A Appendix

A.1 Paper overview

Table 1: Overview of evaluation questions, their GLM-based implementation, and corresponding paper sections.

QUESTION	GLM IMPLEMENTATION	IN
How does my autograder compare to humans?	Include both grader and LLM	Question 1
Do my autograder(s) favour their own output?	Include an interaction between grader and LLM	Question 2
Is there a general human vs autograder difference?	Use a hierarchical GLM with grader-level effects nested in grader type (e.g., human vs. autograder).	Question 3
Are some graders more lenient or strict than others?	Estimate individual grader effects; inspect variation across graders.	Question 3
Do some items receive higher or lower scores?	Include item as predictor; test whether some items are systematically easier/harder	Question 4
Do graders disagree more on some items than others?	Include grader-item interaction; test for grader-specific scoring patterns	Question 4
What is the uncertainty around inter-rater agreement metrics?	Simulate scores from the model and compute agreement (e.g., Krippendorff's α) with uncertainty.	Question 4
Do grader(s) favour longer responses?	Include token length (or token length difference) as a predictor.	Question 5
Do my grader(s) exhibit intransitive	Estimate pairwise probabilities and	Question 5
Is my grading scale well calibrated?	Inspect cutpoints from ordered regression to analyse spacing and interpretability of score intervals.	Appendix A.4

A.2 PRIORS

Below are the priors used across the models described in this paper. They were selected to reflect weakly informative assumptions about effect sizes and score thresholds.

- Intercept: $\beta_0 \sim \mathcal{N}(0, 1)$
- Main effects of grader: $\beta^{\text{grader}} \sim \mathcal{N}(0, 1)$
- Main effects of LLM: $\beta^{\text{LLM}} \sim \mathcal{N}(0, 1)$
- Interaction effects: $\beta^{\text{interaction}} \sim \mathcal{N}(0, 1)$
- Group-level mean for grader type: $\mu_{\text{graderType}} \sim \mathcal{N}(0, 3)$
- Group-level standard deviation: $\sigma^2_{\text{graderType}} \sim \text{HalfCauchy}(1)$
- First cutpoint: $c_1 \sim \mathcal{N}(-4.0, 0.2)$
- Cutpoint differences: $c_j - c_{j-1} \sim \text{LogNormal}(-0.5, 0.3)$ for $j = 2, \ldots, K - 1$

To ensure ordered and well-separated cutpoints, the cutpoint differences are shifted by a small constant before summing: $\Delta_j = (c_j - c_{j-1}) + 0.3$.

A.3 SUPPLEMENTARY FIGURES: MODEL COMPARISONS

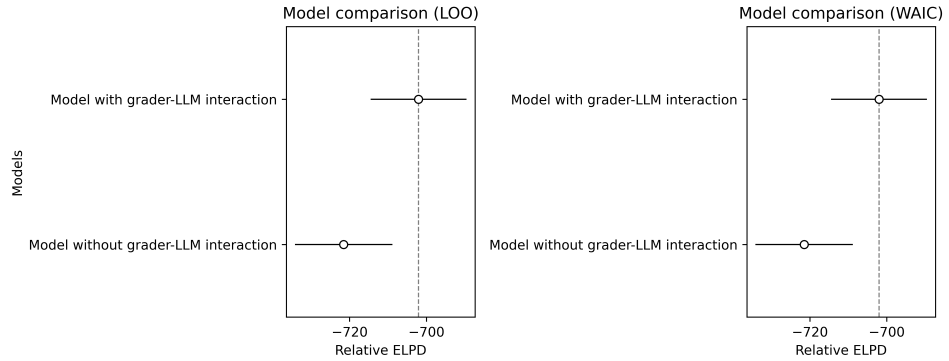

Figure 7: Model comparison for the statistical analysis in Question 2 (Do autograders favour their own generation?). Left panel: Leave-One-Out cross-validation (LOO) scores. Right panel: Widely Applicable Information Criterion (WAIC). Both metrics approximate the Expected Log Predictive Density (ELPD), a measure of predictive accuracy (higher values indicate better performance). Comparing models helps determine whether the added complexity of including an interaction term is justified by improved predictive performance. In this case, the model with the graderLLM interaction (Equation 2) performs slightly better than the model without interaction, supporting a closer examination of potential self-bias effects.

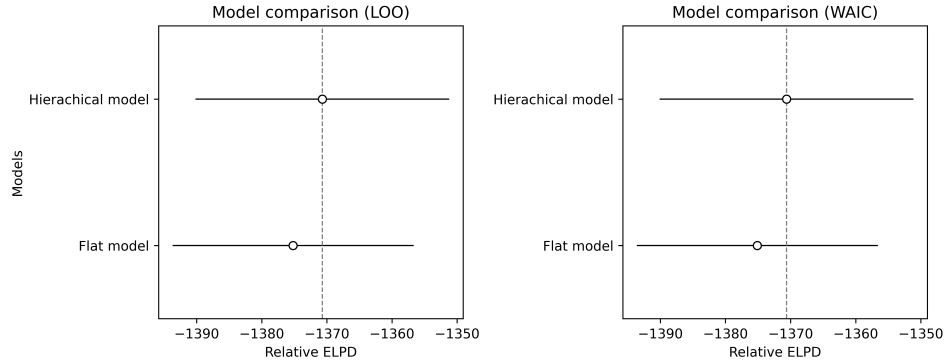

Figure 8: Model comparison for the statistical analysis in Question 3 (Do autograders differ systematically from human experts?). Left panel: Leave-One-Out cross-validation (LOO) scores. Right panel: Widely Applicable Information Criterion (WAIC). Both metrics approximate the Expected Log Predictive Density (ELPD), a measure of predictive accuracy (higher values indicate better performance). The models perform similarly, which is expected given that the data is simulated without an explicitly hierarchical structure. Here we choose the hierarchical model (Equation 3) to demonstrate how to interpret its parameters. In practice, when models perform similarly, researchers should favour the simpler model unless theoretical or interpretability considerations justify the added complexity.

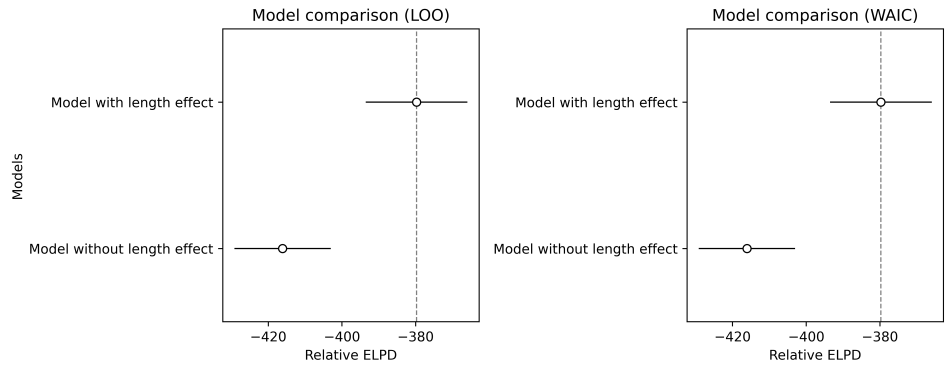

Figure 9: Model comparison for the statistical analysis in Question 5 (Do autograders favour longer outputs?). Left panel: Leave-One-Out cross-validation (LOO) scores. Right panel: Widely Applicable Information Criterion (WAIC). Both metrics approximate the Expected Log Predictive Density (ELPD), a measure of predictive accuracy (higher values indicate better performance). Comparing these models tests whether including a length-effect term (Equation 5) significantly improves predictive performance. The model including the length effect clearly outperforms the simpler model without it, justifying a closer investigation into grader-specific length biases.

Table 2: Example of a rubric score that a researcher might create to grade open-ended question.

POINTS	DESCRIPTION
1	Completely off-topic or no relevant content.
2	Minimal response with no clear concepts or severe confusion.
3	Mentions a relevant idea but largely undeveloped or inaccurate.
4	States one relevant concept with limited clarity or major misconceptions.
5	Mentions key concepts but lacks depth and contains notable flaws.
6	Covers some key concepts with partial accuracy and development.
7	Addresses key concepts clearly, with minor omissions or inaccuracies.
8	Explains most important concepts with accuracy and reasonable depth.
9	Thorough and accurate response with clear development and insight.
10	Complete responses with deep understanding and insightful connections.

A.4 Supplementary question: Is the grading scale well calibrated?

In this section we explore the grading process, and why we chose an ordered logistic regression model. To ensure consistent evaluation, Florence needs to establish standardised scoring across all graders. One simple approach would be to instruct everyone (humans and autograders) to count the number of relevant keywords from a predefined list. Besides obvious issues (e.g., synonyms need to be accounted for), this approach is not scalable as it require building explicit keyword lists for each open-ended question. Florence, as often done in practice, will instead develop a grading rubric that can be applied to each question. She might come up with something like in Table 2.

Having an ordinal scale, instead of just a description, is of course very useful. But this scale is not as simple to interpret as an interval scale (e.g., temperature). If Florence wants to make claims like "autograder A increases the score by 1 point," she needs to be aware of whether the intervals between categories are equivalent. If they are not (which is often the case with ordinal scales), a 1-point increase will mean something different depending on where it occurs on the scale. For example, moving from 5 to 6 might represent crossing some fundamental threshold, while moving from 9 to 10 might just be a small qualitative improvement between two already good responses.

In an ordered logistic regression model, the scores are considered individual categories with a meaningful order. The model maps the observed scores onto a continuous latent scale and, if necessary, can estimate the category boundary values (called cutpoints) on this latent scale. By examining the

Table 3: Cutpoints and interpretations for score intervals

CUTPOINT	VALUE	RANGE	SCORE	THRESHOLD
c1	-4.07	–	0–1	Starting point
c2	-3.25	0.82	1–2	Narrow
c3	-2.39	0.86	2–3	Narrow
c4	-1.28	1.11	3–4	Moderate
c5	-0.20	1.08	4–5	Moderate
c6	1.29	1.49	5–6	Wide
c7	3.11	1.82	6–7	Wide
c8	4.71	1.60	7–8	Wide
c9	5.60	0.89	8–9	Narrow
c10	6.22	0.62	9–10	Narrow

spacing between these cutpoints, we can determine whether the intervals in our grading scale are equivalent across the range. If they are not, we can make inferences from the distances between cutpoints. For example, if some cutpoints are widely spread, it could indicate that we are capturing an important threshold (i.e., substantial changes in the underlying latent score is required to move between categories), that there is a gap in the measurement scale or that graders are reluctant to use certain portions of the scale. Conversely, if some cutpoints are close together, it could indicate that the scale is very sensitive in that region (i.e., small changes in the latent score result in different observed scores), that models have similar latent abilities in that "region" or that the scale contains redundant categories.

Building great scales is by no means an easy task. It is a well-established challenge that has been thoroughly studied in the field of psychometrics, but is beyond the scope of the presented paper. However, using a GLM with an ordered logistic regression is a useful tool to examine the properties of a given scale and can help identify areas that require caution during interpretation.

Going back to Florence, to better capture differences in model capabilities, she decides to examine her grading scale. To do this, she can look at the learned cutpoint parameters of the ordered logistic regression. Taking Equation 1 for example, we can write the cumulative probability more explicitly as:

$$\begin{aligned} \phi_i &= \beta_0 + \beta_1 \cdot X_i^{\text{grader}} + \beta_2 \cdot X_i^{\text{LLM}} \\ p_{ij} &= \text{logit}^{-1}(c_j - \phi_i) \end{aligned} \tag{6}$$

where ϕ_i is the linear predictor, representing the location of response i on an unobserved latent scale. This latent scale is assumed to underlie the observed ordinal scores (e.g., 110). The cutpoints c_j divide the latent scale into discrete intervals corresponding to the observed score categories. The probability p_{ij} represents the cumulative probability that the score assigned to response i is less than or equal to category j, and is calculated as the inverse logit of the difference between the cutpoint c_j and the linear predictor ϕ_i.

Conceptually, this means that the probability of scoring at (or below) a certain grade j (on the observed scale) is calculated by comparing the linear predictor to the cutpoint (on the latent scale). These cutpoints are the mechanism through which ordered logistic models maintain the ordinal structure of the data. They were therefore naturally present in the previously discussed models, but were not mentioned as the focus was on the predictor variables.

Once Florence fits this model, she can analyse the inferred cutpoint values. Those can be found in Table 3. She observes that the distance c1-c2, and c2-c3, is below 1 unit, whereas the distance c6-c7 and c7-c8 is above 1.5 units. This large jump suggests that her scale lacks sensitivity around c6-c8 (i.e., many performance scores are clustering there). From this, she could decide that she wants to better capture difference in that region and therefore adds some intermediate categories in this area.