Generalised Probabilistic Modelling and Uncertainty Estimation in Comparative LLM-as-a-judge

Anonymous ACL submission

Abstract

This paper explores generalised probabilistic modelling and uncertainty estimation in comparative LLM-as-a-judge frameworks. We show that existing Product-of-Experts methods are specific cases of a broader framework, allowing for diverse modelling options. Furthermore, we propose improved uncertainty estimates for individual comparisons, enabling more efficient selection and achieving strong performance with fewer evaluations. We also introduce a method for estimating overall ranking uncertainty. Finally, we demonstrate that combining absolute and comparative scoring enhances performance. Experiments show that the specific expert model has a limited impact on final rankings but our proposed uncertainty estimates, especially the probability of reordering, significantly improve the efficiency of systems. Furthermore, ranking-level uncertainty metrics can be used to identify low-performing predictions, where the nature of the probabilistic model has a notable impact on the quality of the overall uncertainty.

1 Introduction

Instruction-tuned Large Language Models (LLMs) have shown impressive zero-shot performance on a wide range of natural language processing and generation tasks (Wei et al., 2021; Ouyang et al., 2022; Bai et al., 2022; Zhou et al., 2023; Chung et al., 2024). While the number of downstream applications of aligned LLMs increases (Brown et al., 2020; Touvron et al., 2023; Achiam et al., 2023; Dubey et al., 2024), so does the need to evaluate their performance on bespoke tasks, which could lack labelled data or are costly for humans to judge at scale (Zheng et al., 2023b; Wang et al., 2022; Taori et al., 2023). As an alternative, instruction-tuned LLMs have increasingly been used as a replacement for humans to evaluate the quality of natural language generations that demonstrate high correlations with human judgements (Zheng et al., 2023b; Liusie et al., 2024b; Bubeck et al., 2023; OpenAI; Dubois et al., 2023; Wang et al., 2023b; Chiang and Lee, 2023).

There are two standard approaches to using LLM in judging responses. **Absolute scoring**: Prompt an LLM to evaluate a certain attribute of a response on a defined scale (e.g., a scale of 1 to 10). **Comparative scoring**: Prompt an LLM to choose which of the two responses to a given query displays higher quality with respect to a given attribute. Absolute scoring is a straightforward and effective method for evaluating a variety of responses to a query and ranking them. However, the scores obtained are unreliable and may vary significantly between different LLM judges. Alternatively, the more expensive comparative scoring approach has consistently demonstrated higher correlations with human judgements (Zheng et al., 2023b; Liusie et al., 2024b; Qin et al., 2023). However, a drawback of this approach is that it scales quadratically with the number of response candidates, which can become prohibitively expensive due to the inference costs of LLMs.

To address the computational restraints of comparative assessment, various approaches have been proposed. Notably, it is possible to extract more information from the LLM-as-a-judge rather than just a binary decision. Various works have, as opposed to using simple win-ratio, resorted to using the average probability output from the LLM (Qin et al., 2023; Zheng et al., 2023b; Liusie et al., 2024b; Park et al., 2024; Molenda et al., 2024). Building on this idea, Liusie et al. (2024c) introduced a Product-of-Experts (PoE) (Hinton, 1999; Welling, 2007) framework in modelling comparative scoring. In principle, the joint distribution of candidate scores can be broken down into arbitrarily chosen experts that model the score differences of two instances at a time, allowing for a partial set of comparisons to model the full joint distribution. This directly allows one to obtain a ranking of candidates with-

out having to perform all possible comparisons, showing that only a fraction of the total number of comparisons is needed to obtain highly competitive performance.

Contributions: In this paper, we generalise the expert in the comparative framework and derive how there is a wide range of viable options. Starting from the Beta distribution in modelling arbitrary probabilities, we show that prior works are specific instances of this choice. We also propose improved estimates for the uncertainty in individual comparisons and show how these updated uncertainties allow us to make even fewer comparisons without loss of performance. In addition to the uncertainty of a comparison, we also propose an uncertainty in the overall ranking of a set of candidates. Finally, we show that the Product-of-Experts framework easily lends itself to combinations of various scoring approaches. Specifically, we show that absolute scoring can complement comparative scoring efficiently and cheaply and lead to improved overall performance.

2 Background and Related Work

NLG Evaluation with LLMs: The extensive natural language generation capabilities of instruction-tuned large language models (Achiam et al., 2023; Ouyang et al., 2022; Chung et al., 2024; Dubey et al., 2024) have prompted recent work on open-ended generation evaluation using LLMs. Methods such as GPTScore (Fu et al., 2023) rank responses based on the likelihood of generation and G-Eval (Liu et al., 2023) which uses chain-of-thought and form-filling to evaluate the quality of a response. Furthermore, LLM-as-a-judge (Zheng et al., 2023b) approaches score responses on an absolute scale (Wang et al., 2023a; Kocmi and Federmann, 2023) or comparative manner by comparing responses against each other (Qin et al., 2023; Liusie et al., 2024b,c) and building an overall ranking through the set of pairwise comparisons.

LLM-Based Comparative Assessment: The work by Liusie et al. (2024b) showed that comparative assessment yields superior performance compared to absolute scoring methods and various custom baselines. By making all possible $N(N-1)$ pairwise comparisons of N candidate responses, and computing the win-ratio, an overall ranking can be obtained. This style of approach has found its applications in many places. Qin et al. (2023) utilised pairwise comparisons to retrieve relevant sources, using both the full set of comparisons and sorting-based algorithms. Park et al. (2024) employed comparative assessment for dialogue evaluation, calculating the average probability across a randomly sampled set of comparisons to determine score quality. Finally, Liu et al. (2024b) demonstrated the limitations of LLM scoring, and resorted to using pairwise comparisons. They introduced PAirwise-preference Search, a variation of the merge sort algorithm which utilises LLM probabilities.

Ranking from Pairwise Comparisons: The idea of generating a full ranking from pairwise comparisons has been extensively studied. Arguably the most well-known example of this is the ranking of tennis players based on the outcomes of games and this kind of problem has applications in many different areas. Anything from sports and gaming (Beaudoin and Swartz, 2018; Csató, 2013; Motegi and Masuda, 2012), web search (Cossock and Zhang, 2006; Dwork et al., 2001), social studies (Manski, 1977; Louviere et al., 2000) to information retrieval (Cao et al., 2007; Liu et al., 2009) requires modelling through pairwise events. The most common approach to model pairwise comparisons is the Bradley-Terry (BT) model (Bradley and Terry, 1952). By assigning each candidate a latent score, the probability of one candidate winning over another is based on the underlying skill difference. The latent scores can then be obtained by maximising the log-likelihood of the data (David, 1963; Davidson and Farquhar, 1976; Cattelan, 2012). Finally, the TrueSkill model (Herbrich et al., 2006; Minka et al., 2018) generalises the Bradley-Terry model by incorporating uncertainties in candidate scores within a Bayesian framework in a sports context.

Product-of-Experts: Each comparison $\mathcal{C}_k$ made in a comparative assessment framework provides information about the distribution of scores $\boldsymbol{s} = s_{1:N}$ of the candidates. The Product-of-Experts (PoE) (Hinton, 1999; Welling, 2007) approach presents a simple and effective way of combining the information from comparisons according to:

$$\mathrm{p}(\boldsymbol{s}|\mathcal{C}_{1:K}) \propto \prod_k \mathrm{p}(\boldsymbol{s}|\mathcal{C}_k) = \prod_k \mathrm{p}(s_i - s_j|\mathcal{C}_k)$$

The distribution can be simplified into a product of individual experts, and be further simplified as it involves comparisons made exclusively between pairs of candidates. In the work of Liusie et al.

(2024c) there were two choices made for the expert: The Gaussian motivated from its algebraic tractability and the soft Bradley-Terry model as an extension to the standard BT model:

$$\mathrm{p}(s_i - s_j|\mathcal{C}_k) \propto \sigma(s_i - s_j)^{p_{ij}} (1 - \sigma(s_i - s_j))^{1-p_{ij}}$$

Where the probability p_{ij} is obtained from the LLM when comparing candidate i against j. The more experts/comparisons are included the better the resulting estimate of the scores should be. The scores can then be retrieved by optimising $\mathrm{p}(\boldsymbol{s}|\mathcal{C}_{1:K})$ either through iterative (Zermelo, 1929; Dykstra, 1956; Newman, 2023), algebraic (Liusie et al., 2024c) or standard gradient-based approaches.

3 Generalised Expert Modelling

This section will focus on the nature of the expert for modelling absolute and pairwise comparisons. In general, we have N candidates with associated scores $s_{1:N}$. Given a set of comparisons $\mathcal{C}_{1:K}$, the aim is to predict a set of scores $\hat{s}_{1:N}$ which ranks the candidates as closely as possible to the true ranking. Each comparison $\mathcal{C}_k = (i, j, p_{ij})$ contains the ids of the candidates and the corresponding probability produced by the LLM that i is better than j for a certain attribute.

3.1 Comparative Expert Modelling

The experts in prior work all modelled the score difference of candidates $\mathrm{p}\left(s_i - s_j|\mathcal{C}_k\right)$. Through a simple change of variables we propose a more generalised version of the expert in comparative modelling:

$$\mathrm{p}\left(s_i - s_j|\mathcal{C}_k\right) = f'(s_i - s_j)\tilde{\mathrm{p}}\big(f(s_i - s_j)\big|\mathcal{C}_k\big)$$

where the $f(\cdot)$ is a generic monotonically increasing function. Any choice of f is viable as long as the distribution $\tilde{\mathrm{p}}$ supports it. Through this view, we can easily obtain the Gaussian expert in Liusie et al. (2024c), by using an identity f and a simple Gaussian for $\tilde{\mathrm{p}}$. Alternatively, by letting f be the sigmoid function and using an underlying Beta distribution:

$$\tilde{\mathrm{p}}\big(f\big|\mathcal{C}_k\big) = \mathcal{B}\big(f; p_{ij}, 1 - p_{ij}\big)$$

we regain the soft Bradley-Terry model. However, from this point of it is clear that there are many more viable options for modelling a pairwise event. In this work, we will investigate several other combinations starting with a sigmoid and a general:

$$\tilde{\mathrm{p}}\big(f\big|\mathcal{C}_k\big) = \mathcal{B}(f; p_{ij} + \alpha, 1 - p_{ij} + \beta) \quad (1)$$

model. We will also try an unconventional choice by replacing the identity function in the Gaussian model with a sigmoid-like function:

$$\mathrm{p}\left(s_i - s_j|\mathcal{C}_k\right) = \mathcal{N}(s_i - s_j; 0, 1)\mathcal{N}\big(\Phi(s_i - s_j); p_{ij}, 1\big) \quad (2)$$

where $\Phi(\cdot)$ is the cumulative density of the Gaussian. Ablations will study the impact of the choice of f and the underlying distribution $\tilde{\mathrm{p}}$.

3.2 Expert Combinations

As has been mentioned several times, absolute scoring is cheaper but worse than comparative scoring. However, it is possible that absolute scoring can provide complementary information so we propose combining the two approaches into a single model:

$$\mathrm{p}(\boldsymbol{s}|\mathcal{C}_{1:K}, \mathcal{A}_{1:N}) \propto \prod_k \mathrm{p}(s_i - s_j|\mathcal{C}_k) \prod_n \mathrm{p}(s_n|\mathcal{A}_n)$$

where $\mathcal{A}_n$ contains the information from an absolute scoring prompt. When prompting an LLM on a scale of 1 to 10 it contains the probabilities of those values. Unfortunately, absolute scoring will provide a discrete score and while it is possible to obtain the associated logits for each value (of 1 to 10) from the LLM (and construct a categorical distribution) it remains difficult to combine the continuous pairwise experts with the discrete absolute experts. Therefore, we also propose using moment matching to transform the categorical expert into a Gaussian $\mathcal{N}(\mu, \sigma^2)$:

$$\mu = \sum_c c p_c, \quad \sigma^2 = \sum_c (c - \mu)^2 p_c$$

where c and p_c represent the class and the associated probability that the LLM would output that class. In our running example, we would have $c \in \{1, \ldots, 10\}$. Finally, the absolute expert can be written as:

$$\mathrm{p}(s_n|\mathcal{A}_n) \approx \mathcal{N}(s_n; \mu_n, \sigma_n^2), \ \forall n = 1, \ldots, N$$

and would allow us to operate with continuous values and optimise the skill scores using gradient-based approaches. In this work, we rely solely on simple absolute scoring but there exist many more sophisticated pointwise scoring approaches like G-Eval (Liu et al., 2023) which can provide further improvements.

3.3 Home Advantage

A big issue plaguing LLM-based approaches is bias in the system. In our case, the probability outputs of an LLM are inconsistent $p_{ij} \neq 1 - p_{ji}$, meaning that the LLM-based judge can assign conflicting probabilities when comparing i to j as opposed to j to i. This stems from positional biases in the system (Zheng et al., 2023b; Chen et al., 2024a; Liusie et al., 2024a; Zheng et al., 2023a; Wang et al., 2023c; Dominguez-Olmedo et al., 2024; Zhu et al., 2023; Chen et al., 2024b; Liu et al., 2024a). To resolve such an issue, we rely on one of the two approaches. **Permutation Debiasing:** For each comparison we make two LLM calls for both i vs j and j vs i to obtain a final debiased probability $\tilde{p}_{ij} = \frac{1}{2}(p_{ij} + (1 - p_{ji}))$ which would ensure consistency: $\tilde{p}_{ij} = (1 - \tilde{p}_{ij})$. **Home Advantage:** An alternative approach is to directly incorporate the positional bias into the comparative expert model. Since we already know that a certain position will be preferred over another we can introduce a 'home advantage' (Agresti, 1990; Caron and Doucet, 2012) parameter to model the inconsistency through our function f:

$$f(s_i - s_j; \Delta) = f(s_i - s_j - \Delta)$$

While prior approaches have developed the theory for home advantage in specific use-cases such as Bradley-Terry (Caron and Doucet, 2012) and Gaussian experts (Liusie et al., 2024c), our parameterisation through the generic function f allows us to straightforwardly incorporate home advantage into any type of expert. Furthermore, while the work of Liusie et al. (2024c) estimated the advantage parameter Δ through bespoke rules for each expert, we estimate it by maximising the likelihood $\mathrm{p}(\boldsymbol{s}|\mathcal{C}_{1:K}, \Delta)$.

4 Uncertainty Estimation

This section will explore how to estimate uncertainty when ranking examples. Two levels of uncertainty will be explored. **Pairwise uncertainty:** The uncertainty in the score difference of a pair of candidates. **Ranking uncertainty:** The uncertainty in the overall ranking of a set of candidates. Being able to estimate these uncertainties robustly can help in reducing the number of comparisons needed to achieve good performance and in understanding whether the overall ranking is trustworthy.

4.1 Laplace's approximation

Unfortunately, it is generally analytically intractable to derive uncertainties from the distribution $\mathrm{p}(\boldsymbol{s}|\mathcal{C}_{1:K})$ and we will therefore, in all cases, apply Laplace's approximation:

$$\mathrm{p}(\boldsymbol{s}|\mathcal{C}_{1:K}) \approx \mathcal{N}\left(\boldsymbol{s}; \boldsymbol{\mu}^{(K)}, \boldsymbol{\Sigma}^{(K)}\right)$$

where we set (and dropped the superscript for conciseness):

$$\boldsymbol{\mu} = \arg\max_{\boldsymbol{s}} \ln \mathrm{p}(\boldsymbol{s}|\mathcal{C}_{1:K})$$

$$\boldsymbol{\Sigma}^{-1} = -\nabla\nabla \ln \mathrm{p}(\boldsymbol{s}|\mathcal{C}_{1:K})\big|_{\boldsymbol{\mu}}$$

There are more advanced approaches to approximating an intractable distribution but we will rely on this simple and efficient scheme in this work.

4.2 Pairwise Uncertainty Estimation

Being able to estimate the uncertainty in a pair of candidates, in their score difference, can allow us to decide which comparisons are useful and which or not. The better the quality of the uncertainty estimate, the fewer comparisons are needed to achieve good performance. Following Liusie et al. (2024c), the aim is to iteratively add additional comparisons to improve the product-of-experts model and the overall ranking of scores.

Given the Gaussian approximation, Liusie et al. (2024c) posed that the next comparison that should be selected should induce **minimum overall uncertainty** in the resulting distribution, giving the following selection criteria under the soft BT model:

$$\hat{i}, \hat{j} = \arg\max_{i,j} \ \sigma(\mu_i - \mu_j)\,\sigma(\mu_j - \mu_i) \cdot \left(\Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj}\right) \quad (3)$$

In this work, we propose two additional pairwise uncertainty metrics which apply to any modelling choice: A simple **variance in score differences** and the **probability of reordering**. The variance $\mathbb{V}[s_i - s_j|\mathcal{C}_{1:K}] = \Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj}$ is straightforward and easy to compute. Since we want to select the pairs with the highest variance (or uncertainty) we get the following selection mechanism:

$$\hat{i}, \hat{j} = \arg\max_{i,j} \ \Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj} \quad (4)$$

However, since we are interested in the correct ranking of candidates, a potentially more appropriate uncertainty metric is the probability that two

candidates are reordered. Assuming that $s_i > s_j$ the probability of reordering then becomes:

$$\mathtt{P}(s_i < s_j | \mathcal{C}_{1:K}) = \Phi\left(\frac{\mu_j - \mu_i}{\sqrt{\Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj}}}\right)$$

As shown in the Appendix C, this can be simplified to a similar selection form since we want to pick examples with the highest reordering probability:

$$\hat{i}, \hat{j} = \arg\max_{i,j} \frac{\Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj}}{(\mu_i - \mu_j)^2} \quad (5)$$

Furthermore, since all selection mechanisms take the following form:

$$\hat{i}, \hat{j} = \arg\max_{i,j} \; w(\mu_i - \mu_j)\left(\Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj}\right)$$

we run ablation studies on changing the weight function $w(\cdot)$.

4.3 Ranking Uncertainty Estimation

In addition to pairwise uncertainties, there is a need to understand the level of uncertainty in the overall ranking $\mathtt{P}(s_1 < \cdots < s_N | \mathcal{C}_{1:K})$ but this is difficult to estimate even under a Gaussian model. Instead we use the following metrics to represent the overall uncertainty. The entropy in the Gaussian approximation:

$$\mathcal{H}[\boldsymbol{s}|\mathcal{C}_{1:K}] = \frac{N}{2}(1 + \ln(2\pi)) + \frac{1}{2}\det(\boldsymbol{\Sigma})$$

The lower the entropy, the more certain we should be in our predicted ranking. While the entropy of the rank distribution should in reality replace the entropy of the score distribution, the issue is that there are $N!$ different possible rankings and obtaining accurate estimates means needing to sample an order of magnitude more than possible rankings.

5 Experimental Setup

5.1 Datasets

We mainly perform experiments on the summary evaluation SummEval dataset (Fabbri et al., 2021) which contains 100 articles, each with 16 machine-generated summaries evaluated on four different attributes: coherency (COH), consistency (CON), fluency (FLU), and relevancy (REL). We will also use the much larger HANNA dataset (Chhun et al., 2022) which has 1056 machine-generated stories annotated by humans on six different attributes. These are averaged to a single overall quality score.

5.2 Methodology

We will be relying on Flan-T5 (Chung et al., 2024) (3B) system to evaluate performance on the SummEval dataset and Qwen2.5-{3B, 7B}-Instruct (Qwen Team, 2024) systems on the larger HANNA dataset. Appendix A will detail our choices, how we structure the prompts and how the probabilities are extracted from each model.

Probabilistic Models: In almost all experiments we will rely on the soft Bradley-Terry extension as our baseline expert model following. This will be compared against our proposed extensions to this approach: (1) The generalised Beta distribution in Eq. (1), (2) the extended Gaussian distribution in Eq. (2) and (3) the combination of comparative and absolute outputs in a single PoE model. We will not include simple baselines such as average win-ratio and average probability since these have been shown to be inferior on a wide range of tasks (Liusie et al., 2024c; Raina et al., 2024).

Iterative Selection: We will also compare the above probabilistic models in an active learning framework where each model needs to select the comparisons that will induce the best performance. The baseline will be the **minimum uncertainty** approach given in Eq. (3) using the soft BT model. We will compare this against our proposed **variance** (Eq. (4)) and **probability of reordering** (Eq. (5)) selection mechanisms.

Ranking Uncertainties: The ranking uncertainty metrics will be investigated on how well they correlate with the actual performance of the predicted rankings, whether they can identify high-performing predictions.

Evaluation Metrics: Since we are interested in predicting the ranking of candidate responses given a context/query, the main performance metric is Spearman rank correlation between the predicted and the human labelled scores. In SummEval ($N = 16$) we perform absolute and comparative scoring and evaluate the average Spearman across all contexts. For HANNA ($N = 1056$) we rank all generated stories. Furthermore, we assess the quality of the various iterative selection schemes by the number of comparisons needed to achieve good Spearman rank correlation. The ranking uncertainties are also evaluated using the area under the receiver operating characteristic curve (AUROC) to detect well-performing rankings.

6 Results

6.1 Form of Distribution

This section investigates a wide range of PoE models conditioned on the full set of comparisons. For SummEval with $N = 16$ summaries per context, there is a total of $K = 240$ number of comparisons. Unless reported otherwise, all results are based on the direct biased outputs of judges. Surprisingly, all experts perform similarly when evaluated on the full set of comparisons. The only model that performs noticeably better is the combined comparative-absolute model. Absolute scoring seems to extract complementary information and give the overall system a performance boost. While the ranks predicted by models are similar, the ranking uncertainties of these systems differ.

Table 1: Spearman correlations (%) for SummEval using Flan-T5 (3B) as a judge.

Function f	Distribution $\tilde{\mathrm{p}}$	COH	CON	FLU	REL
x	Gaussian	49.1	45.2	32.5	42.2
σ	Gaussian	49.2	45.3	32.5	42.2
Φ	Gaussian	49.2	45.3	32.5	42.3
σ	Beta	49.2	45.2	32.5	42.2
Φ	Beta	49.2	45.3	32.6	42.3
Comparative (σ-Beta) + Absolute Experts		50.7	45.9	32.9	43.4

To benchmark, an uncertainty is predicted for each context, meaning to represent how well the predicted ranks performs. The Spearman rank performance of each context is thresholded by the median score and mapped to a binary value so that contexts are classed as '0' or '1'. This allows us to use the standard AUROC score to evaluate detection performance. Table 2, shows the AUROC performance of a range of probabilistic models. Unlike previous results, the nature of the probabilistic model seems to have a significant impact on performance. While the performance of various PoE models are similar, the predicted entropy, and by equivalence, the hessian (curvature) of the log-likelihood seems to differ between certain models. This seems to stem mainly from the choice of f but further investigations are needed to understand how to predict robust uncertainties.

Table 2: AUROC (%) detection performance of $\mathcal{H}$, the entropy of Laplace's approximation.

Function f	Distribution $\tilde{\mathrm{p}}$	COH	CON	FLU	REL
x	Gaussian	53.4	51.8	54.9	58.4
σ	Gaussian	65.5	67.3	60.8	65.8
Φ	Gaussian	65.0	67.3	61.0	65.9
σ	Beta	63.8	68.3	60.6	63.8
Φ	Beta	64.8	67.4	60.9	65.7
Comparative (σ-Beta) + Absolute Experts		64.5	69.9	62.0	67.0

The linear-Gaussian model performs significantly worse, only marginally outperforming a random classifier. Furthermore, all models that map skill differences through a sigmoid-like function f perform notably better, with the combined model being the best system similar to previous results. While this is only a preliminary investigation into detecting well-performing examples, it is a good start into understanding the problem and what models produce better uncertainties.

6.2 Iterative Selection

In this section, we explore the quality of various models and uncertainty metrics when iteratively selecting comparisons. Furthermore, this investigation is performed on both biased and debiased probabilities. Four main points can be observed from the results in Figure 1:

(1) Model Convergence: As reported in the table above, all models converge towards the same final performance since the rankings are predicted from the same full set of comparisons.

(2) Quality of Uncertainties: There is a significant gap in performance between various uncertainty metrics. Probability of reordering is shown to outperform the minimum uncertainty metric in all attributes of SummEval for both biased and debiased cases. We expect this performance difference to originate from how each metric was derived. While minimum uncertainty is simply focused on choosing the next comparison that would induce the least uncertainty, the probability of reordering is directly linked to achieving the correct ranking.

(3) Expert Model Invariance: Probability of reordering under two different models, σ-Beta (soft BT) and Φ-Gaussian (ngaussian), perform almost identically in all cases. This again reinforces the idea that the nature of the expert model does not matter as much as the uncertainty modelling used in selecting the comparisons.

(4) Performing better with less: Performance is expected to increase as one adds more and more comparisons to the PoE model. However, in many of the cases above, performance drops until the full

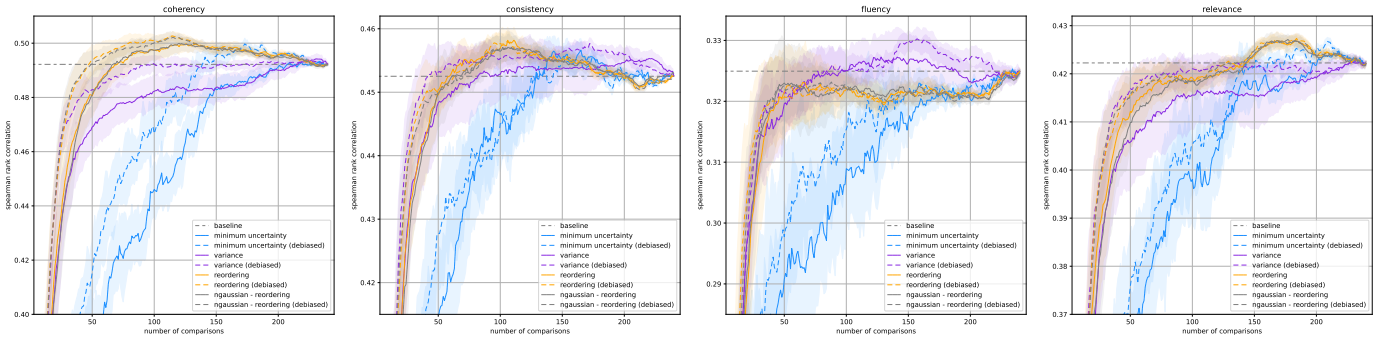


Figure 1: The Spearman Rank Correlation when iteratively selecting the next examples of lowest confidence/highest uncertainty. The baseline is the soft Bradley-Terry model with the minimum uncertainty metric. We also report the proposed variance and probability of reordering under the soft BT model. Furthermore, the Φ-Gaussian model is referred to as "ngaussian". Debiased refers to permutation debiased probabilities.

set of comparisons is reached. This is related to an overconfidence issue plaguing the uncertainty estimates. We show in Appendix B how the LLM-as-a-judge is miscalibrated, and how temperature annealing is not enough to calibrate and solve the overconfidence issue.

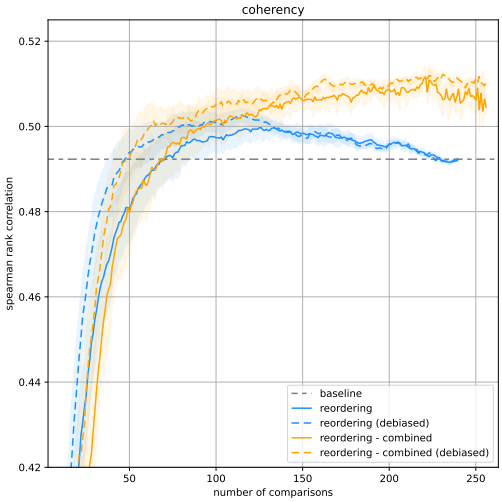


Figure 2: Comparing the soft BT model to the combined model which relies on additional absolute experts. Both use the probability of reordering as the selection criteria.

In Figure 2, we compare the soft BT model with the combined comparative-absolute model. Due to the cost of obtaining $N = 16$ absolute experts, the combined model initially performs worse. However, both biased and debiased combined models outperform the final soft BT model performance using a fraction of comparisons.

6.2.1 Ablation Studies

This section will explore various nuances in the modelling choices. In Figure 3, we vary the parameters of the underlying Beta distribution in a generalised soft BT model. Evaluated on SummEval

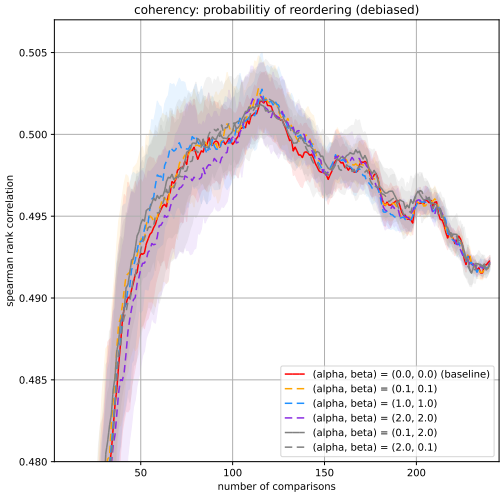


Figure 3: Selection performance. Varying the parameters α and β of the Beta distribution in Eq. (1).

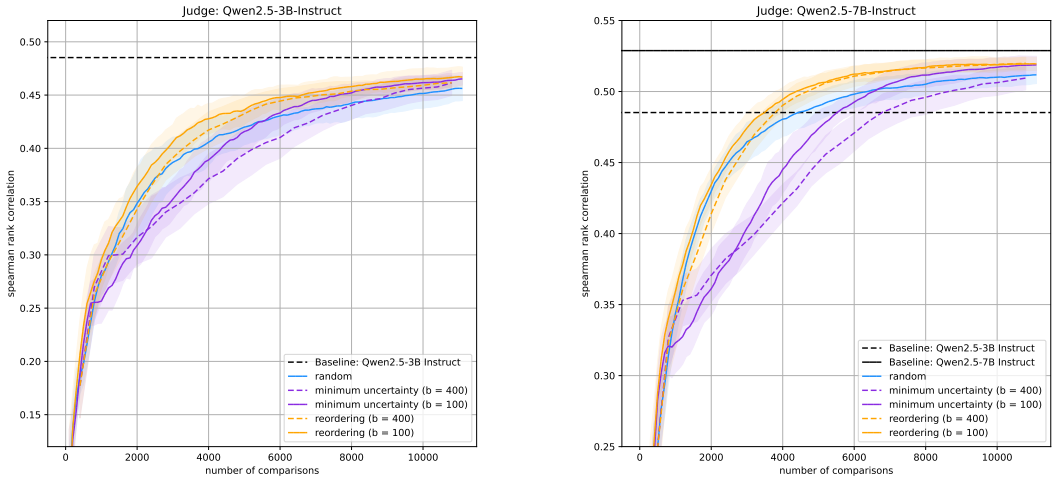
(COH), it is clear that the underlying Beta distribution has negligible impact on both the selection process and the final performance. Furthermore, we explore generalising the selection metrics with the following:

$$\hat{i}, \hat{j} = \arg\max_{i,j} \frac{\Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj}}{|s_i - s_j|^{\epsilon}} \qquad (6)$$

where we vary the exponent ϵ. Setting $\epsilon = 0$ returns variance while $\epsilon = 2$ gives probability of reordering. The exponent is swept on the same benchmark in Figure 5. It is evident that while $\epsilon = 0$ suffers in performance, most of the other values perform similarly. The best-performing option is $\epsilon = 0.5$ which slightly outperforms other options including the probability of reordering.

6.3 Large-Scale Selection

All results have been focused on ranking a small set of candidates $N = 16$ which has $K = 240$



(a) Using Qwen2.5-3B-Instruct as the judge.

(b) Using Qwen2.5-7B-Instruct as the judge.

Figure 4: The Spearman Rank Correlation when iteratively selecting the next batch b of examples of lowest confidence/highest uncertainty. The baselines are Qwen2.5-{3B-7B} models with all comparisons selected. Furthermore, the efficient baseline is set to the soft BT model with the minimum uncertainty metric.

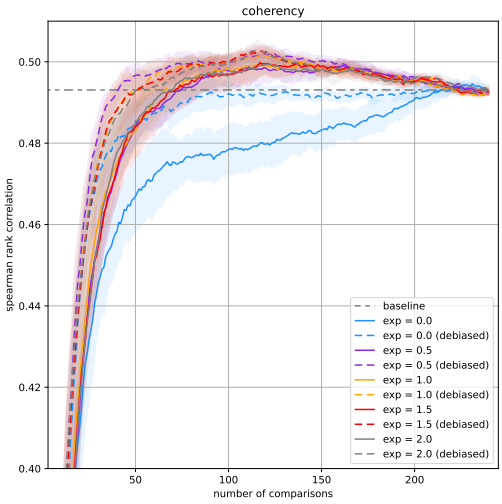


Figure 5: Selection performance. Varying the parameter ϵ in Eq. (6).

possible comparisons. In this section we scale up to the HANNA dataset with $N = 1056$ stories and $K = 1114080$ possible comparisons. Furthermore, due to the cost of iteratively selecting a single comparison, re-estimating Laplace's approximation, and repeating this process, we opt to perform batch acquisitions with $b = \{100, 400\}$. This will showcase in the extreme case how well our best-proposed uncertainty metric, probability of reordering, performs compared to random and minimum uncertainty selection. We will only evaluate the iterative process up until 1% of all possible comparisons to test the efficiency of approaches.

In Figure 4, various iterative schemes are reported. Furthermore, the performance of the soft BT model with all possible comparisons is reported under both backbone judges. In these results, one can observe minimum uncertainty to suffer significantly compared to the simplest baseline of random selection when using both Qwen2.5-3B and 7B as backbone judges. This is caused due to the lack of diversity when selecting a batch of comparisons. While reducing the batch size of acquisitions helps performance, it still lacks significantly compared to the probability of reordering which is far more robust to larger batch acquisitions.

7 Conclusions

This paper generalised probabilistic modelling for comparative LLM-as-a-judge, demonstrating that existing approaches are specific instances of a broader framework. We introduced improved uncertainty estimates for individual comparisons and overall rankings, leading to more efficient iterative selection strategies. Notably, the probability of reordering proved to be a superior metric for selecting informative comparisons. We also showed the benefits of combining absolute and comparative scoring within a Product-of-Experts framework, achieving enhanced performance. While the specific expert model had limited impact on final rankings given sufficient comparisons, the choice of uncertainty estimation and the incorporation of absolute scoring significantly improved efficiency and accuracy. Our findings highlight the importance of robust uncertainty estimation in LLM-based evaluation and provide a more flexible and efficient framework for comparative assessment.

8 Limitations

The main concern lies in the quality of the estimated uncertainties, which are crucial for the efficiency of the proposed iterative selection methods. The reliance on Laplace's approximation to derive these uncertainties introduces potential inaccuracies. This approximation assumes that the posterior distribution over model parameters is approximately Gaussian, which may not hold true in all scenarios, particularly when the true posterior is multimodal or exhibits significant skewness. Consequently, the derived uncertainty metrics, such as the variance and probability of reordering, might not perfectly reflect the true uncertainty in the model's predictions.

Furthermore, the calibration of the LLM-as-a-judge is a non-trivial challenge. Although we demonstrate that temperature annealing is insufficient to fully address the overconfidence issue, the development of more sophisticated calibration techniques could enhance the reliability of the probability outputs and, consequently, the accuracy of the uncertainty estimates. Additionally, the computational cost of obtaining comprehensive pairwise comparisons for large-scale datasets remains a practical constraint. While the proposed methods improve efficiency, exploring alternative approaches could further reduce the cost of evaluations. Finally, the generalisation of these findings to other domains and tasks beyond summary and story evaluation should be approached with caution, as the performance of LLM-as-a-judge can vary depending on the specific evaluation criteria and the nature of the generated text.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alan Agresti. 1990. *Categorical data analysis*. John Wiley & Sons.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

David Beaudoin and Tim Swartz. 2018. A computationally intensive ranking system for paired comparison data. *Operations Research Perspectives*, 5:105–112.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136.

Francois Caron and Arnaud Doucet. 2012. Efficient bayesian inference for generalized bradley–terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196.

Manuela Cattelan. 2012. Models for paired comparison data: A review with emphasis on dependent data.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or llms as the judge? a study on judgement biases. *Preprint*, arXiv:2402.10669.

Xinyun Chen, Ryan A. Chi, Xuezhi Wang, and Denny Zhou. 2024b. Premise order matters in reasoning with large language models. *Preprint*, arXiv:2402.08939.

Cyril Chhun, Pierre Colombo, Fabian Suchanek, and Chloé Clavel. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5794–5836.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*.

David Cossock and Tong Zhang. 2006. Subset ranking using regression. In *Learning Theory: 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22-25, 2006. Proceedings 19*, pages 605–619. Springer.

László Csató. 2013. Ranking by pairwise comparisons for swiss-system tournaments. *Central European Journal of Operations Research*, 21:783–803.

Herbert Aron David. 1963. *The method of paired comparisons*, volume 12. London.

Roger R Davidson and Peter H Farquhar. 1976. A bibliography on the method of paired comparisons. *Biometrics*, pages 241–252.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. *Preprint*, arXiv:2306.07951.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 30039–30069. Curran Associates, Inc.

Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. 2001. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622.

Otto Dykstra. 1956. A note on the rank analysis of incomplete block designs–applications beyond the scope of existing tables. *Biometrics*, 12(3):301–306.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems*, 19.

Geoffrey E. Hinton. 1999. Products of experts. In *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, volume 1, pages 1–6. IET.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024b. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *Preprint*, arXiv:2403.16950.

Adian Liusie, Yassir Fathullah, and Mark JF Gales. 2024a. Teacher-student training for debiasing: General permutation debiasing for large language models. *arXiv preprint arXiv:2403.13590*.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024b. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Adian Liusie, Vatsal Raina, Yassir Fathullah, and Mark Gales. 2024c. Efficient llm comparative assessment: a product of experts framework for pairwise comparisons. *arXiv preprint arXiv:2405.05894*.

Jordan J Louviere, David A Hensher, and Joffre D Swait. 2000. *Stated choice methods: analysis and applications*. Cambridge university press.

Charles F Manski. 1977. The structure of random utility models. *Theory and decision*, 8(3):229.

Tom Minka, Ryan Cleven, and Yordan Zaykov. 2018. Trueskill 2: An improved bayesian skill rating system. *Technical Report*.

Piotr Molenda, Adian Liusie, and Mark J. F. Gales. 2024. Waterjudge: Quality-detection trade-off when watermarking large language models. *Preprint*, arXiv:2403.19548.

Shun Motegi and Naoki Masuda. 2012. A network-based dynamical ranking system for competitive sports. *Scientific reports*, 2(1):904.

MEJ Newman. 2023. Efficient computation of rankings from pairwise comparisons. *Journal of Machine Learning Research*, 24(238):1–25.

OpenAI. Evals is a framework for evaluating llms and llm systems, and an open-source registry of benchmarks. Accessed: 2024-12-04.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

ChaeHun Park, Minseok Choi, Dohyun Lee, and Jaegul Choo. 2024. Paireval: Open-domain dialogue evaluation with pairwise comparison. *arXiv preprint arXiv:2404.01015*.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Vatsal Raina, Adian Liusie, and Mark Gales. 2024. Finetuning llms for comparative assessment tasks. *Preprint*, arXiv:2409.15979.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. Large language models are not fair evaluators. *Preprint*, arXiv:2305.17926.

Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. 2023c. Primacy effect of ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 108–115, Singapore. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

M. Welling. 2007. Product of experts. *Scholarpedia*, 2(10):3879. Revision #137078.

Ernst Zermelo. 1929. Die berechnung der turnierergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1):436–460.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023a. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges.

A Prompting

When prompting an LLM for the score of a candidate, or which of two candidates is better, the important information lies in the logits of the tokens we are interested in. In this section we detail the design the of our prompts for both Flan-T5 and Qwen2.5. Since the former is an encoder-decoder foundation model and the latter is a decoder-only foundation model the prompts need to be designed slightly differently. Note, we use Flan-T5 on SummEval since it has been shown to be very competitive on the dataset (Liusie et al., 2024b,c) and Qwen2.5 models on HANNA due to an increased context size needed and complexity of the dataset.

Absolute prompting: For the Flan-T5 system, we give the encoder the following prompt:

```
Article: <context>\n\nSummary: <A>
\n\nScore the response between 1
and 10 based on how coherent the
summary is.
```

where we are scoring the coherency of a summary. The `<context>` and `<A>` are replaced by the article and summary. Following this we extract the logits corresponding to 1 to 10 from the decoder. The probability of each class is then:

$$p_c = \frac{\exp(z_c)}{\sum_{i=1}^{10} \exp(z_i)} \quad i = 1, \dots, 10$$

The choice of 1-10 is arbitrary and any other range could have been chosen.

Comparative prompting: For the Flan-T5 system, we give the encoder the following prompt:

```
Article: <context>\n\nSummary A: <A>
\n\nSummary B: <B>\n\nWhich Summary
is more coherent, Summary A or
Summary B?
```

The `<context>`, `<A>` and `<B>` are replaced by the article and two different summaries. Following this we give the following prefix to the decoder:

```
Summary
```

and extract the logits corresponding to A and B from the decoder. The prefix ensures that the probability mass of the next token is concentrated into the options "A" and "B". From these logits we extract the probability that A will win:

$$p = \frac{\exp(z_A)}{\exp(z_A) + \exp(z_B)}$$

Similarly, we prompt the Qwen2.5 system in the following matter when we want to rank various stories from HANNA:

```
{"role": "system", "content": "You
are an expert story assessor."},

{"role": "user", "content": "Story A:
<A>\n\nStory B: <B>\n\nWhich story
is better overall, Story A or B?
Answer only with Story A or Story B."}

{"role": "assistant", "content": "Story "}
```

These are then prepared by the Qwen2.5 tokenizer in the instruction following format and fed into the model. Following on, the logits corresponding to A and B are then extracted for the next token and converted into a probability.

B Calibration

This section reports the calibration error and reliability diagrams for the different metrics under a biased and debiased setup. The main point is to address the overconfidence issue related to our results in Section 6.2 and why temperature annealing is not enough to solve the problem.

The calibration is based on the confidence scores of individual comparisons $\max(p, 1-p)$. Therefore, when calibrating using temperature annealing, the resulting (binary) predictions remain the same:

$$\tilde{p} = \frac{p^{1/T}}{p^{1/T} + (1-p)^{1/T}}$$

To understand the impact of calibration, we find the optimal temperature on the SummEval dataset by minimising the expected calibration error, see Table 3. Each attribute has its optimal temperature. We also report the corresponding reliability diagrams in Figure 6. From these results, it is evident that simple temperature annealing can almost entirely resolve the miscalibration in the systems.

Table 3: Expected calibration error (%).

Method	Debiased	COH	CON	FLU	REL
-	✗	9.80	3.77	9.87	11.86
	✓	2.83	1.82	4.84	6.20
Calibrated	✗	1.02	0.68	1.28	0.98
	✓	2.58	1.72	1.08	1.07

This next part will check how temperature annealing affects the solution of a soft Bradley-Terry model. Starting from the gradient of the log-likelihood:

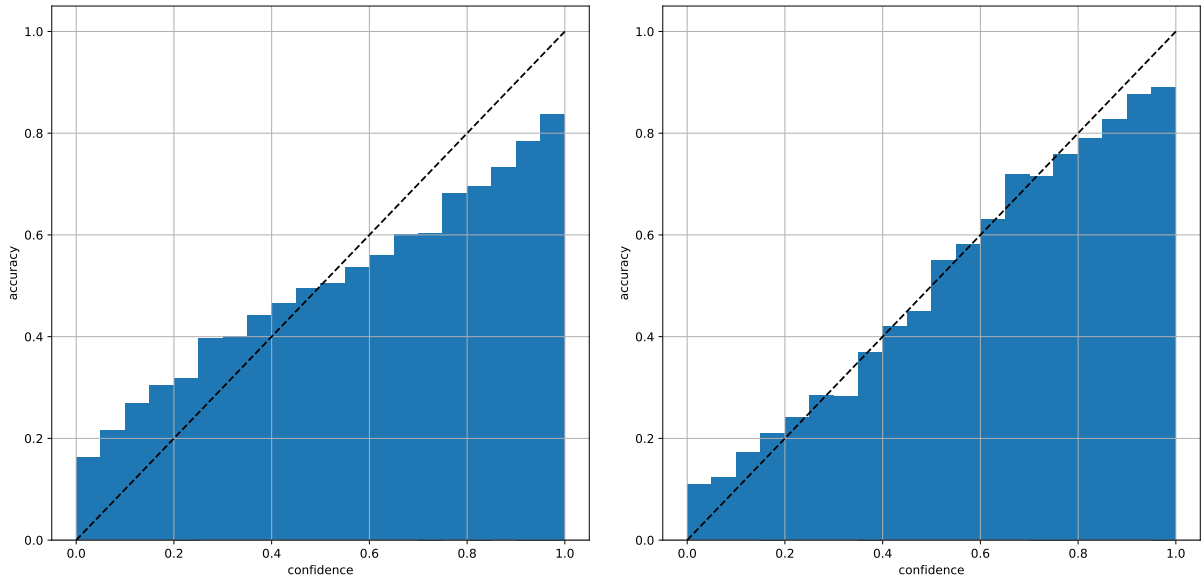
$$\nabla \ln \mathrm{p}(\boldsymbol{s}|\mathcal{C}_{1:K}) = \sum_{i,j \in \mathcal{C}_{1:K}} p_{ij} - \sigma(s_i - s_j) = 0$$

Looking at a single element of the sum, and under calibrated probabilities the new solution becomes:

$$\begin{aligned}
\tilde{p}_{ij} &= \sigma(\tilde{s}_i - \tilde{s}_j) \iff \\
\frac{p^{1/T}}{p^{1/T} + (1-p)^{1/T}} &= \sigma(\tilde{s}_i - \tilde{s}_j) \iff \\
\frac{1}{1 + \left(\frac{1-p}{p}\right)^{1/T}} &= \sigma(\tilde{s}_i - \tilde{s}_j) \iff \\
\frac{1}{1 + \exp\left(\frac{1}{T}\ln\left(\frac{1-p}{p}\right)\right)} &= \sigma(\tilde{s}_i - \tilde{s}_j) \iff \\
\sigma\left(\frac{1}{T}\ln\left(\frac{p}{1-p}\right)\right) &= \sigma(\tilde{s}_i - \tilde{s}_j) \iff \\
\frac{1}{T}\ln\left(\frac{p}{1-p}\right) &= \tilde{s}_i - \tilde{s}_j \iff \\
\frac{p}{1-p} &= \exp(T(\tilde{s}_i - \tilde{s}_j)) \iff \\
p &= \frac{\exp(T(\tilde{s}_i - \tilde{s}_j))}{1 + \exp(T(\tilde{s}_i - \tilde{s}_j))} \iff \\
p &= \frac{1}{1 + \exp(-T(\tilde{s}_i - \tilde{s}_j))} \iff \\
p &= \sigma(T(\tilde{s}_i - \tilde{s}_j))
\end{aligned}$$

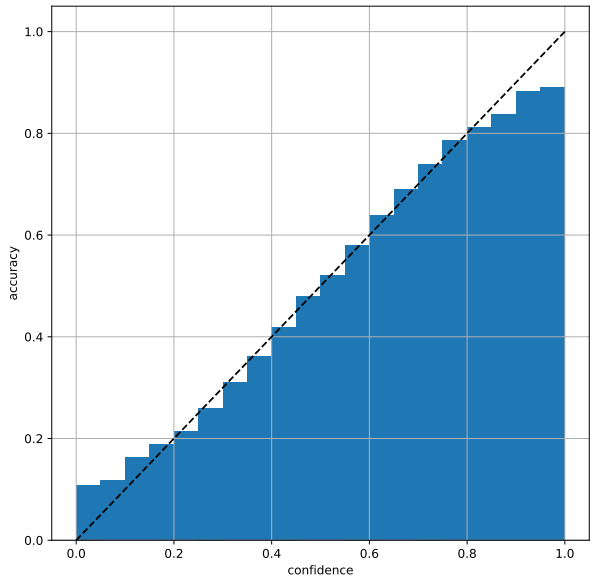
This shows that temperature annealing leads to a new solution of scores that are linearly scaled by the temperature T. Therefore, even if temperature annealing is enough to calibrate a system, it has no impact at all on the predicted rankings.

Instead, we report a different result, the diagrams in Figure 7. We obtain either the confidence of each comparison or the probability of reordering. Then the aim is to compute the accuracy of comparisons on a filtered dataset when removing the examples of lowest confidence/highest uncertainty. What one expects from high quality uncertainties is for the accuracy of the filtered dataset to improve as much as possible. While we observe that the accuracy improves as we reject samples, both metrics display a significant overconfidence issue; accuracy reduces when rejecting samples with the highest confidence and lowest uncertainty. This could partially explain why our results in Section 6.2 showcase a 'bump', where adding more comparisons decreases the system's performance. This also justifies using more advanced methods for calibrating the outputs of LLM-judges when using them to rank candidates.

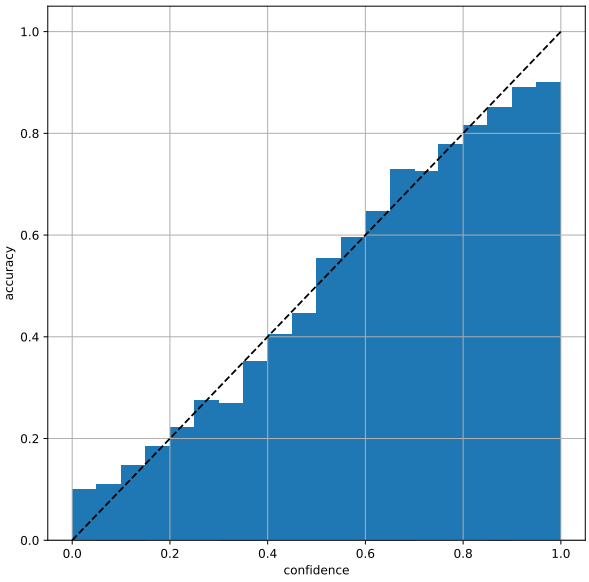


(a) `Biased`

(b) `Debiased`

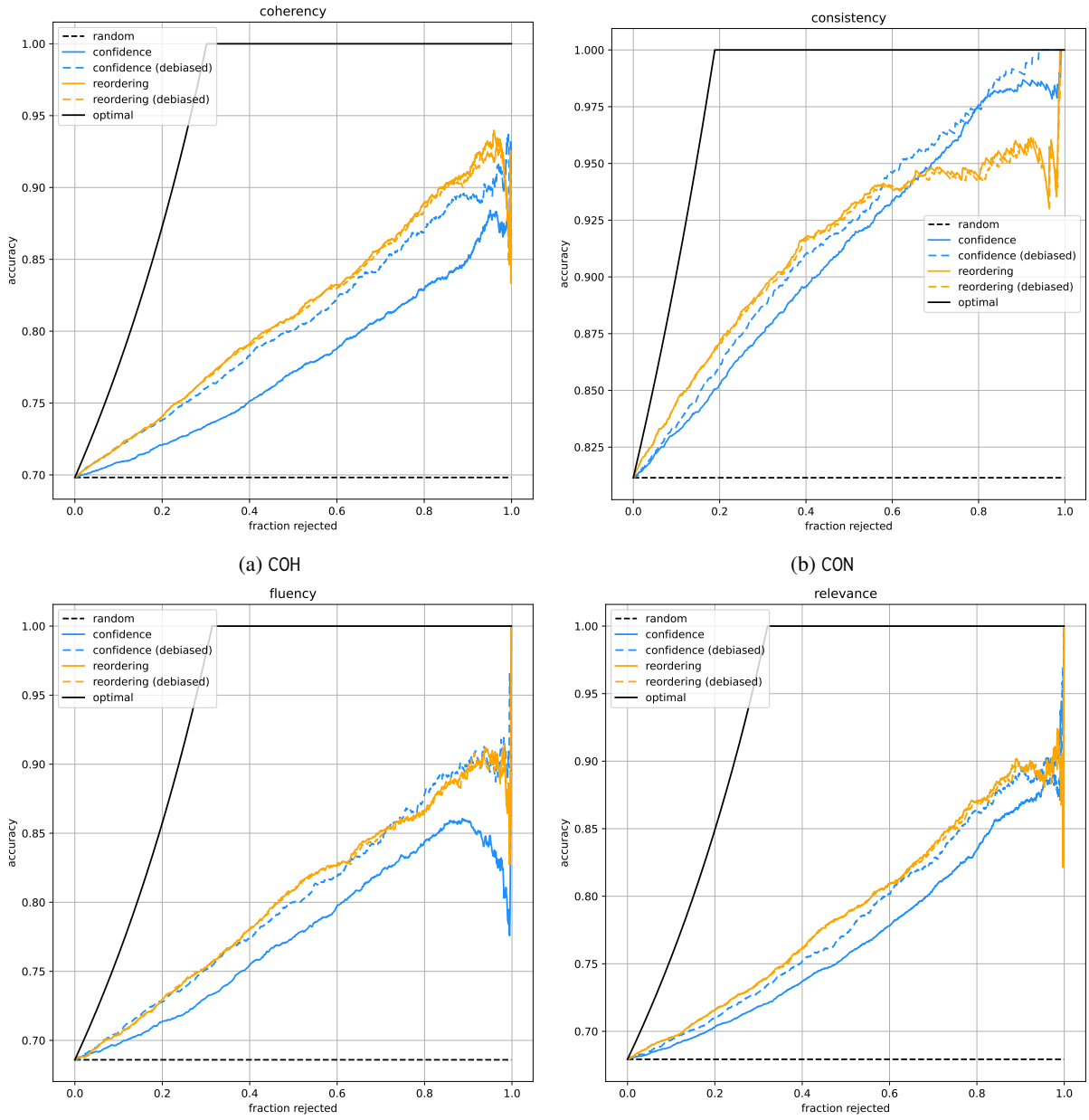


(c) `Biased - calibrated`



(d) `Deiased - calibrated`

Figure 6: The reliability diagram of biased and debiased, standard and calibrated systems on the `coherency` metric.



(a) COH (b) CON

(c) FLU (d) REL

Figure 7: The accuracy at a comparison-level when the examples of lowest confidence/highest uncertainty are rejected.

C Probability of Reordering

In this section we showcase how the probability of reordering can be rephrased to a familiar form: Assuming that $s_i > s_j$ ($\mu_i > \mu_j$) the probability of reordering becomes:

$$\mathtt{P}(s_i < s_j | \mathcal{C}_{1:K}) = \Phi\left(\frac{\mu_j - \mu_i}{\sqrt{\Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj}}}\right)$$

The selection is based on picking the examples with highest probability of reordering:

$$\begin{aligned}
\hat{i}, \hat{j} &= \arg\max_{i,j} \Phi\left(\frac{\mu_j - \mu_i}{\sqrt{\Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj}}}\right) \\
&= \arg\max_{i,j} \frac{\mu_j - \mu_i}{\sqrt{\Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj}}} \\
&= \arg\min_{i,j} -\frac{\sqrt{\Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj}}}{\mu_i - \mu_j} \\
&= \arg\max_{i,j} \frac{\Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj}}{(\mu_i - \mu_j)^2}
\end{aligned}$$

Similarly, assuming $s_j > s_i$ returns the exact same expression.