QUANTIFYING BIASES IN LLM-AS-JUDGE EVALS

Anonymous authors
Paper under double-blind review

ABSTRACT

The evaluation of large language models (LLMs) is increasingly performed by other LLMs, a setup commonly known as "LLM-as-a-judge", or autograders. While autograders offer a scalable alternative to human evaluation, they are not free from biases (e.g., favouring longer outputs or generations from their own model family). Here we propose a statistical framework based on Bayesian generalised linear models (GLMs) that enables researchers to address their primary research questions (e.g., LLM capability or risk assessment), while simultaneously identifying, quantifying and mitigating various biases in their autograders. Our approach can be applied to various evaluation formats (e.g., absolute scores or pairwise preferences) and augments traditional metrics (e.g., inter-rater agreement) by providing precise uncertainty estimates and clarifying sources of disagreement between graders. This framework also enables efficient counterfactual simulations without costly re-evaluation (e.g., assessing agreement after removing systematic biases). We demonstrate these capabilities through simulated examples, with all methods available in an open-source software package. Overall, we introduce a novel framework for autograder evaluation which allows researchers to detect, quantify and correct for various biases in a systematic way.

1 INTRODUCTION

Imagine a typical scenario: a researcher, Florence[1], is assessing how well an LLM does on a given task. Outside of true/false questions, the outputs can be quite complex, e.g., open-ended answers, agentic trajectories or intrinsic preferences. Techniques have been developed to assess these outputs, e.g., scoring rubrics or collecting preferences. Due to the high stochasticity of LLMs, typical LLM evaluations require collecting a lot of samples meaning that manually scoring each response would be very time-consuming. Florence, like many researchers, decides to build autograders to automate this task. As she is interested in grading open-ended questions, she creates a rubric and prompts an autograder to apply it. Being a careful researcher, she wants to assess how well the autograders scores align with her own. As commonly done, she decides to assess this using an inter-rater agreement, e.g., Krippendorff's α (Tam et al., 2024; Bavaresco et al., 2025). She might get a value close lower than zero, indicating substantial disagreement between her and the autograders. But what does this mean? Is this just random noise or is there a way to explain this disagreement?

Recent studies suggest that such disagreement may not just be noise, as autograders can exhibit systematic biases. For instance self-bias, where LLM-based graders assign higher scores to responses generated by the same LLM family (Panickssery et al., 2024; Liu et al., 2024b), or more broadly to machine-generated content over human-written responses (Liu et al., 2023). Another common issue is length bias, where longer answers are preferred regardless of their actual quality (Zheng et al., 2023; Dubois et al., 2024). Additional biases include preferences for certain writing styles, answer structures, or the presence of certain keywords (Koo et al., 2024; Wang et al., 2024; Stureborg et al., 2024; Wu & Aji, 2025).

Through careful observation of the outputs, Florence might identify that the autograders consistently assigns lower scores than she does. She suspects that there is actually no fundamental disagreement on what constitutes good or bad responses, but rather slightly different scoring thresholds. One way to test this would be to adapt the scoring rubric to encourage higher scores and rerun the evaluation,

[1] In tribute to the pioneering work of two Florence Nightingales in statistics: the 19th-century nurse who applied statistical methods to public health and the 20th-century statistician

but this approach would be resource intensive. A more efficient alternative is to simulate a counterfactual: what would the scores look like if we removed the systematic shift? By adjusting for this bias in the existing data and recomputing Krippendorff α on the simulated scores, Florence can test her hypothesis without collecting new data. Doing this, she might find a higher value which would confirm that the apparent disagreement was largely due to a systematic shift rather than fundamental differences in quality assessment.

By performing such analyses, researchers can transform vague notions of autograder unreliability into precise and actionable insights about specific biases. To achieve this at scale, researchers need a framework that can decompose disagreement into interpretable components, quantify each bias with uncertainty, and predict how removing specific biases would affect evaluation outcomes without requiring costly re-evaluation. Our Bayesian GLM framework provides exactly these capabilities: (1) jointly modeling multiple bias sources (self-bias, length bias, grader severity, item effects) to identify which biases are present and their relative importance, (2) providing posterior distributions that quantify not just whether biases exist but their precise magnitude and uncertainty, (3) supporting both absolute scoring and pairwise preference formats to handle diverse evaluation setups, and (4) enabling counterfactual simulations that reveal how evaluation outcomes would change if specific biases were removed.

In the following sections, we will demonstrate this framework by addressing five common evaluation challenges. To facilitate wider adoption, all statistical models presented in this paper are implemented in the open-source HiBayes package.

2 METHODS

We begin by explaining how autograder scores can be compared to human scores and how this comparison can be integrated into an LLM evaluation analysis (Question 1). We then show how to assess whether the graders are biased towards certain models being evaluated (Question 2), how to quantify individual differences in the case of multiple graders (Question 3), and how to analyse item-level patterns (Question 4). Finally, we show how this framework can be applied to pairwise judgments settings, how to quantify intransitive (e.g., cyclic) preferences and how to assess whether graders have biases towards longer formats (Question 5). For a summary of the evaluation questions along with their corresponding formalisations, cf 1.

2.1 HOW DO SCORES FROM AN AUTOGRADER COMPARE TO SCORES FROM AN EXPERT?

Florence needs to assess how well two LLMs do at answering open-ended questions. Because she is automating their grading with autograders, she essentially needs to answer two questions: 1) Is LLM A or LLM B better at answering open-ended questions? 2) Can my autograder reliably evaluate LLM responses relative to a human annotator?

A GLM framework allows to answer these questions with a single analysis. Suppose that each LLM answered $N = 50$ items which were each graded by Florence and an autograder on a scale from 1-10 (simulated scores depicted on Figure 1).

GLMs extend linear regressions to handle non-normal outcomes while preserving the familiar regression structure. This is particularly useful as it allows researchers to include multiple predictors, control for potential confounders, and isolate the contribution of each variable to an outcome.

To answer the questions above, Florence can fit a regression model with an intercept β_0, which represents the overall average latent score, a coefficient β_1 which quantifies the effect of the grader, and coefficient β_2 which quantifies the effect of LLM A versus LLM B on the score (cf., linear predictor ϕ_i in Equation 1).

$$\begin{aligned} \phi_i &= \beta_0 + \beta_1 \cdot x_i^{\text{grader}} + \beta_2 \cdot x_i^{\text{LLM}} \\ \text{score}_i &\sim \text{OrderedLogistic}(\phi_i, \boldsymbol{c}) \end{aligned} \tag{1}$$

In this model, the linear predictor ϕ_i combines the effects of both the grader type and LLM identity. Because scores take discrete values from 1-10, we use an ordered logistic likelihood function. The

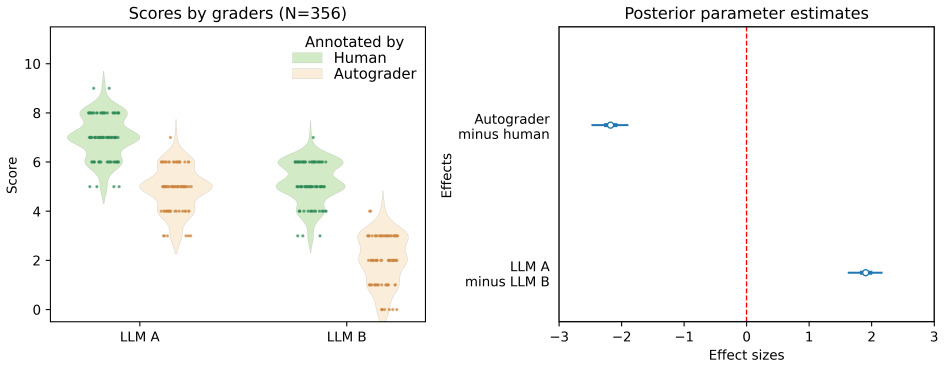

Figure 1: Addressing Question 1: how do scores from an autograder compare to scores from an expert?. Left panel: Simulated scores for LLM-generated answers graded by a human expert (Florence) and an autograder. Right panel: Posterior distributions of estimated effects. The horizontal blue lines represent 95% credible intervals. The dashed red vertical line indicates a null effect ($\beta = 0$). The coefficient for autograder minus human is negative, with a credible interval that does not include zero, indicating that the autograder tends to assign lower scores. The coefficient for LLM A minus LLM B is similarly positive, suggesting that LLM A receives higher scores than LLM B on average.

linear predictor produces a continuous latent value, which the ordered logistic model maps to discrete 1-10 scores through estimated cutpoints $\boldsymbol{c}$. These cutpoints are estimated during model fitting along with the β coefficients (cf. Appendix A.3). The variables x_i^{grader} and x_i^{LLM} encode the identity of the grader and the LLM respectively. Each variable takes a value of $+1$ or -1 (i.e., effect coding) to distinguish between the two levels (e.g., autograder vs. human, LLM A vs. LLM B).

After fitting, we can make two inferences based on the effect sizes of the coefficients β_1 and β_2 (right panel of Figure 1):

1. The "Autograder minus human" effect is negative with credible intervals excluding zero, indicating that the autograder gives lower scores than the human expert

2. The "LLM A minus LLM B" effect is positive with credible intervals excluding zero, indicating that LLM A receives higher scores on average than LLM B.

Crucially, because the GLM accounts for both sources of variation simultaneously, Florence can confidently select LLM A for her task while remaining aware of the autograder's conservative scoring tendency. This illustrates how a GLM framework enables researchers to both answer substantive research questions and validate their evaluation methods within a single, principled analysis.

2.2 Do autograders favour their own generation?

Recent literature has raised concerns that autograders may demonstrate self-bias, a tendency to assign better scores to outputs generated by the same base model (Panickssery et al., 2024; Liu et al., 2024b; Koo et al., 2024) or outputs from models vs. humans (Liu et al., 2023). Similarly to above, a GLM framework allows to quantify autograder self-bias while evaluating LLMs.

Florence is concerned that her autograder (from model family A) might unfairly favour outputs from LLM A (also from model family A). To assess such self-bias, she uses a second autograder (from model family B). She wants to investigate whether responses from LLM A receive higher scores when graded by the autograder A compared to when graded by autograder B (and vice versa). The resulting data are shown in the left panel of Figure 2.

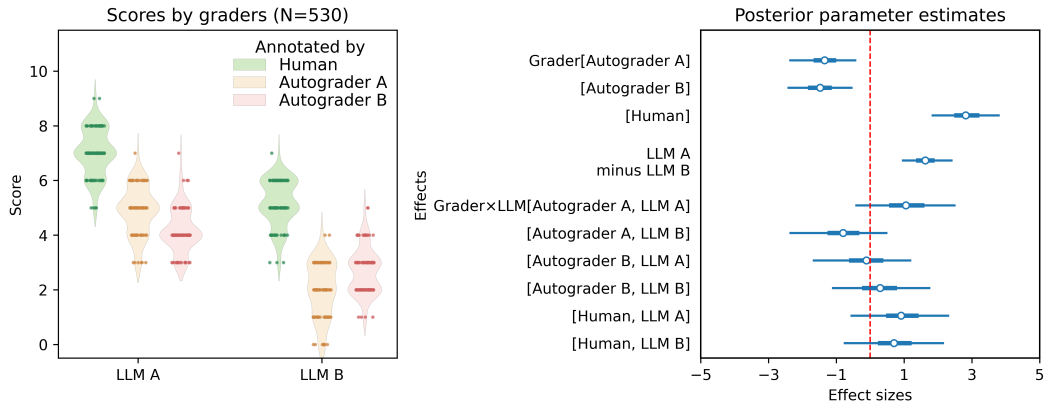

Figure 2: Addressing Question 2: Do autograders favour their own generation?. Left panel: Simulated scores for LLM-generated answers by LLM A and LLM B. Scores were given by a human expert (green) and autograders (yellow and red). Right panel: Posterior distributions of estimated effects from the GLM. The horizontal blue lines represent 95% credible intervals, and the dashed red vertical line indicates a null effect ($\beta = 0$). The grader effect β_1 shows how each grader deviates from the average score across all graders and LLMs. The LLM effect β_2 is positive, indicating that LLM A generally receives higher scores than LLM B. The graderLLM terms β_3 represent a set of parameters (one for each graderLLM combination). Autograder A seems to have a tendency to prefer LLM A vs LLM B, suggesting a potential self-bias.

We extend the previous model (Equation 1) by adding a term that captures whether specific graders systematically favour outputs from specific LLMs. This is implemented as a set of graderLLM interaction effects[2], denoted by β_3.

$$\begin{aligned} \phi_i &= \beta_0 + \beta_{1,g_i} + \beta_2 \cdot x_i^{\text{LLM}} + \beta_{3,g_i,\ell_i} \\ \text{score}_i &\sim \text{OrderedLogistic}(\phi_i, \boldsymbol{c}) \end{aligned} \tag{2}$$

As we now have more than two graders, the main effect coefficients $\beta_{1,1}, \beta_{1,2}, \beta_{1,3}$ represent each grader's deviation from the grand mean score, estimated using effect coding. The LLM variable still has two levels and is binary-coded as before (see Equation 1). The interaction term $\beta_{3,j,k}$ represents a set of parameters estimated using index-based coding, with one distinct coefficient independently estimated for each grader j and LLM k combination.

After selecting the best-fitting model using model comparison techniques (Figure 7) in Section A.3), we examine the estimated effects (right panel of Figure 2. The interaction parameters β_3 show that Autograder A assigns somewhat higher scores to LLM A than to LLM B (positive vs. negative effects), suggesting potential self-bias toward outputs from its own model family (although this is not definitive as the credible intervals overlap with zero).

With these findings, Florence confidently answers her main question (LLM A performs better on open-ended questions), uncovers that the autograders assign lower scores than she does, and identifies a potential self-bias from Autograder A.

2.3 Do autograders differ systematically from human experts?

Florence might ask a few colleagues to help grade some of the responses (Human X, Y and Z in the left panel of Figure 3.), and try different an additional autograder (Autograder A, B and C in the left panel of Figure 3.).

[2]Strictly speaking, this is not a single interaction effect (e.g., $\beta_3 \cdot X^{\text{grader}} \cdot X^{\text{LLM}}$), but a set of parameters estimated independently using index-based coding (where each graderLLM combination has a unique integer index). This allows direct comparisons across specific combinations rather than relying on a single coefficient.

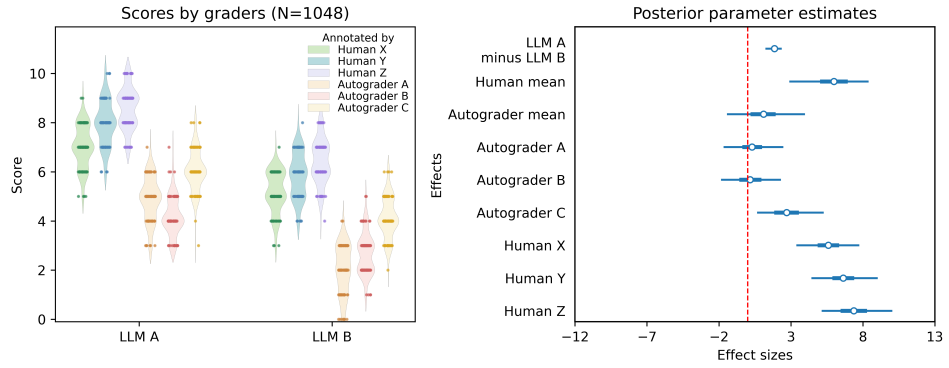

Figure 3: Addressing Question 3: Do autograders differ systematically from human experts?. Left panel: Simulated scores on LLM A and LLM B answers, as graded by multiple human experts (green) and autograders (yellow and red). Right panel: Posterior distributions of estimated effects from the hierarchical model. The horizontal blue lines represent 95% credible intervals, and the dashed red vertical line indicates a null effect ($\beta = 0$). Individual grader effects show how each grader deviates from their respective group-level average (human or autograder). Group-level means for human and autograder graders ($\mu_{\text{graderType}}$) indicate that, on average, human graders assign higher scores than autograders.

She might then be interested in assessing: 1) whether autograder scores, on average, differ systematically from human scores, and 2) how much individual graders vary within each group.

To capture both group-level and individual-level differences, we can define what is known as a hierarchical GLM (cf. Equation 3). In this model, each grader has their own scoring tendency (β_{grader_i}), drawn from a group-level distribution (human or autograder). This allows to estimate group-level means for humans versus autograders while also capturing how individual graders deviate from their group's average. Through partial pooling (sharing information across graders of the same type), the model makes efficient use of limited data, which is particularly helpful when some graders have few observations.

$$
\begin{aligned}
\phi_i &= \beta_0 + \beta_{1,g_i} + \beta_2 \cdot x_i^{\text{LLM}} \\
\text{score}_i &\sim \text{OrderedLogistic}(\phi_i, \boldsymbol{c}) \\
\beta_{1,g_i} &\sim \mathcal{N}(\mu_{\text{graderType}_i}, \sigma^2_{\text{graderType}_i}) \\
\mu_{\text{graderType}_i} &\sim \mathcal{N}(0, 3) \\
\sigma^2_{\text{graderType}_i} &\sim \text{HalfCauchy}(1)
\end{aligned}
\tag{3}
$$

As before, β_2 is a scalar coefficient applied to $x_i^{\text{LLM}} \in \{-1, +1\}$, which indicates whether the response was generated by LLM A or LLM B. Unlike before, β_{1,g_i} represents the effect of the individual grader who assigned score i, and is drawn from a group-level distribution based on grader type (human or autograder). Specifically, $\beta_{1,g_i} \sim \mathcal{N}(\mu_{\text{graderType}_i}, \sigma^2_{\text{graderType}_i})$, where $\mu_{\text{graderType}_i}$ represents the average score tendency for each grader type, and $\sigma^2_{\text{graderType}_i}$ captures variability within each type. The prior distributions for the group-level means and variances ($\mu_{\text{graderType}}$ and $\sigma^2_{\text{graderType}}$) are specified in Section A.2. This hierarchical structure enables the model to estimate both the average difference between human and autograder scores and the variation among individual graders.

To formally assess whether this group-level difference exists, Florence should of course compare this hierarchical approach against a simpler flat model (cf. Figure 8 in Section A.3 for a model comparison). Here we select the hierarchical model to demonstrate how to examine both group-level differences between grader types and individual grader characteristics.

In the right panel of Figure 3 we see the individual grader effects (β_1; Autograder AC and Human XZ in the plot) and the group-level means ($\mu_{\text{graderType}}$; human mean and autograder mean in the plot). Using this method, Florence can confidently conclude that there is a general tendency for humans to

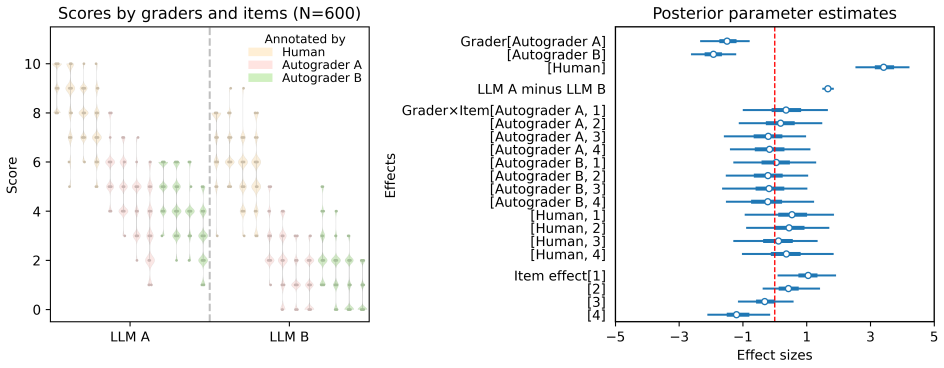

Figure 4: Addressing Question 4: How do scores differ at an item level?. Left panel: Simulated scores for each item (14), grouped by LLM and grader identity. Each cell shows the distribution of scores assigned by a given grader to responses from a particular model on a given item. Right panel: Posterior distributions of estimated effects from the item-level GLM (Equation 4). The plot shows main effects for grader and LLM identity (top), item main effects (bottom), and graderitem interactions (middle). Horizontal blue lines represent 95% credible intervals, and the dashed red vertical line indicates a null effect ($\beta = 0$). Item 1 has a strong positive effect, suggesting it consistently receives higher scores. In contrast, Item 4 has a negative effect, indicating that it receives lower scores. Grader - item interaction terms are small and uncertain, indicating no evidence of systematic grader disagreement on specific items.

give higher scores than autograders. Additionally, she can visualise individual-level differences and make informed decisions. For example, she might observe that Autograder C produces scores that are more closely aligned with those of the human graders. If consistency with human judgment is a key objective, she may choose to use this autograder in future evaluations.

2.4 How do scores differ at an item level?

Florence now becomes interested in whether variation arises at the level of individual evaluation items (i.e., open-ended questions). She wonders whether some items consistently receive higher or lower scores, and whether graders agree more on certain items than others.

To answer these questions, she needs repeated responses for the same items. Until now, we have assumed that each data point corresponds to a different item. Lets instead imagine that Florences dataset consists of four items, with each model answering each item 25 times. The data split by items can be seen in the left panel of Figure 4 (different items are represented by violin plots of the same colour).

To answer Question 4, we extend Equation 1 by including two additional terms. The first term, β_{3,m_i}, accounts for a main effect of items, capturing whether some items receive systematically higher or lower scores. The second term, β_{4,g_i,m_i}, represents a graderitem interaction, allowing us to test whether particular graders behave differently on specific items.

$$\begin{aligned} \phi_i &= \beta_0 + \beta_{1,g_i} + \beta_2 \cdot x_i^{\text{LLM}} + \beta_{3,m_i} + \beta_{4,g_i,m_i} \\ \text{score}_i &\sim \text{OrderedLogistic}(\phi_i, \boldsymbol{c}) \end{aligned} \tag{4}$$

The term β_{1,g_i} captures the main effect of grader g_i, and β_2 models the effect of LLM identity (e.g., whether the response was produced by LLM A or B). As mentioned above, the new term β_{3,m_i} represents the main effect of item m_i, which captures whether some questions tend to receive higher or lower scores overall. The final term, β_{4,g_i,m_i}, captures grader - item interactions and is implemented in the same way as the interaction term in Equation 2, i.e., a coefficient for combination. This allows to directly compare individual combinations and detect whether certain graders are more lenient or harsh on specific items. As before, all main categorical effects (grader, item) are encoded using effect coding, so that the resulting coefficients reflect deviations from the overall mean.

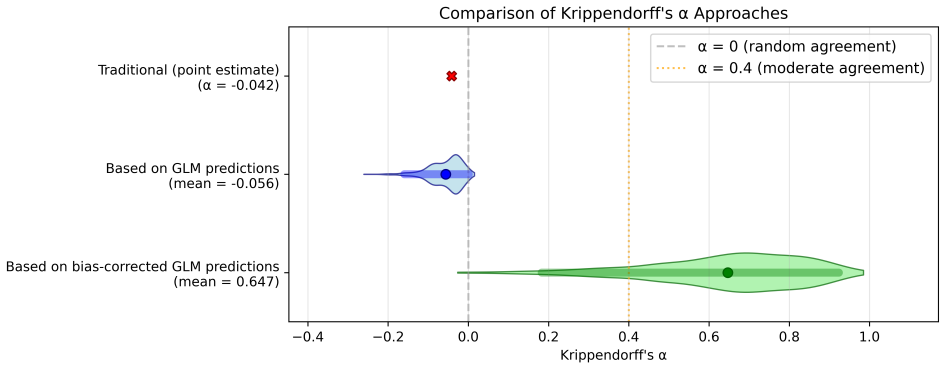

Figure 5: Posterior distributions of Krippendorffs α under different modeling assumptions. The red cross shows the traditional α computed directly from the observed scores, suggesting strong disagreement between graders. The blue distribution shows α values estimated from posterior simulations of a fitted GLM, incorporating uncertainty in the predictions. The green distribution shows α values after removing the main effect of grader identity, revealing what agreement might look like in a counterfactual scenario where graders do not differ systematically in scoring scale.

After fitting the model, Florence inspects the estimated effects (right panel of Figure 4).

1. Looking at the main effect of item (β_3 in Equation 4), she observes that item 1 leads to higher scores and item 4 to lower scores. This suggests that the former is easier to answer, and the later more challenging.

2. Looking at the grader - item interaction term (β_4 in Equation 4), she does not see evidence that specific graders differ on individual items.

From these results Florence concludes that while some items appear easier than others, grader disagreement is not concentrated on any particular question. The grader main effects reveals that humans consistently score higher than autograders, but Florence wonders about the consistency of their judgments beyond this systematic bias, i.e., do graders at least agree on which responses are relatively better or worse? Is the disagreement between graders due to fundamental differences in quality judgment, or merely due to correctable systematic biases?

This is where combining GLMs and inter-rater agreement metrics like Krippendorff's α becomes particularly valuable. While GLMs quantify systematic biases and α measures overall agreement, α alone cannot distinguish between random disagreement and systematic biases. For example, if she were to compute α directly she would get $\alpha = -0.2$ (red cross in Figure 5), indicating substantial disagreement, but telling her little about why graders disagree or whether the disagreement is fixable.

The GLM framework offers two advantages here. First, by computing α on posterior samples from the fitted GLM, Florence can obtain not just a point estimate but a full distribution (blue in Figure 5). Second, and more importantly, GLMs enable counterfactual simulations. Florence can ask: "What would agreement look like if graders didn't have systematic biases?" To answer this, she removes each grader's estimated bias (the β_1 coefficients) from the linear predictor before mapping to categorical scores, then recomputes α on these bias-adjusted predictions.

This counterfactual α (green distribution in Figure 5) jumps to approximately 0.7, which is substantially higher than the observed α. This reveals that most disagreement stems from systematic scoring differences rather than inconsistent judgments about quality. The wider green distribution reflects increased uncertainty after removing predictable grader variation.

Without the GLM framework, Florence would only know the agreement is poor. With it, she can decompose the disagreement into systematic biases (which can be corrected through calibration) versus fundamental disagreements about quality. The ability to simulate counterfactuals without collecting new data makes GLMs particularly valuable for understanding and improving evaluations.

2.5 DO AUTOGRADERS FAVOUR LONGER OUTPUTS?

So far, we have focused on evaluation settings where graders assign absolute scores. However, many LLM evaluations rely on pairwise comparisons, where graders are asked to choose which of two outputs better satisfies a target criterion (e.g., correctness). The same statistical modeling framework can be applied in this setting, with the outcome modeled as a binary preference. We use such a setup to illustrate how pairwise comparisons can be modeled and to examine length bias, which has often been observed in pairwise evaluation settings. Of course, length bias can also be captured in absolute score setups similarly to other biases in previous sections.

Let's imagine that Florence wants to compare the quality of outputs generated by three different LLMs. She chooses a pairwise evaluation format, where each grader (e.g., herself or an autograder) is repeatedly shown two responses to the same prompt - each generated by a different LLM - and must select the better response. An example of such data can be seen in the left panel of Figure 6. Each bar represents a pairwise comparison (e.g., "LLM A vs. LLM B"), and its height reflects how frequently the first listed model (e.g., LLM A) was chosen. To model this binary outcome, we switch from the ordered logistic regression used previously to a binomial GLM with a logit link function. The outcome variable y_i indicates whether the first model in the pair was chosen ($y_i = 1$) or not ($y_i = 0$), and we include a categorical effect in the model to denote the LLM pair being compared.

Florence's younger brother, always up-to-date with ML controversies, recently told her that some autograders may systematically prefer longer outputs even if those outputs are not of higher quality, a phenomenon commonly referred to as length bias (Zheng et al., 2023; Dubois et al., 2024). To capture this bias, she adds a continuous predictor capturing the token-length difference between the two outputs. As there are two graders (herself and the autograder), which might have different biases, she computes one such predictor per grader. To test the existence of the length bias formally, she compares the model with and without this term (cf. Figure 9 in Section A.3). For demonstration purposes, let's look at the model with grader-specific length bias here:

$$
\begin{aligned}
\text{logit}(p_i) &= \beta_0 + \beta_{1,\pi_i} + \beta_{2,g_i} + \beta_{3,g_i} \cdot x_i^{\text{lengthDiff}} \\
y_i &\sim \text{Binomial}(1, p_i) \\
\beta_{3,g_i} &\sim \mathcal{N}(\mu_{\text{lengthDiff}}, \sigma^2_{\text{lengthDiff}}) \\
\mu_{\text{lengthDiff}} &\sim \mathcal{N}(0, 0.5) \\
\sigma_{\text{lengthDiff}} &\sim \text{HalfNormal}(1.0)
\end{aligned}
\tag{5}
$$

where y_i is a binary outcome indicating whether the first-listed LLM was preferred. The intercept β_0 is the overall tendency to prefer the first-listed model, β_{1,π_i} captures pair-specific preferences (e.g., "LLM A vs. LLM B") where π_i indexes the LLM pair, and β_{2,g_i} captures grader g_i's overall tendency to prefer the first-listed model. The predictor $x_i^{\text{lengthDiff}}$ is the token-length difference between the two responses. The grader-specific slope coefficient β_{3,g_i} quantifies how sensitive grader g_i is to length differences and is drawn from a hierarchical distribution with mean $\mu_{\text{lengthDiff}}$ and standard deviation $\sigma_{\text{lengthDiff}}$. Positive values of $\mu_{\text{lengthDiff}}$ indicate a preference for longer outputs. The hierarchical structure captures both the average length bias across graders and the variability among individual graders.

Importantly, once we have estimated the probability of choosing an LLM over another, we can compare these probabilities across pairs. This allows to identify rational (transitive) and irrational (intransitive) patterns of decision-making, such as cyclic dependencies (e.g., preferring A over B, B over C, but C over A). Such intransitivities exist in LLM evaluations (Xu et al., 2025). Traditional models like the BradleyTerry model implicitly assume transitivity and thus cannot capture these cycles. Recent approaches have proposed either removing intransitivities from datasets (Yu et al., 2025) or explicitly quantifying them (Zhang et al., 2025; Liu et al., 2024a; Zhao et al., 2024). GLMs naturally capture these intransitivities alongside grader biases and differences in LLM performance.

Inspecting the estimated effects in the right panel of Figure 6, we observe a positive effect for the grader-specific length bias parameter (β_3), particularly for the autograder. From this, Florence can infer that the autograder is more likely to select longer outputs, irrespective of their intrinsic quality, meaning that the autograder implicitly associated output length and perceived correctness. By explicitly quantifying such biases within the model, Florence can more reliably interpret differences

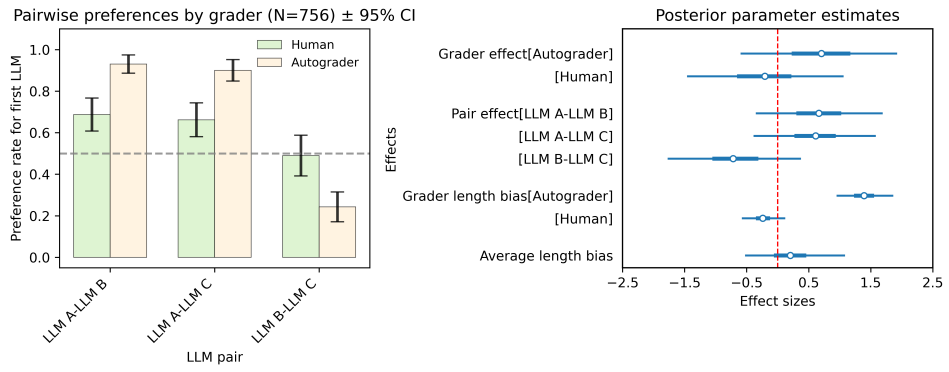

Figure 6: Addressing Question 5: Do autograders favour longer outputs?. Left panel: Proportion of pairwise preferences across three LLM pairs (A vs. B, A vs. C, B vs. C). Preference rate values represent the fraction of cases where the first listed model was selected over the second. Error bars represent 95% confidence intervals. Right panel: Posterior distributions of estimated effects from the GLM for pairwise comparisons (log-odds scale). Horizontal blue lines represent the 95% credible intervals, and the dashed red vertical line indicates a null effect (β=0). The pair effect terms represent the relative preference between specific pairs of LLMs, indicating which LLM is generally preferred. The grader length bias terms quantify each grader's sensitivity to token-length differences when making choices. A positive length bias indicates a preference towards longer outputs.

in LLM rankings. For example, if LLM A wins most comparisons but consistently produces longer outputs, Florence might question: "Is LLM A genuinely better, or simply more verbose? Can I really trust the autograders judgements?". Additionally, by examining the estimated LLM pair parameters (β_1), she can verify whether the observed preferences follow a consistent ranking or if there are intransitive (cyclical) patterns. Here, the estimated parameters indicate a consistent ordering: LLM A tends to be preferred to LLM B and LLM C, and LLM B tends to be preferred to LLM C. This integrated statistical framework empowers her to disentangle and quantify these systematic biases and assess preference consistency, leading to deeper and more reliable conclusions.

CONCLUSION

In this paper, we introduced a statistical framework for evaluating autograders using Bayesian GLMs. By jointly modelling the evaluation outcome and the scoring process, this approach enables researchers to assess both LLM performance and autograder behaviour within a single analysis. Through a series of examples, we followed the journey of a fictional researcher toward evaluating autograders. We used simulated data throughout to explore a wide range of evaluation questions and settings, and to illustrate key modelling principles in a controlled and reproducible way.

Specifically, we showed how this framework can quantify various types of biases (e.g., self-bias and length bias), capture individual-level differences both among graders and items, and improve the estimation of group-level trends through hierarchical modelling. Crucially, the framework enhances traditional inter-rater agreement metrics in two ways: it provides uncertainty quantification through posterior distributions, and it enables counterfactual analysis to understand the sources of disagreement. Researchers can still compute familiar statistics such as Krippendorff's α, Cohen's κ, or Kendall's τ, but now with credible intervals and the ability to decompose disagreement into systematic biases versus fundamental differences in judgment. Additionally, we demonstrated the method's flexibility across different evaluation formats, including absolute scoring and pairwise comparisons.

The examples presented are by no means exhaustive. Many other applications and extensions of GLMs are possible, and we hope this work provides a clear and practical starting point for researchers seeking to adapt the framework to their own evaluation scenarios. To support practical adoption, we have summarised common evaluation questions and their implementation in a GLM setup in Table 1. All statistical models presented in this paper are implemented in the open-source HiBayes package.

REFERENCES

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernndez, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andr F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2025. URL https://arxiv.org/abs/2406.18403.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled AlpacaEval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 517–545, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.29. URL https://aclanthology.org/2024.findings-acl.29/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using Gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Yinhong Liu, Zhijiang Guo, Tianya Liang, Ehsan Shareghi, Ivan Vulić, and Nigel Collier. Aligning with logic: Measuring, evaluating and improving logical consistency in large language models. *arXiv preprint arXiv:2410.02205*, 2024a.

Yiqi Liu, Nafise Moosavi, and Chenghua Lin. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12688–12701, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.753. URL https://aclanthology.org/2024.findings-acl.753/.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7f1f0218e45f5414c79c0679633e47bc-Paper-Conference.pdf.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators, 2024. URL https://arxiv.org/abs/2405.01724.

Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V Stolyar, Katelyn Polanska, Karleigh R McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ digital medicine*, 7(1):258, 2024.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 297–312, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.21/.

Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. Investigating non-transitivity in llm-as-a-judge. *arXiv preprint arXiv:2502.14074*, 2025.

Yan Yu, Yilun Liu, Minggui He, Shimin Tao, Weibin Meng, Xinhua Yang, Li Zhang, Hongxia Ma, Chang Su, Hao Yang, et al. Elspr: Evaluator llm training data self-purification on non-transitive preferences via tournament graph reconstruction. *arXiv preprint arXiv:2505.17691*, 2025.

Yifan Zhang, Ge Zhang, Yue Wu, Kangping Xu, and Quanquan Gu. Beyond bradley-terry models: A general preference model for language model alignment. In *Forty-second International Conference on Machine Learning*, 2025.

Xiutian Zhao, Ke Wang, and Wei Peng. Measuring the inconsistency of large language models in preferential ranking. *arXiv preprint arXiv:2410.08851*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.