Comparing AI Tools and Human-Graded Literature Review Scores in a Veterinary Research Course

Dr. Ibrahim Elsohaby *(DVM, MVSc, GradCert One Health, PhD)*

Assistant Professor of Public Health and Epidemiology
City University of Hong Kong

Email: ielsohab@cityu.edu.hk

PetEpiLab

VetEd ASIA 2025 – Veterinary Education Research Session
23rd AAVS Annual Meeting, Yogyakarta, Indonesia
8 November 2025

香港城市大學
City University of Hong Kong

Contents

香港城市大學
City University of Hong Kong

01 Introduction

Introduction

01 Literature reviews

- Synthesizing and communicating scientific information is a core competency in veterinary education.
- Literature reviews are a key tool for developing these critical thinking and writing skills.
- Despite advances in AI, automated grading of scientific writing remains largely unexplored in veterinary contexts.

(Van Der Vleuten, 1996)

Introduction

02 Large language models (LLMs)

- Modern large language models (LLMs) like ChatGPT enable zero-shot grading, requiring no prior training, making them accessible and easy to use.
- LLMs are being studied across disciplines (e.g., medicine, law, humanities), but their accuracy, reliability, and fairness vary.
- Rigorous, context-specific evaluation is essential before integrating LLMs into veterinary education assessment.

(Choi et al., 2021; Kung et al., 2023)

香港城市大學
City University of Hong Kong

02 Objectives

Objectives

1

Assess the reliability and agreement between AI and human grading

To evaluates how consistently and accurately four large language models (ChatGPT, Qwen, DeepSeek, and Copilot) align with expert human raters in grading veterinary literature reviews.

2

Explore student perceptions of receiving AI-generated feedback

To understand how veterinary students experience and evaluate feedback generated by artificial intelligence on their academic writing, particularly literature reviews.

香港城市大學
City University of Hong Kong

03 Materials & Methods

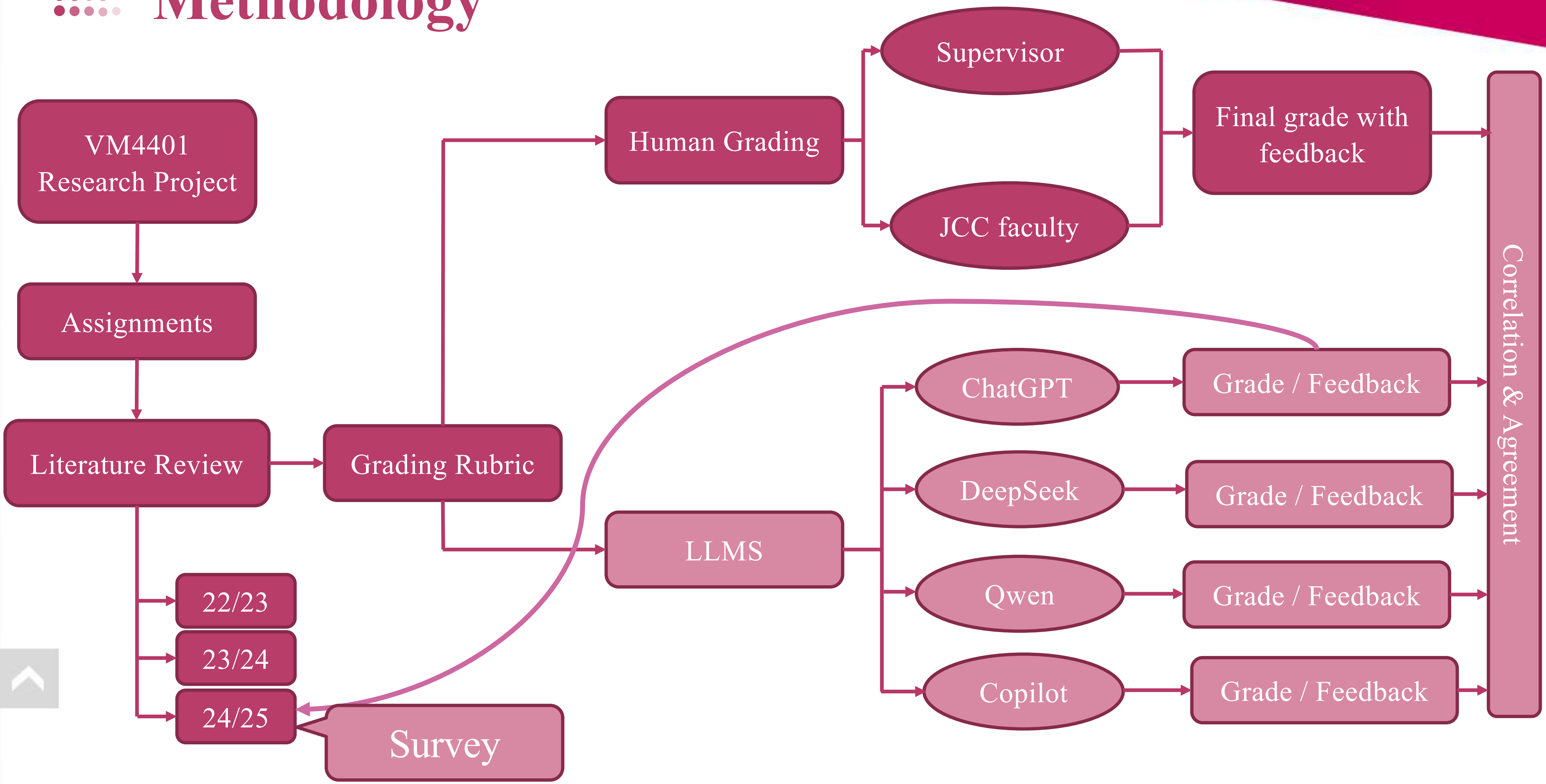
Materials

VM4401 - Research Project

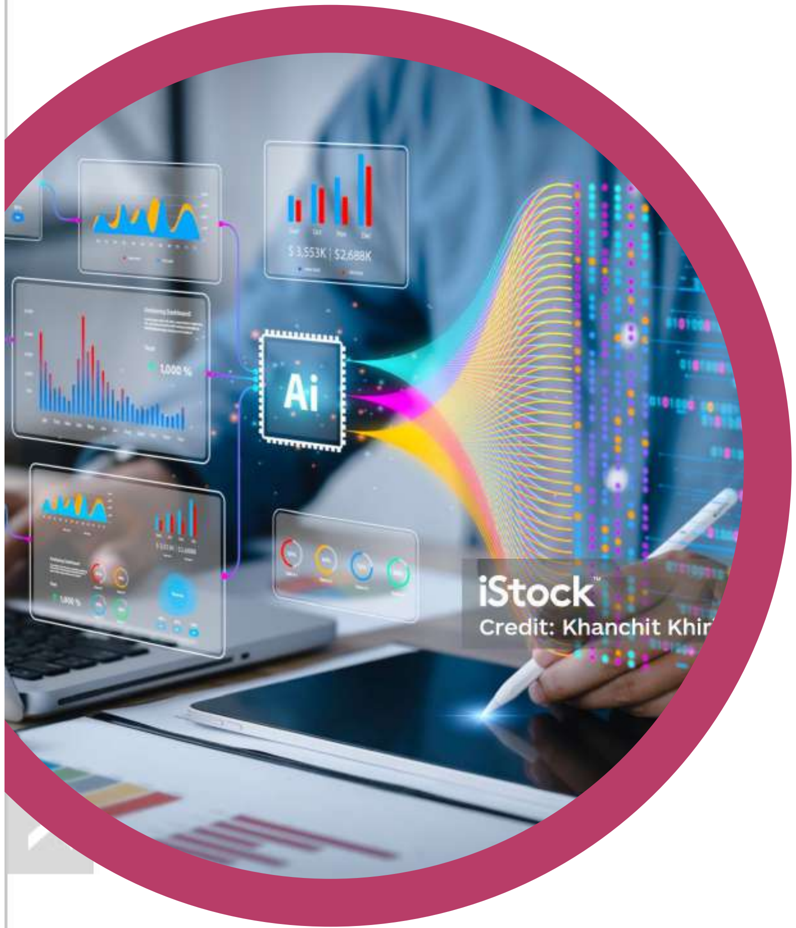
Prof. Colin McDERMOTT
Clinical Assistant Professor
Department of Veterinary Clinical Sciences

- ❑ Fifth year BVM students
- ❑ Two semesters (A and B)
- ❑ CityU/JCC faculty provided a pool of research topics.
- ❑ Students select research topics matching their interests
- ❑ Course assessment (write a literature review).
- ❑ Grading rubric (on a scale from 1 to 100)
  - ✓ Content compliance (20 points),
  - ✓ Reference formatting (10 points),
  - ✓ Quality of references (10 points),
  - ✓ Scientific content (30 points),
  - ✓ Clarity of writing (10 points),
  - ✓ Organized progression (10 points),
  - ✓ Spelling, punctuation and grammar (10 points).

Methodology



Methodology



01 Study Design

- ✓ Cross-sectional study
- ✓ Literature reviews **(N = 61)**
- ✓ Three academic years (2022/23–2024/25)

02 Human Grading

- ✓ Two CityU/JCC faculty members
- ✓ Reviewers applied the rubric independently
- ✓ The final score (mean of the two reviewers)

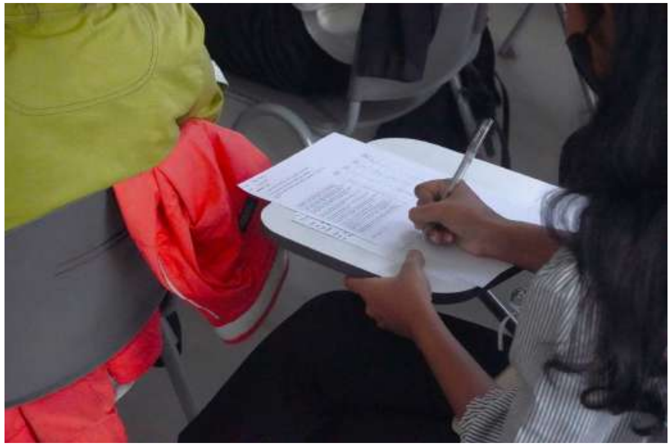
03 AI Grading

- ✓ Four LLMs, including ChatGPT-4o (OpenAI), Qwen 2.5 (Alibaba Cloud), DeepSeek R1 (DeepSeek), and Copilot (Microsoft).

Methodology

04 Student Survey

- 2024/25 cohort (n = 23 students)
- Survey included 22 questions:
  - ✓ Experience with AI-generated feedback
  - ✓ Comparison with human feedback
  - ✓ Specific feedback on AI tools



05 Statistical analysis

- Descriptive statistics
- Wilcoxon signed-rank test
- Spearman's correlation coefficients (ρ)
- Bland-Altman plots

香港城市大學
City University of Hong Kong

04 Results

Human grading

- ✓ The frequency distributions of scores assigned by each reviewer
- ✓ Reviewer scores showed a weak correlation (***ρ*** **= 0.21**)
- ✓ Bland-Altman plot indicated no systematic bias, with a mean difference of **+1.1 point**.

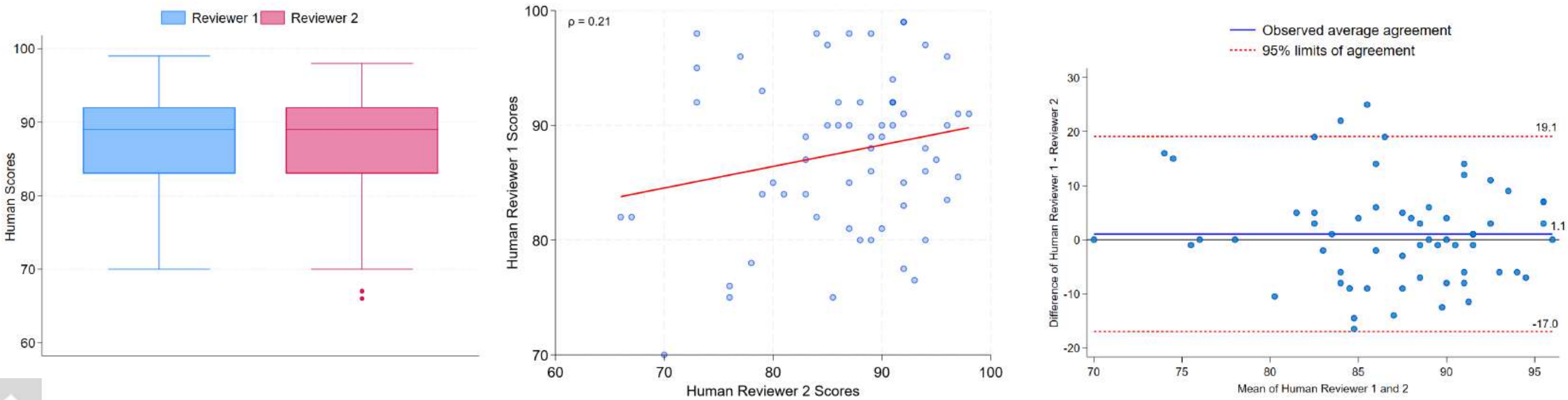


Figure 1. Human grading scores assigned by two reviewers. (**A**) Boxplots showing the distribution of scores assigned by each reviewer; (**B**) Scatter plot depicting the relationship between the scores assigned by the two reviewers; and (**C**) Bland–Altman plot illustrating the level of agreement between the reviewers' scores.

AI grading

- ✓ Scores from AI models significantly different compared to human grading (P <0.05).
- ✓ ChatGPT and Copilot (P = 0.09) and Qwen and DeepSeek (P = 0.21).
- ✓ ChatGPT and Copilot higher scores than human.
- ✓ Qwen and DeepSeek lower scores than human.

Table 1. Descriptive statistics of the grading scores from human reviewers and AI models.

Grading tools	Mean	SD	25%	Median	75%	*P*-value[2]
Human[1]	87.1	5.7	84	88	91	--
ChatGPT	89.6	1.6	89	90	91	0.005
Qwen	84.6	6.0	80	85	90	0.013
DeepSeek	83.1	5.8	81	84	86	<0.001
Copilot	90.5	4.3	87	91	94	<0.001

[1] Represent average scores of two reviewers.
[2] *P*-value from Wilcoxon signed-rank test in comparison with human scores.

Correlations between Human and AI scores

- ✓ All AI models showed weak correlations with human scores.
- ✓ Copilot was the only model that showed significant correlation (ρ = 0.39, P = 0.05)

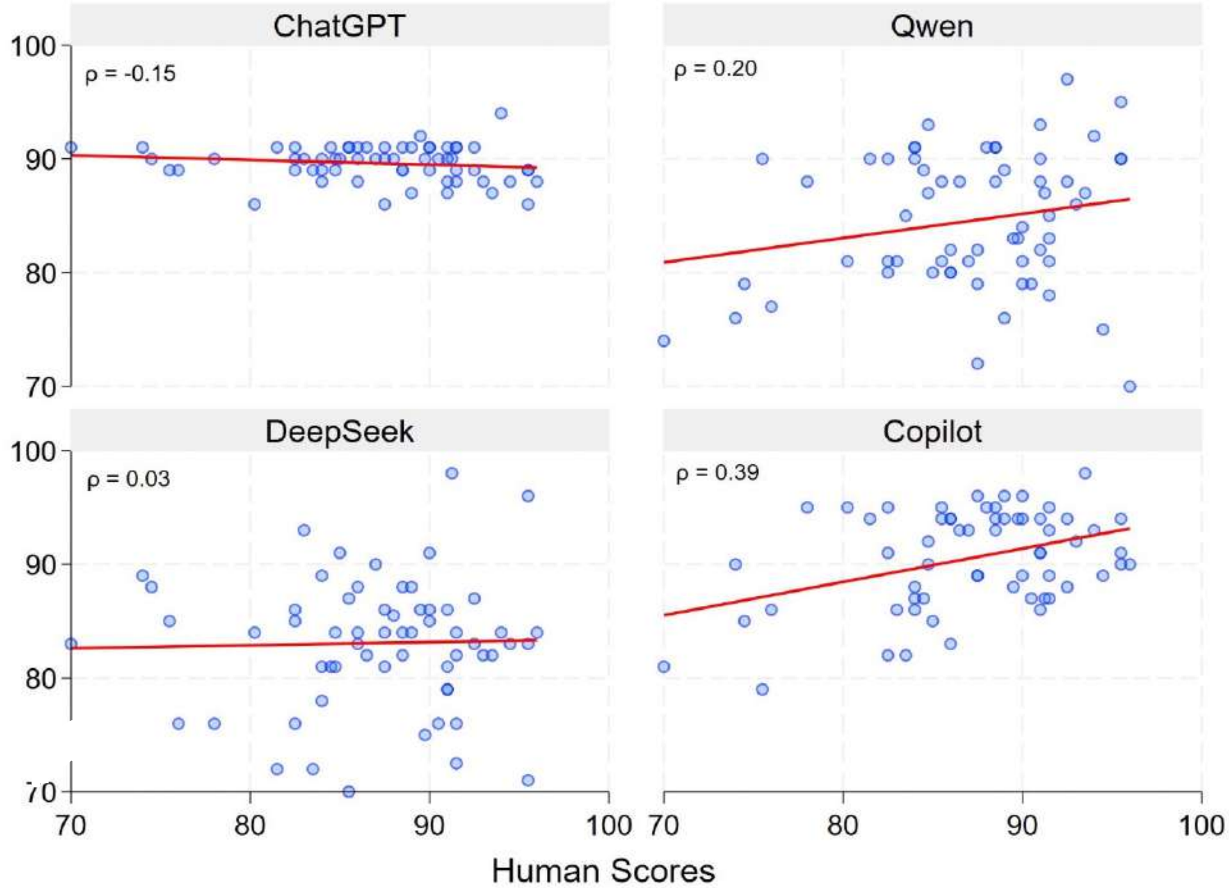


Figure 2. Scatter plots illustrating the relationships between grading scores assigned by human reviewers and those generated by AI models.

Agreement between Human and AI scores

- ✓ All AI models showed some bias and variability
- ✓ Overestimate scores
  - ○ **ChatGPT** (mean bias = **+2.5**)
  - ○ **Copilot** (mean bias = **+3.4**)
- ✓ Underestimate scores
  - ○ **Qwen** (mean bias = **–2.6**)
  - ○ **DeepSeek** (mean bias = **–4.1**)
- ✓ **DeepSeek** exhibited the greatest variability

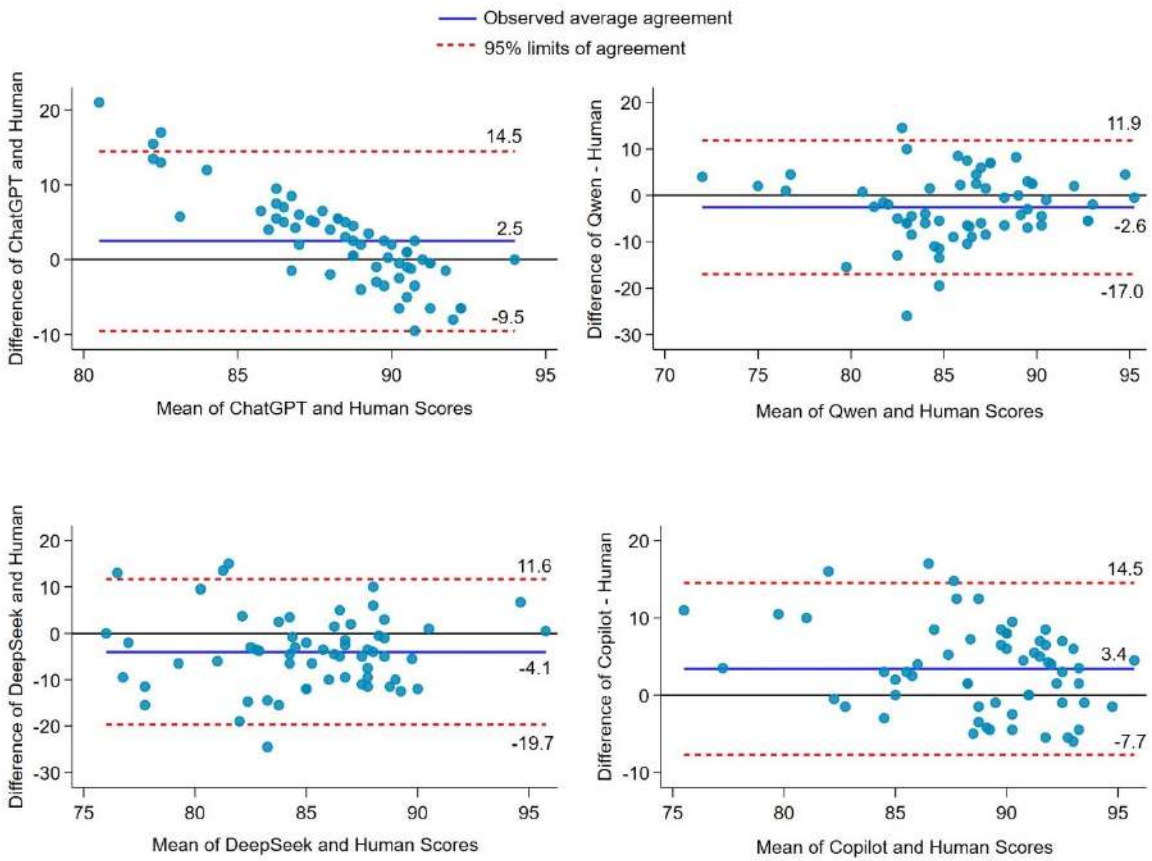


Figure 3. Bland-Altman plot illustrating the agreement between grading scores assigned by human reviewers and those generated by AI models.

Student feedback

- ✓ Participation rate: 78.3% (18/23)
- ✓ **55.6%** had prior experience with AI-generated feedback.
- ✓ **88.8%** rated clarity of AI reports as “Good”
- ✓ **61.1%** rated accuracy of AI reports as “Good”
- ✓ **61.1%** rated the depth of analysis as “Average”
- ✓ **72.2%** found AI feedback “somewhat” helpful for identifying strengths and areas for improvement
- ✓ **38.9%** reported being “Satisfied” with the feedback

Questions	Categories	Frequency	%
I. Experience with AI-Generated Feedback			
Have you previously received feedback from AI tools for academic assignments?			
	Yes	10	55.6
	No	8	44.4
How would you rate the clarity of the AI-generated grading report?			
	Excellent	1	5.6
	Good	16	88.8
	Average	1	5.6
	Poor	0	0.0
	Very Poor	0	0.0
How would you rate the accuracy of the AI-generated grading report?			
	Excellent	2	11.1
	Good	11	61.1
	Average	5	27.8
	Poor	0	0.0
	Very Poor	0	0.0
How would you rate the depth of analysis provided by the AI-generated grading report?			
	Excellent	2	11.1
	Good	3	16.7
	Average	11	61.1
	Poor	2	11.1
	Very Poor	0	0.0
Did the AI-generated grading report help you understand your strengths and areas for improvement?			
	Yes, significantly	2	11.1
	Yes, somewhat	13	72.2
	Neutral	2	11.1
	No, not really	1	5.6
	No, not at all	0	0.0
How satisfied are you with the AI-generated feedback you received for your literature review?			
	Very Satisfied	0	0.0
	Satisfied	7	38.9
	Neutral	11	61.1
	Dissatisfied	0	0.0
	Very Dissatisfied	0	0.0

Student feedback

- ✓ 50.0% rated AI feedback as "slightly worse" than human feedback
- ✓ Human feedback seen as:

50% More detailed/thorough

50% More actionable/useful

77.8% More personalized

- ✓ 55.6% found AI feedback "more formal" in tone

Most useful AI aspects	66.7% Identified strengths
	61.1% Highlighted areas for improvement
Least useful AI aspects	55.6% Generic/repetitive content
	16.7% felt AI covered all key areas

83.3% preferred combined AI + human feedback in the future

CityU
香港城市大學
City University of Hong Kong

05 Conclusions

Conclusions

- LLMs (ChatGPT, Qwen, DeepSeek, Copilot) show promise in enhancing efficiency, consistency, and timeliness of grading literature reviews in veterinary education.

- **However, they do not yet match human expertise in reliability or depth of assessment.**

Future work should:

- Evaluate newer LLMs
- Optimize prompts and rubrics
- Develop hybrid AI–human assessment models

A human-in-the-loop framework, with **AI handling linguistic/mechanical feedback** and **humans providing conceptual judgment**, is the most viable path forward for assessing scientific writing in veterinary education.

Thanks for your attention.

For questions!

Dr. Ibrahim Elsohaby

Email: ielsohab@cityu.edu.hk