-assisted quality scores.
- We used paired quotations to show what students retained or lost.

What remains untested

- The comparisons do not establish authorship, intent or why a gap arose.
- Automated matching has not been run.
- The rubric, cutoffs and weighting rule have not been validated for grading.

Part A restricted substantive AI authorship. Openly permitted co-production is the proposed extension.

Style and Content Consistency: Four Types

	High content consistency	Lower content consistency
High style consistency	Type 1 · 11 students	Type 3 · 1 student
Lower style consistency	Type 2 · 2 students	Type 4 · 4 students

All four students in Type 4 recorded no AI disclosure.

For this assignment, the take-home essay counted 80% when either style or content consistency was high. When both were lower, the in-class follow-up counted 80%.

High consistency = at least 4/5. Disclosure is an observed characteristic, not a separate type.

Calculating the Overall Mark

The scoring rule used in this case study

1. Score each part

Part A
80 / 100

Part B
60 / 100

→

2. Apply consistency weights

Types 1–3: consistency in at least one dimension
A 80% + B 20%
Example: combined mark = 76

Type 4: lower consistency in both dimensions
A 20% + B 80%
Example: combined mark = 64

→

3. Curve and grade

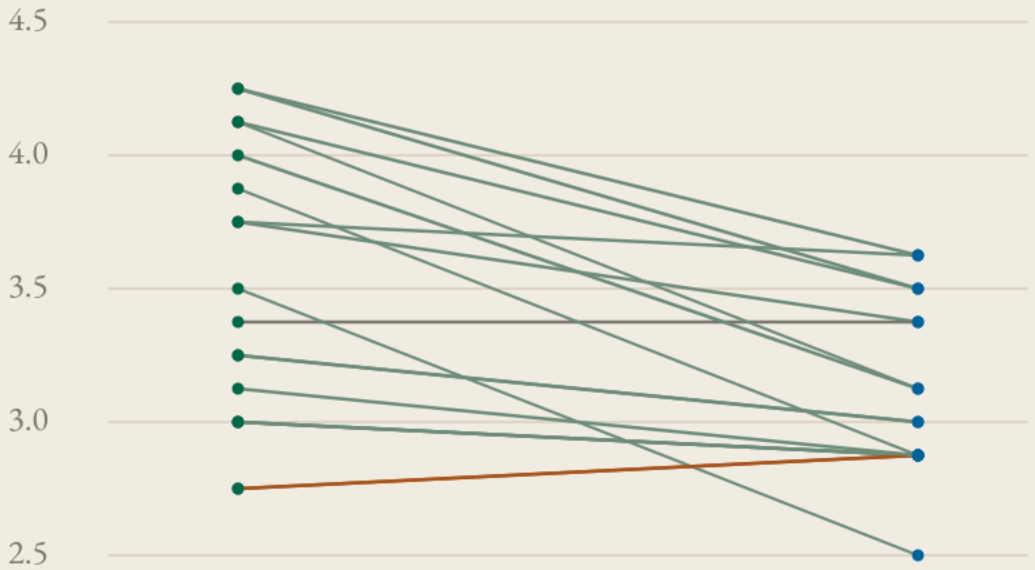
Add 9 points

76 becomes 85 (A)
64 becomes 73 (B)

Content scores reflect assignment fit, accuracy and sophistication. Example inputs shown above.

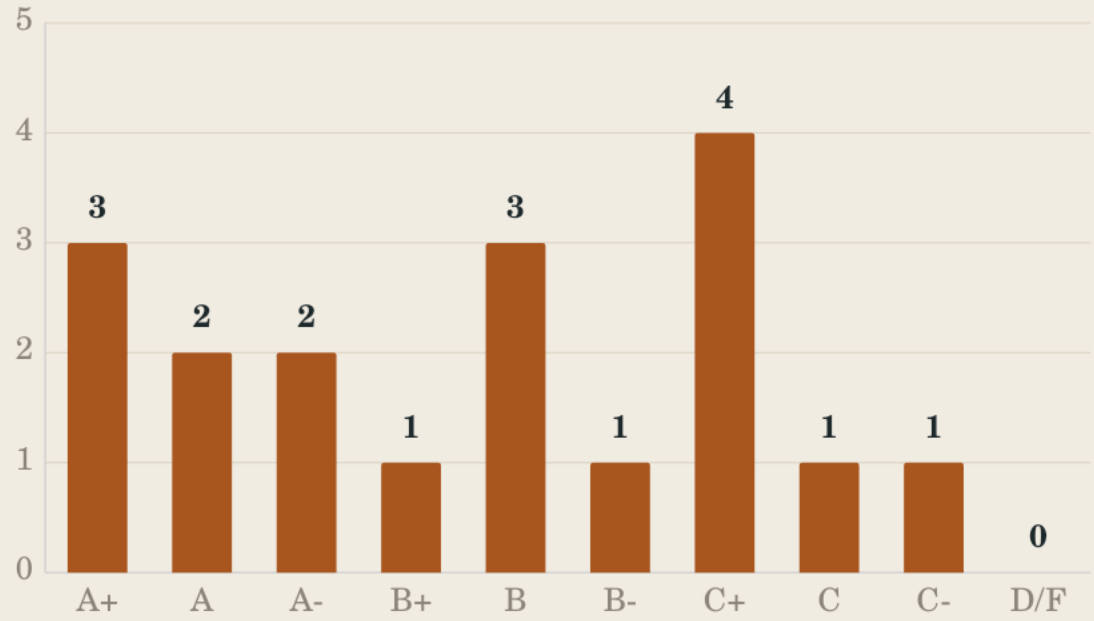
A-to-B Quality Changes and Overall Grades

Four-criterion quality average



15 lower · 1 equal · 2 higher

Curved overall grades



Overall Trends

15 of 18

Quality falls from A to B

Students scored lower in B on the four-criterion quality average: 3.51 to 3.10 out of 5. Continuity can coexist with a clear difference in polish or quality.

25%

An A-range grade remains possible

One of four students with lower consistency in both style and content received an A−, after consistency weighting and the nine-point curve.

r = −0.15

Little relationship in this sample

No disclosure has a weak negative correlation with final marks (18 pairs). All four students with lower style and content consistency recorded no disclosure. General attentiveness to guidelines is a hypothesis, not an established cause.

Consistency in the Overall Thesis

Student 12

Strong continuity

The deeper test is whether the same explanatory argument survives, including its handling of alternative outcomes.

Take-home essay (A)	**In-class follow-up (B)**	*Reading notes*
"Actions and relations **generate structures** that, once sufficiently durable, reshape the opportunity set for later action."	"Economic actions produce spatial resulting structure, while **accumulated structures** subsequently provide opportunities available to economic actors."	*The same recursive argument* Both explain how relations produce structures that shape later action. The explanatory mechanism survives.
A, p.6	B, script p.1	
"The **two outcomes** locate persistence at different points in the recursive relationship between economic relations and accumulated spatial structures."	"The second result **would not invalidate** the relational perspective. Relational economic geography does not require constant change or predicts that every disturbance produce new trajectory."	*The same alternative-outcome logic* Both explain why restoration can fit a relational account. The distinction between a theory and one predicted result persists.
A, p.5	B, script p.3	

Theoretical sophistication persists in the mechanism and the interpretation of rival findings.

Clear Consistency Across A and B

Student 01

Consistency 4.50 / 5

Distinct studies: A mines online NEV comments. B surveys attitudes to vehicle-to-grid technology.

Take-home essay (A)	**In-class follow-up (B)**	*Reading notes*
“the **hydrogen safety paradox**, which refers to the concerns of the explosive nature of hydrogen fuel cells in HFCVs”	“the "**hydrogen safety paradox**" (worry of the explosions of hydrogen fuel cells)”	*The same specific example* The recap preserves a distinctive hydrogen-risk example and its role in the debate.
A, p.2	B, script p.1	
“the **VIF** value should be kept below 5 ... the **p-value** should be kept within 0.05 ... This ensures the robustness and reliability”	“The robustness of the regression model will be validated via **p-value** (below 0.05) and **VIF** (below 5).”	*The same inferential shortcut* Both treat thresholds as sufficient assurance of robustness. A shared error makes the correspondence especially specific.
A, p.6 · ellipses omit text	B, script p.3	

Specific conceptual recall and the same reasoning habit persist after a real change of study.

A Strong Essay with Mixed Consistency

Student 17

Consistency 3.25 / 5

A strong take-home essay: 4.5/5 for adherence and sophistication, 4/5 for accuracy.

Take-home essay (A)	In-class follow-up (B)	*Reading notes*
"Strategies A and B lower predicted costs using standardised and locally augmented information. Strategies C and D use the same data but have **distribution safeguards**."	"For both scenarios, total costs and environmental effects e.g. carbon emissions and biodiversity gain and loss are quantified."	*Some continuity remains* A separates information from objectives. B bundles local knowledge with ethical benefits.
A, p.4	B, script p.3	
"Such payments would **not be assumed** to remedy involuntary losses."	"With the recognition and participation of all stakeholders, especially the vulnerable communities, **compensations for local people are exempted**."	*A substantive qualification weakens* A qualifies what payments remedy. B infers waived compensation from participation without establishing consent.
A, p.8	B, script p.4	

A's careful qualifications are only partly sustained in B. Lower syntactic control adds to the gap.

High Consistency at a Lower Level of Quality

Student 04

Consistency 4.25 / 5

A and B both score 2.5/5 for syntax and 3/5 for accuracy. The correspondence is stronger than the quality.

Take-home essay (A)	In-class follow-up (B)	*Reading notes*
"The pattern of archaeological data **be considered** as the past pattern"	"This research result can **be further adopt** by the relics department in China."	*Comparable verb-form errors* Both leave the passive verb form incomplete. The grammatical limitation recurs across writing conditions.
A, p.1 · sentence excerpt	B, script p.2	
"we will measure **publication intensity** independently of stock size, divide each province's publication by its known site stock."	"we can calculate the **publication intensity** with the renewed scale, to have a control variable when we predict the population."	*The same data logic* The same unusual adjustment persists in a new application. Both texts leave causal claims underdeveloped.
A, p.4	B, script p.2	

The same data knowledge and recurrent language limitations support a high consistency judgment.

Factual Details Retained in the Follow-Up

Student 04

The comparison preserves a named author and year, plus the location of another cited study.

Take-home essay (A)	In-class follow-up (B)	*Reading notes*
"**Rick (1987)** believed that under a systematic sampling method, we could use proxies from archaeology to reflect the population trend"	"Back to **1987, Rick** proposed his idea during a systemic survey ... radiocarbon dating samples are a good proxy for population."	*Author, year and proposition* Rick and 1987 recur, with the same link between archaeological samples and population.
A, p.1	B, script p.1 · ellipsis omits text	
"Codd et al. (2024) used the radiocarbon dates and ethnographic population density from **California**"	"Codding use the population data from **California**."	*The same empirical case* The author naming varies, but California remains the location of the population-validation example.
A, p.2	B, script p.1	

These are specific recalled details, beyond repeating the broad controversy.

A Factual Inconsistency in the Follow-Up

Student 06

B explicitly recalls A's five cities. Four names carry across, but the Ukrainian city changes.

Take-home essay (A)

"Donetsk, Ukraine"

Table 1 identifies the Ukrainian case as Donetsk. The later case discussion uses the same city.

A, pp.3, 12–13

In-class follow-up (B)

"Part A five cities (Detroit, Yubari, Jilin, Shaoyang, **Kyiv**)"

B, script p.3

Reading notes

A concrete mismatch

A says Donetsk. B says Kyiv while naming the earlier study, rather than proposing a new case.

How much weight should this carry?

The discrepancy is clear, but local. The other four cities agree, and the overall consistency score remains 4.00/5. A memory slip is a plausible explanation.

A factual inconsistency can coexist with substantial overall correspondence.

Consistency in Research Methods

Student 09

Partial continuity

The research structure and outcome concepts recur, while the analytical detail is less explicit in B.

Take-home essay (A)	**In-class follow-up (B)**	*Reading notes*
"a **Difference-in-Difference (DID)** approach ... combining firm-level panel data with spatially explicit climate extremes ..." A, p.2 · ellipsis omits text	"This study will use **description analysis and time-series analysis**. We will collect all data related to pre-disaster and post-disaster periods." B, script p.1	*Causal design is less explicit* A specifies DID and panel data. B describes before/after observations without an explicit control comparison.
"**resource-firm networks** will be constructed to capture resource sharing relations" A, p.3 · sentence excerpt	"we will conduct field-based **observation and interviews**." B, script p.2	*Mechanisms remain, detail changes* Both investigate resources and relationships. A defines networks analytically, while B lists observations and interview topics.

Compare variables, comparisons and inferential steps, as well as the names of techniques.

Specificity in Describing the Literature

Student 03

Substantial continuity

The question is whether B retains particular authors, their positions and the distinctions between them.

Take-home essay (A)	**In-class follow-up (B)**	*Reading notes*
"It also distinguishes between a 'generalising perspective' and a '**context-specific perspective**,' giving greater consideration to local knowledge alongside Western science"	"Supporters argue that NCP gives more attention to local knowledge and the different ways people value nature (**Díaz, 2018; Pascual, 2017**; IPBES, 2019)."	*Specific supporting camp* B recalls supporting authors and their local-knowledge claim. A's finer terminology is compressed.
A, p.2 · quotation marks normalised	B, script p.1	
"**Braat (2018)** argues that NCP offers limited conceptual innovation and mainly reflects a political compromise."	"Critics argue that ecosystem services research can already do this (**Braat, 2018; Faith**, 2018, 2021)."	*Specific opposing camp* Braat and Faith remain identified as critics, but B combines objections that A distinguishes.
A, p.2	B, script p.1	

B recalls a structured disagreement between named authors, rather than only the topic.

Accuracy in Describing the Literature

Student 03

Accuracy largely maintained

The same disputed claim: does NCP introduce perspectives that ecosystem-services research already includes?

Take-home essay (A)	**In-class follow-up (B)**	*Reading notes*
"Critics question the extent of these improvements. Braat (2018) argues that NCP offers **limited conceptual innovation**"	"Critics argue that **ecosystem services research can already do this** (Braat, 2018; Faith, 2018, 2021)."	*Same critical position* Both question NCP's claimed novelty. B's preceding sentence specifies local knowledge and different ways of valuing nature.
A, p.2 · sentence excerpt	B, script p.1	

Independent source check

Braat's editorial argues that ecosystem-services research already includes social-science perspectives and attends to stakeholder interests and cultural knowledge. This supports the shared criticism of claimed novelty.

B preserves the direction of the criticism, while summarising it more briefly.

IV

Vulnerabilities and Questions

Subjective style judgments

Permitted editing and fairness

Assessment design and teaching



Vulnerability: Subjective Style Judgments

Readers may disagree about whether two texts sound like the same writer.

- Ground each judgment in paired passages rather than a general impression.
- Use moderation to check whether different readers reach similar conclusions.
- Workplace judgments across formats do not by themselves establish fair grading criteria.

Vulnerability: May Penalize Permitted Assistance

Using tools in Part A can potentially exaggerate the gap.

- Proofreading, translation support and grammar tools can improve the take-home essay.
- Such assistance can make the in-class follow-up sound less polished even when the student has retained the argument.
- Judge the permitted assistance and intended learning outcome for each task.

Vulnerability: The Consistency Cutoff

Illustration: take-home essay = 80, in-class follow-up = 60. Hold the other decisions fixed.

Below the consistency cutoff

A 20% / B 80% = 64

At the consistency cutoff

A 80% / B 20% = 76

A borderline call can change this mark by 12 points. Moderation, a review band or sliding weights deserve testing before use.

Teaching Consistency Across Formats

Practise before grading.

Learn to use assistance while retaining command of what you submit.

Extend the format

Pair a report with an oral explanation, presentation questions or a structured interview.

Rehearse the transfer

Restate a thesis, explain one method, defend a source and respond to a new objection.

Teach the comparison

Use matched A/B examples, give feedback on gaps and repeat before the graded task.

Two Pedagogical Questions

1 Does consistency-weighted scoring reward memorization too much?

(Tradeoff between grading consistency and incentives to memorize)

2 What are the guidelines for consistent writing?

(Open question)

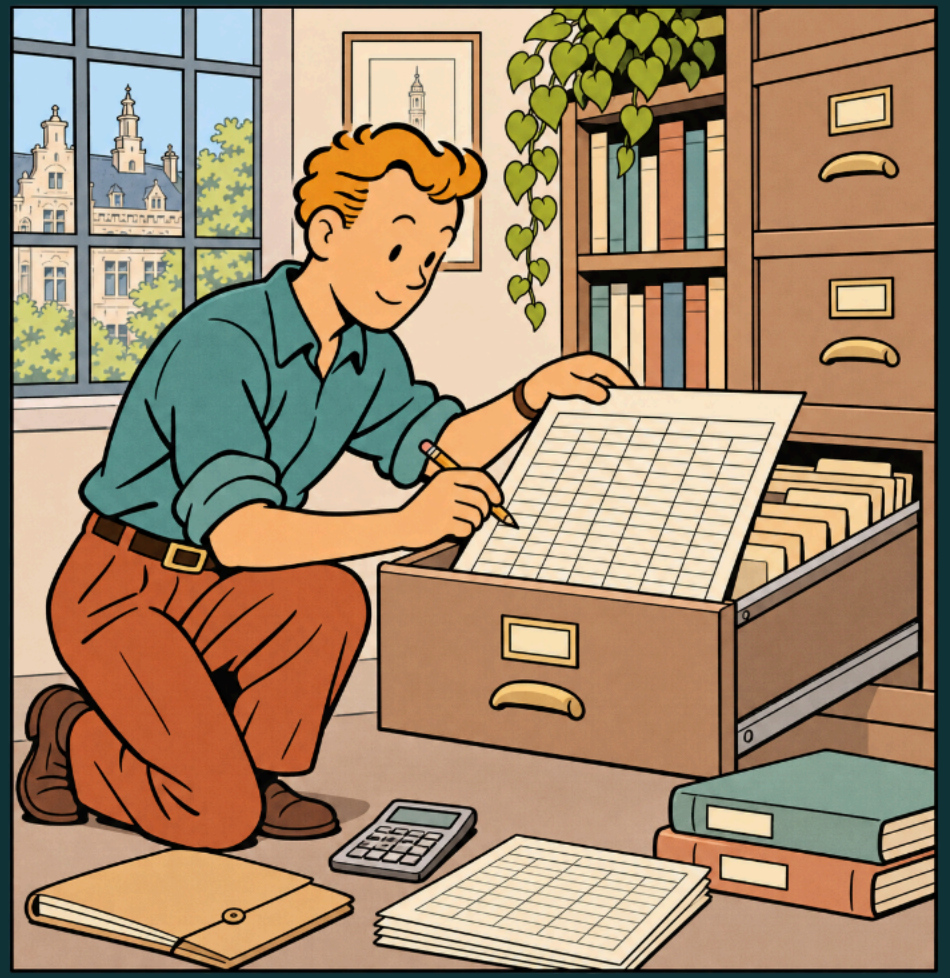
Questions?

adler@hku.hk

APPENDIX

Detailed scoring tables

Anonymised marks by consistency type and overall score.



A1 Scores by consistency type

Types 1–3: A 80%, B 20%. Type 4: A 20%, B 80%. Add 9 points to each weighted total, giving a top mark of 90.

Student	Consistency type	A content /100 (grade)	B content /100 (grade)	Style consistency	Content consistency	Overall consistency	AI in A	A:B weight	Curved total /100 (grade)
01	1 High consistency in style and content (11)	77 (B+)	70 (B-)	5	4.5	4.5	None	80:20	**85 (A)**
02		73 (B)	73 (B)	4	4.5	4.25	Full prompts	80:20	**82 (A-)**
03		83 (A-)	73 (B)	4	4.5	4.25	Substantial	80:20	**90 (A+)**
04		63 (C)	60 (C-)	5	4	4.25	Limited	80:20	**71 (B-)**
05		67 (C+)	63 (C)	5	4	4.25	None	80:20	**75 (B)**
06		50 (D)	57 (D+)	4	4	4	Substantial	80:20	**60 (C-)**
07		67 (C+)	70 (B-)	4	4	4	None	80:20	**77 (B+)**
08		60 (C-)	60 (C-)	4	4	4	None	80:20	**69 (C+)**
09		83 (A-)	63 (C)	4	4.5	4	None	80:20	**88 (A)**
10		83 (A-)	73 (B)	4	4	4	Substantial	80:20	**90 (A+)**
11		60 (C-)	60 (C-)	4	4	3.75	None	80:20	**69 (C+)**
12	2 Lower style consistency / high content consistency (2)	83 (A-)	73 (B)	3	4.5	4	Full prompts	80:20	**90 (A+)**
13		70 (B-)	50 (D)	2	4.5	3.5	Full prompts	80:20	**75 (B)**
14	3 High style consistency / lower content consistency (1)	57 (D+)	53 (D)	4	2.5	2.75	Limited	80:20	**65 (C)**
15	4 Lower consistency in style and content (4)	80 (A-)	60 (C-)	3	3.5	3.25	None	20:80	**73 (B)**
16		53 (D)	60 (C-)	3	3.5	3.25	None	20:80	**68 (C+)**
17		87 (A)	67 (C+)	2	3.5	3.25	None	20:80	**80 (A-)**
18		57 (D+)	60 (C-)	2	2.5	2.75	None	20:80	**68 (C+)**

Style consistency = syntax/voice match. Content consistency = concepts/reasoning match. High consistency = 4/5 or above.

Type 4: lower style and content consistency. All four students recorded no AI disclosure.

Content = 20 × mean(adherence, accuracy, sophistication). Weight displayed A/B marks, round, then add 9 points and assign the letter.

A2 Scores by overall mark

Uniform curve: weighted total + 9 points. Students 03, 10 and 12 share the top score of 90 (A+). Student numbers stay fixed.

Rank	Student	Type	A content /100 (grade)	B content /100 (grade)	A:B weight	Weighted before curve	Curved total /100 (grade)	AI disclosure in A
1	03	1	83 (A-)	73 (B)	80:20	81	**90 (A+)**	Substantial
1	10	1	83 (A-)	73 (B)	80:20	81	**90 (A+)**	Substantial
1	12	2	83 (A-)	73 (B)	80:20	81	**90 (A+)**	Full prompts
4	09	1	83 (A-)	63 (C)	80:20	79	**88 (A)**	None
5	01	1	77 (B+)	70 (B-)	80:20	76	**85 (A)**	None
6	02	1	73 (B)	73 (B)	80:20	73	**82 (A-)**	Full prompts
7	17	4	87 (A)	67 (C+)	20:80	71	**80 (A-)**	None
8	07	1	67 (C+)	70 (B-)	80:20	68	**77 (B+)**	None
9	05	1	67 (C+)	63 (C)	80:20	66	**75 (B)**	None
9	13	2	70 (B-)	50 (D)	80:20	66	**75 (B)**	Full prompts
11	15	4	80 (A-)	60 (C-)	20:80	64	**73 (B)**	None
12	04	1	63 (C)	60 (C-)	80:20	62	**71 (B-)**	Limited
13	08	1	60 (C-)	60 (C-)	80:20	60	**69 (C+)**	None
13	11	1	60 (C-)	60 (C-)	80:20	60	**69 (C+)**	None
15	16	4	53 (D)	60 (C-)	20:80	59	**68 (C+)**	None
15	18	4	57 (D+)	60 (C-)	20:80	59	**68 (C+)**	None
17	14	3	57 (D+)	53 (D)	80:20	56	**65 (C)**	Limited
18	06	1	50 (D)	57 (D+)	80:20	51	**60 (C-)**	Substantial

Types: 1 high style + content consistency; 2 lower style / high content consistency; 3 high style / lower content consistency.

Type 4: lower consistency in both. All four recorded no AI disclosure. Types describe consistency, not content quality.

Students 19–20: Part B absent, so no overall mark or rank. Ties share a rank; student number orders tied rows.