LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models

Anonymous ACL submission

Abstract

Current developments in large language models (LLMs) have enabled impressive zero-shot capabilities across various natural language tasks. An interesting application of these systems is in the automated assessment of natural language generation (NLG), a highly challenging area with great practical benefit. In this paper, we explore two options for exploiting the emergent abilities of LLMs for zero-shot NLG assessment: absolute score prediction, and comparative assessment which uses relative comparisons between pairs of candidates. Though comparative assessment has not been extensively studied in NLG assessment, we note that humans often find it more intuitive to compare two options rather than scoring each one independently. This work examines comparative assessment from multiple perspectives: performance compared to absolute grading; positional biases in the prompt; and efficient ranking in terms of the number of comparisons. We illustrate that LLM comparative assessment is a simple, general and effective approach for NLG assessment. For moderate-sized open-source LLMs, such as FlanT5 and Llama2-chat, comparative assessment is superior to prompt scoring, and in many cases can achieve performance competitive with state-of-the-art methods. Additionally, we demonstrate that LLMs often exhibit strong positional biases when making pairwise comparisons, and we propose debiasing methods that can further improve performance.

1 Introduction

With the current rapid advances in generative AI, pre-trained models are increasingly utilized in a range of NLP tasks, necessitating reliable evaluations of these models. Human evaluation, where annotators critically assess the quality of the outputs of natural language generation (NLG) systems, has been the gold standard approach (Lita et al., 2005; Belz and Reiter, 2006; Lai and Tetreault, 2018; Fabbri et al., 2021). However, human evaluation has its drawbacks, and is notably labor-intensive, time-consuming, and costly. As such, automating the evaluation process and assessing NLG systems without human intervention is highly desirable.

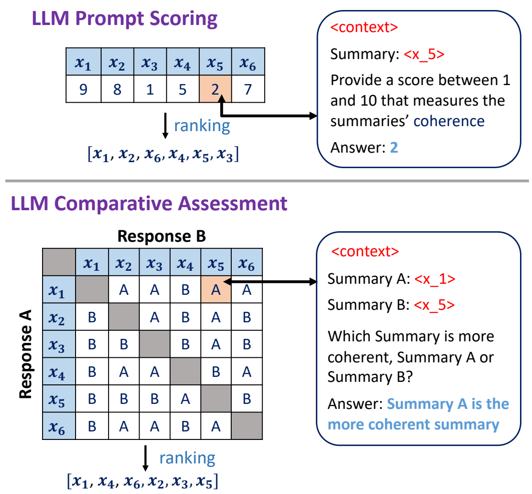


Figure 1: Prompt Scoring v.s. Comparative Assessment. Comparative Assessment prompts an LLM to compare candidates in a pairwise manner, and the comparisons are subsequently converted into scores or ranks.

Though there has been considerable progress in automatic evaluation methods, many proposed approaches have certain restrictions that limit their effectiveness. A large body of existing work use evaluation methods designed for particular tasks and attributes (Mehri and Eskenazi, 2020a; Rei et al., 2020; Manakul et al., 2023b), for example, measuring the consistency of summaries (Wang et al., 2020; Manakul et al., 2023a). Though effective within their domain, these approaches are not extensible to different NLG aspects and cannot be used by practitioners wishing to evaluate systems on inputs or properties that are less common.

The recent development in the emergent abilities of LLMs (Wei et al., 2022) has enabled LLMs to achieve impressive zero-shot performance for a slew of language tasks. This has led to general prompt-based assessment approaches, such as prompt-scoring where an LLM is probed to score outputs on a particular aspect (Wang et al., 2023; Kocmi and Federmann, 2023). These approaches are often only effective with massive LLMs with 175B+ parameters, which may limit the applicability of the approach, especially when access is limited to API access.

With the insight that for humans, it is often easier to select which of two options is better than it is to score options independently, we question whether pairwise comparisons may be more effective at leveraging the impressive emergent ability of LLMs. In this work, we consider LLM comparative assessment, where an LLM is prompted to compare pairs of NLG candidates and predict which one is better. We demonstrate empirically that comparative assessment performs much better than prompt-scoring for FlanT5 and Llama style models, and enables moderate-sized open-source LLMs to achieve near (or above) state-of-the-art performance across a range of NLG language tasks, for a diverse set of attributes. Our approach is general and can be applied to a diverse range of tasks and textual attributes, is simple and requires minimal prompt engineering. Further, we demonstrate that pairwise LLM comparisons often exhibit strong positional biases, where the ordering of candidates impacts the decisions. We introduce a simple debiasing method and empirically illustrate that debiasing can provide further performance improvements, especially when large biases are present.

Our contributions are 1) We are the first work that comprehensively analyzes pairwise comparative assessment for NLG evaluation; 2) We demonstrate that comparative assessment is far more effective than prompt-scoring for moderately-sized LLMs, and yields performance that is state-of-the-art for particular attributes; 3) We demonstrate that positional bias impacts comparative decisions, and introduce a method to debias LLMs which leads to performance boosts, especially when only a subset of comparisons are considered.

2 Background and Related Work

2.1 Reference-based Evaluation

In NLG evaluation, a standard approach is the comparison of annotator-provided gold-standard references with the generated response. Established heuristics, such as the N-gram overlap metrics ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005), have extensively been applied for assessing summarization and machine translation respectively. Recently, the paradigm has evolved to incorporate embedding-based methods like BERTScore (Zhang et al., 2019), which not only compares generated texts with references, but also factors in semantic considerations beyond word overlap.

2.2 Tailored NLG Evaluation Approaches

Tailored approaches have been proposed for assessing specific properties of generated texts. For example, question-answering systems are used for summary consistency assessment (Wang et al., 2020; Scialom et al., 2021) to probe information consistency. For Dialogue quality assessment, the language model probability from a DiaoloGPT system is used as a proxy for response quality (Mehri and Eskenazi, 2020b). A survey for NLG evaluation methods was conducted by Celikyilmaz et al. (2020).

2.3 Zero-shot LLM Evaluation

Given the current capabilities of LLMs such as ChatGPT and GPT4, the zero-shot ability of these systems for a wide range of tasks, including NLG evaluation, has been investigated. Existing works have looked at using LLM to evaluate open-ended story generation and adversarial attacks (Chiang and Lee, 2023) and using ChatGPT to score the quality of texts along a certain axis (Wang et al., 2023; Kocmi and Federmann, 2023), demonstrating that ChatGPT can be used in a zero-shot setting and achieve reasonable performance.

2.4 LLM Pairwise Comparisons

Pairwise comparative judgement (Thurstone, 1927) has been a popular approach of assessing candidates for exams, however where typically human assessors are used. Investigating the ability and application of pairwise comparisons via LLMs has been relatively underexplored, with concurrent work using pairwise rankings for information text retrieval (Qin et al., 2023) and separately for assessing LLM-based chat assistants on open-ended questions where outputs are compared to that of a baseline system (Chiang et al., 2023; Zheng et al., 2023).

3 Comparative Assessment

3.1 Notation

In this work, we investigate using LLM comparative judgements for NLG assessment. Assume that there is a context d (e.g., a text passage or dialogue) and a set of N candidate responses, $x_{1:N}$. For a given attribute (e.g., coherence, consistency, fluency) the N candidates have true underlying scores, $s_{1:N}$. As scores often only have relative meaning, in this work only the ranks of the candidates will be evaluated. The objective is therefore to accurately predict the true ranks, $r_{1:N}$, of the candidate scores. In comparative assessment, one uses pairwise comparisons to determine which of the two input responses is better. Let $y_{ij} \in \{0, 1\}$ represent the true outcome of whether x_i is higher ranked than x_j, such that $y_{ij} = \mathbb{1}(s_i > s_j)$. Here, an LLM is used to model the probability that response i is better than response j, p_{ij},

$$p_{ij} = P(y_{ij}|x_i, x_j, d) \tag{1}$$

Which can alternatively be converted into hard decisions, $\hat{y}_{ij}$, by selecting the most likely outcome.

$$\hat{y}_{ij} = \begin{cases} 1, & \text{if } p_{ij} > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Let $\mathcal{C} = \{c_k\}_{k=1...R}$ represent a set of comparisons, where R is the total number of comparisons, and each comparison $c = (i, j)$ indicates the indices of the two considered candidate responses. For example, the set of all possible comparisons, $\mathcal{C} = \{(i, j) \mid i, j \in [1...N], i \neq j\}$, could be used, or alternatively a smaller subset of comparisons.

3.2 Prompt Design

To leverage the emergent ability of LLMs, we use comparative prompts that probe a model to decide which of the two candidates is better. Let T be a prompt template that converts candidate responses x_i and x_j as well as context d into an output text, prompt $\mathcal{P} = T(x_i, x_j, d)$. This work aims to find a simple, general and robust assessment method, and as such extensive prompt engineering is not in the scope of this work (despite possible performance gains). We evaluate two simple and suitable prompts in our initial investigations. Our prompts for comparative assessment are shown in Figure 2.

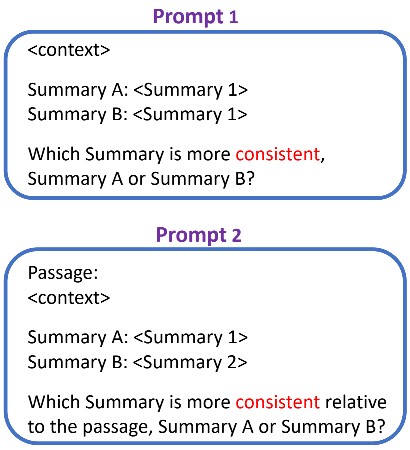


Figure 2: Comparative prompt template 1 and 2. When assessing different attributes, only the attribute is changed (e.g., consistent → engaging) and for response assessment, the word 'summary' is replaced with 'response'.

3.3 Comparative Decisions

A central aspect of LLM comparative assessment is the methodology of getting comparative decisions. In this section, we consider two approaches for leveraging LLMs for comparative assessment; First for when one has output token-level probabilities (Prompt-Based Classifier), and second for when only the output texts are available.

Prompt-Based Classifier: If one has access to the output probabilities, an efficient method to get probability estimates of the predictions is to leverage prompt-based classifiers. Let $P_\theta(w|x)$ represent an LLM's conditional language model distribution of the output sequence w given the input text x. For prompt-based classifiers, the LM probabilities of specific label words (w_k) are used as a proxy for the class decisions (Liusie et al., 2023). For example in summarization assessment, given a prompt $\mathcal{P}$ ending in '... which summary is better', one can set w_i='*Summary A*' and w_j='*Summary B*' and define the probability that response i is better than response j as:

$$p_{ij} = \frac{P_\theta(w_i|\mathcal{P})}{P_\theta(w_i|\mathcal{P}) + P_\theta(w_j|\mathcal{P})} \tag{3}$$

Text Generation: Alternately, if only limited API access is available, one can sample responses from the conditional LM given the input prompt $\mathcal{P}$,

$$\tilde{w}^{(k)} \sim P_\theta(w|\mathcal{P}) \tag{4}$$

Let $f(\tilde{w}) \in \{0, 1\}$ be a function that maps the text response to the comparative decision. By generating K samples from the LLM, one can estimate the comparative probability p_{ij} by looking at the fraction of the samples that selects x_i over x_j.

$$p_{ij} = \frac{1}{K}\sum_{k=1}^{K} f(\tilde{w}^{(k)}) \tag{5}$$

3.4 Comparisons to Ranks

Although the full set of possible comparisons yields the most information for the rankings, this requires $R{=}N(N{-}1)$ comparisons, which can be computationally expensive. For computational efficiency, we can consider 3 different comparison selection strategies: random, no-repeat and symmetric. For **random**, comparisons are randomly selected from the set of all possible comparisons. For **no-repeat**, if (x_i, x_j) is selected then (x_j, x_i) will not be selected. For **symmetric**, if (x_i, x_j) is selected, then (x_j, x_i) will also be selected.

Given a set of selected comparisons $\mathcal{C}$ and weights of a comparative assessment system $\boldsymbol{\theta}$, one can generate a predicted rank ordering $\hat{r}_{1:N}$ of the candidate responses. A simple but effective approach is to sort the candidates by **win-loss** ratio,

$$\hat{s}_i = \frac{\text{\#wins of } x_i}{\text{\#comparisons involving } x_i} \tag{6}$$

which can then be ordered to convert the scores into predicted ranks $\hat{r}_{1:N}$.

3.5 Debiased Comparative Assessment

Let $\tilde{y}_{ij}$ represent the outcome of the comparison when considered in the opposite ordering, such that $\tilde{y}_{ij} = 1 - \hat{y}_{ji}$. For a positionally unbiased comparator, reversing the ordering should have no impact on the outcome of the comparison

$$\tilde{y}_{ij} = \hat{y}_{ij} \qquad \forall\, (i,j) \in [1...N], i \neq j \tag{7}$$

Systems may, however, have systematic positional biases and could for example favor the first position over the second position. To quantify the level of systematic bias, one can determine $P(A)$, the prior associated with the first position, and $P(B)$ the prior for the second position. This can be estimated for a given set of comparisons by using the statistics over all comparisons, and by calculating the fraction of times that each position is selected.

$$P(A) = \frac{\sum_{i,j\in\mathcal{C}} \hat{y}_{ij}}{|\mathcal{C}|} \quad P(B) = \frac{\sum_{i,j\in\mathcal{C}} \tilde{y}_{ij}}{|\mathcal{C}|} \tag{8}$$

When using a symmetric comparative set $\mathcal{C}$, for an unbiased system, both $P(A)$ and $P(B)$ should be 0.5 and any large deviation is symptomatic of positional bias. To address possible positional bias, one may reweight system probabilities, $\hat{p}_{ij}$, through

$$\hat{p}_{ij} = \frac{\alpha \cdot p_{ij}}{\alpha \cdot p_{ij} + (1 - p_{ij})} \tag{9}$$

where $\alpha \in \mathbb{R}^+$ is a weight that can be set such that $P(A) = P(B) = 0.5$. Reweighting in this fashion is equivalent to,

$$\hat{y}_{ij} = \begin{cases} 1, & \text{if } p_{ij} > \tau \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where $\tau \in [0, 1]$ is a decision threshold corresponding to α, set such that $P(A) = P(B) = 0.5$.

4 Experimental Setup

4.1 Datasets

To investigate the general applicability of comparative assessment, we cover a range of standard NLG evaluation tasks and datasets as follows:

SummEval (Fabbri et al., 2021) is a summary evaluation benchmark of 100 passages, each with 16 machine-generated summaries. Each summary is evaluated for coherency (`COH`), consistency (`CON`), fluency (`FLU`), and relevancy (`REL`).

Podcast (Manakul and Gales, 2022) is for benchmarking podcast summary assessment methods. It contains 179 podcasts each with 15 abstractive summaries. Each summary was evaluated for its overall quality on a 4-point scale.

TopicalChat with the USR annotations (Mehri and Eskenazi, 2020b) is for benchmarking dialogue evaluation. It includes 60 dialogue contexts and six system responses per context. These responses were assessed on coherency (`COH`), continuity (`CNT`), engagingness (`ENG`), and naturalness (`NAT`).

WebNLG (Gardent et al., 2017) is for benchmarking data-to-text evaluation methods. It contains 223 semantic triple groups, each paired with outputs from 8 triple-to-text generation systems. These texts were evaluated for fluency (`FLU`), grammar (`GRA`) and semantic equivalence (`SEM`).

4.2 Base Large Language Models (LLMs)

We investigate two families of open-source instruction-tuned LLMs. The first system is FlanT5 (Chung et al., 2022), T5 (Raffel et al., 2020) that have been instruction tuned on a diverse set of 1000 NLP tasks (Wang et al., 2022). The second system is Llama2-chat (Touvron et al., 2023), which is Llama2 tuned on instruction datasets. We investigate a range of model sizes; 220M, 770M, 3B and 11B for FlanT5, and 3B and 13B for Llama2.

4.3 Baselines

The NLG evaluation methods can be categorized into *reference-based* and *reference-free*. Reference-based methods compare the output against the reference such as n-gram metrics (e.g., BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004)), or embedding based metrics (e.g., BERTScore (Zhang et al., 2019)). In contrast, reference-free methods compare the generated texts against the original source (or context for generation) directly.

4.3.1 Bespoke Methods

Bespoke methods require a specific data which could be supervised labels (e.g., human judgements for the summaries) or data for model training (e.g., question-answering). Although bespoke methods could work in a similar domain (e.g., developed for summarization, but applied on dialogue generation), they are not as general as zero-shot methods.

UniEval (Zhong et al., 2022) convert NLG evaluation into Boolean QA problem. This method uses pre-defined schemes for selected aspects (e.g., coherence) and generates synthetic data to fine-tune a T5 system for NLG assessment. References are used for particular aspects (e.g. relevancy), and schemes/systems are bespoke for a particular attribute (though a sequentially trained system that scores multiple attributes is also explored).

QuestEval (Scialom et al., 2021) and **MQAG** (Manakul et al., 2023a) are QA-based approaches for assessing *consistency* in summarization tasks. QuestEval uses extracted answer spans while MQAG represents information using multiple-choice questions. Both methods are reference-free.

Longformer-SFT: For podcast summarization, we follow Manakul and Gales (2022) in using a Supervised Fine-Tuned longformer (Beltagy et al., 2020) as a baseline. The input is the document and the summary, and human judgement is used as the supervised target label at training, and the performance is reported using 5-fold cross-validation.

4.3.2 Zero-shot Methods

Zero-shot methods can be applied generally to any task without further training or fine-tuning. Comparative assessment is a zero-shot method.

GPTScore (Fu et al., 2023) evaluates texts using conditional language model scores. By conditioning the language model on instruction and context, GPTScore assumes that it will assign a higher probability to a high-quality generated text.

Prompt Scoring. Another baseline is prompt-scoring. With this approach, for a particular attribute, the LLMs is asked to assess the response quality between 1-10. Simple prompts are used with the general templates shown in Figure 3. Prompt-scoring is run for all open-source LLMs considered (FlanT5 and Llama2), and is used as the main baseline to compare comparative assessment against. During generation, the maximum generation length is set to 5 and the temperature is set to 1.0. Similarly, ChatGPT prompt-scoring has recently been proposed in Wang et al. (2023); Kocmi and Federmann (2023), which we also include as a baseline where applicable.

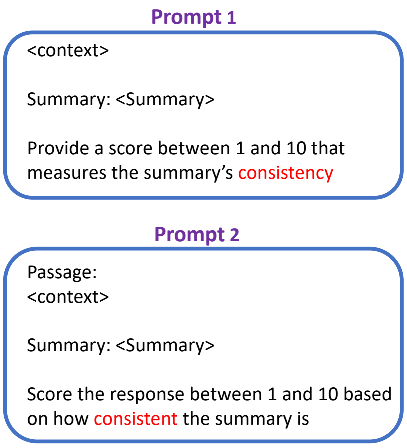


Figure 3: Scoring template 1 and template 2. Only the attribute is changed (e.g., consistent → engaging) and response description ('summary'→ 'response') for different tasks.

4.4 Methodology

Each LLM is used for both prompt-scoring and comparative assessment. For the main comparative assessment results, we consider the full set of possible comparisons, where all pairs of candidates in both permutations are compared by the framework. Comparisons are made using the prompt-based classifier (as described in §3.3) using the prompt templates shown in Fig. 2, where the system outputs a probability for Response A and Response B. The winner of the comparison is the response with the highest probability, where candidates are then ranked in order of the win-ratio (as described in §3.4). For Llama2, comparative prompts are appended with '`Answer:`' while scoring prompts end with 'Score:'. The spearman correlation between predicted scores and human judgements is used as the performance metric.

5 Experiments

5.1 NLG Evaluation Results

Summary Assessment: Table 1 analyzes the effectiveness of comparative assessment on **SummEval**, where the following observations can be made:

(1) Moderate-sized LLMs are ineffective in the prompt-scoring set-up, with the best system (FlanT5-3B) achieving spearman correlations of 10-20. The performance difference with ChatGPT prompt-scoring implies that scoring is likely an emergent ability only effective for larger LLMs.

(2) LLMs are able to achieve considerably higher correlations in the comparative assessment set-up, with performance higher for nearly all systems. Further, comparative assessment leads to more robust performance, with most 3B+ models achieving correlations within the range of 30-50.

(3) Comparative assessment enables LLMs of under 1B to perform well, with FlanT5-770M achieving moderate correlations. However, performance improves significantly when using 3B+ LLMs, although for SummEval there are diminishing (if any) performance gains by scaling up.

(4) The best comparative assessment LLM (FlanT5-3B) is competitive with all other zero-shot methods, including ChatGPT scoring (an LLM with two orders of magnitude more parameters), and achieves the best correlation in 3 of the 4 aspects.

(5) Comparative assessment achieves competitive performance with UniEval. Although UniEval has better overall performance, UniEval was designed for bespoke tasks and aspects (it is fine-tuned on synthetic data created for particular attributes) where the results in Tables 2 and 4 show that UniEval has noticeable degradation in out-of-domain settings. In contrast, comparative assessment is zero-shot and general.

Approach	COH	CON	FLU	REL
Baselines (§4.3)				
BERTScore (w/ Ref)	25.9	19.7	23.7	34.7
QuestEval	18.2	30.6	22.8	26.8
MQAG	17.0	28.8	19.3	16.6
UniEval (single-best)	54.6	47.2	43.3	46.3
UniEval (continual)	57.5	44.6	44.9	42.6
GPTScore FlanT5-3B	47.0	43.6	42.1	34.4
GPTScore FlanT5-11B	45.6	43.8	42.4	34.3
GPTScore GPT3	40.1	47.5	41.0	34.3
ChatGPT scoring†	45.1	43.2	38.0	43.9
Prompt Scoring (§4.3.2)				
FlanT5-220M	4.0	-0.2	0.2	2.8
FlanT5-770M	-3.6	-1.6	-1.5	-0.0
FlanT5-3B	**14.5**	**19.8**	**3.9**	**15.2**
FlanT5-11B	0.7	11.2	3.2	5.7
Llama2-chat-7B	8.6	9.0	1.8	7.8
Llama2-chat-13B	9.9	6.9	1.2	9.2
Comparative Assessment (§3)				
FlanT5-220M	4.0	-0.2	0.2	2.8
FlanT5-770M	29.8	26.3	20.6	35.1
FlanT5-3B	**51.2**	**47.1**	**32.5**	44.8
FlanT5-11B	44.2	37.2	30.2	43.4
Llama2-chat-7B	27.9	24.6	20.2	35.6
Llama2-chat-13B	40.9	39.9	30.8	**45.3**

Table 1: Spearman correlation coefficient for **SummEval**, averaged over both prompts per system (for prompt-scoring and comparative). †ChatGPT performance is quoted from Wang et al. (2023), which use more detailed scoring prompts.

Approach	System-lvl	Summary-lvl
Baselines (§4.3)		
BERTScore (w/ Ref)	73.9	25.1
UniEval (continual)	42.0	22.8
QuestEval	42.5	20.4
MQAG	77.9	12.6
Longformer-SFT	89.6	19.6
Prompt Scoring (§4.3.2)		
Llama2-chat-7B	88.5	2.6
Llama2-chat-13B	80.0	25.3
Comparative Assessment (§3)		
Llama2-chat-7B	88.2	37.4
Llama2-chat-13B	97.1	45.5

Table 2: Spearman correlation coefficient for **Podcast**.

Podcast Assessment: When considering podcast summarization with long inputs of over 5k tokens on average, only Llama2 models (which have a limit of 4k tokens) were used (as FlanT5 has a limit of 1k tokens). Table 2 shows that comparative assessment yields highly impressive performance for long-spoken summarization, with comparative assessment out-competing all other baselines. Further, although prompt-scoring has good system-level correlations, the lack of granularity leads to poor summary-level performance.

Dialogue Assessment: Next, we analyze comparative assessment on TopicalChat, for evaluating conversational responses. Table 3 shows similar findings for TopicalChat as to those in SummEval, where comparative assessment again outperforms the correlations seen from prompt-scoring.

Data-to-Text Assessment: For data-to-text generation, the context is highly abstract and is a list of triples in the form of (object, relation, subject). This makes assessing the semantics challenging, as the LLM needs to parse and understand semantic triples. Table 4 shows that understanding triples is

Approach	COH	CNT	ENG	NAT
Baselines (§4.3)				
UniEval (single-best)	60.7	-	59.6	54.7
UniEval (continual)	61.3	-	60.5	44.4
GPTScore GPT3	56.9	32.9	49.6	52.4
ChatGPT scoring[†]	54.7	57.7	37.9	58.0
Prompt Scoring (§4.3.2)				
FlanT5-220M	-2.2	0.2	-8.4	2.1
FlanT5-770M	3.7	3.1	-4.3	3.8
FlanT5-3B	**31.9**	**28.8**	17.4	**23.7**
FlanT5-11B	15.3	8.0	4.3	24.3
Llama2-chat-7B	16.4	17.0	20.6	21.4
Llama2-chat-13B	21.7	19.9	**31.4**	23.2
Comparative Assessment (§3)				
FlanT5-220M	-0.3	8.2	-10.5	2.2
FlanT5-770M	38.5	36.3	25.3	35.3
FlanT5-3B	49.4	**49.4**	37.3	47.4
FlanT5-11B	**54.3**	42.2	54.7	**54.2**
Llama2-chat-7B	28.9	33.7	36.1	30.3
Llama2-chat-13B	32.4	43.2	**55.5**	33.5

Table 3: Spearman correlation coefficient for **TopicalChat**. [†]ChatGPT is prompted using our prompt-scoring prompts.

Approach	FLU	GRA	SEM
Baselines (§4.3)			
BLEU	36.3	34.7	50.3
METEOR	44.3	42.9	62.7
NLI Model[*]	-	-	63.7
UniEval (continual)	21.7	16.3	-
Prompt Scoring (§4.3.2)			
FlanT5-220M	18.5	17.4	8.0
FlanT5-770M	14.5	13.6	17.1
FlanT5-3B	**30.8**	**32.7**	**38.5**
FlanT5-11B	-0.7	6.9	20.8
Llama2-chat-7B	3.8	2.4	17.0
Llama2-chat-13B	1.8	0.5	5.6
Comparative Assessment (§3)			
FlanT5-220M	-13.6	-17.9	0.1
FlanT5-770M	36.2	35.2	11.4
FlanT5-3B	40.6	41.4	12.8
FlanT5-11B	41.4	44.8	52.4
Llama2-chat-7B	22.9	37.8	-5.3
Llama2-chat-13B	**44.9**	**45.1**	**53.5**

Table 4: Spearman correlation coefficient for **WebNLG**. [*]Quoted from the NLI method with the backoff template in Dušek and Kasner (2020).

an emergent ability of LLMs, where for grammar and fluency the correlations are quite similar between the 3B and 11B/13B systems, however for semantic understanding, the 10B+ systems highly outcompete the 3B+ systems. Note that when evaluating UniEval, we used the closest attribute that they designed for, which was naturalness for both.

5.2 Positional Bias

We investigate whether the comparative prompts have any implicit positional bias, and whether systems prefer the first/second position. Table 5 shows the fraction of comparisons that selected the candidate in the first position for SummEval. Since all comparisons in both permutations are considered, this fraction should be 0.50 for an unbiased system. However, we observe considerably high bias, with some set-ups even selecting the first option 80% of the time. Further, we observe that larger systems appear to be more susceptible to bias than smaller systems, which may explain the similarity in performance for the 3B and 11B/13B systems in the previous main results. Similar results for other datasets are provided in Appendix A.2

System	Prompt	COH	CON	FLU	REL
FlanT5 3B	1	0.37	0.46	0.39	0.41
	2	0.43	0.47	0.40	0.44
FlanT5 7B	1	0.18	0.20	0.13	0.23
	2	0.24	0.24	0.17	0.26
Llama2-chat 7B	1	0.41	0.17	0.26	0.18
	2	0.68	0.56	0.48	0.45
Llama2-chat 13B	1	0.31	0.37	0.18	0.32
	2	0.29	0.30	0.19	0.26

Table 5: Positional bias $P(A)$ for both prompt templates, for various systems in the comparative setup on SummEval.

5.3 Debiasing

The previous section demonstrates that comparative assessment exhibits positional bias which may impact system decisions. We therefore investigate whether debiasing can improve evaluation performance. Table 6 shows standard and debiased LLM comparative assessment performance for the considered tasks and scores, with WebNLG SEM and Podcast omitted due to the required emergent ability and large context length respectively. We observe that debiasing can lead to performance boosts, where we note that the prompts which have a high bias (seen in Table 5 and Table 8 in the appendix) benefit most from debiasing. In particular, for TopicalChat we observe large gains for the FlanT5-11B system, which enables state-of-the-art performance. To explain why debiasing can lead to large performance boosts, consider a very biased system where the first response is always selected as better. Although over both permutations the system is unbiased for any comparison, the bias in the system will cause the system to assume that all candidates are of the same quality. By reducing the bias of each comparison, the system may be able to pick up subtler quality differences between the samples.

System	Debias	SummEval COH	SummEval CON	SummEval FLU	SummEval REL	TopicalChat COH	TopicalChat CNT	TopicalChat ENG	TopicalChat NAT	WebNLG FLU	WebNLG GRA	Avg.
FlanT5-3B	✗	51.2	47.1	32.5	44.8	49.4	49.4	37.3	47.4	41.0	41.8	44.2
	✓	51.8	46.9	33.0	45.3	49.6	50.2	38.0	46.3	40.7	42.3	44.4
FlanT5-11B	✗	44.2	37.2	30.2	43.4	54.3	42.2	54.7	54.2	41.4	44.8	44.7
	✓	45.3	39.7	30.7	44.7	57.2	59.5	59.5	58.8	44.5	44.6	48.5
Llama2-chat-7B	✗	29.4	24.6	19.7	35.2	28.2	33.1	36.3	28.7	22.9	37.8	29.6
	✓	28.8	24.8	19.7	35.5	29.1	34.5	39.7	28.5	24.3	37.1	30.2
Llama2-chat-13B	✗	40.9	39.9	30.8	45.3	32.4	43.2	55.5	33.5	44.9	45.1	41.2
	✓	42.8	40.3	31.9	47.1	32.5	44.5	56.9	38.4	45.9	43.7	42.4

Table 6: Spearman correlation coefficient on different aspects of the NLG evaluation tasks, averaged over all prompts considered, using all pairs and ordering considered (i.e. full matrix comparisons).

System	Debias	COH	CON	FLU	REL
FlanT5-3B	✗	68.6	82.0	68.2	67.2
	✓	69.8	82.1	68.8	67.8
FlanT5-11B	✗	61.6	70.3	60.3	63.3
	✓	66.2	76.7	65.9	67.4
Llama2-chat-7B	✗	59.6	63.8	59.6	61.0
	✓	60.3	65.7	60.4	63.1
Llama2-chat-13B	✗	62.6	75.4	61.1	65.4
	✓	65.8	76.9	67.2	68.5

Table 7: Accuracy of the comparative systems, at a comparison level, for SummEval.

5.4 Comparative Accuracy

One can also measure the accuracy of the comparative system at a comparison level. Table 7 shows the pairwise comparison accuracy for Summeval, over all candidate pairs where the true score of the candidate response varies. We observe accuracies between 60-80% across all tasks and observe that debiasing can substantially increase accuracy. This highlights that LLMs are able to compare the quality of responses fairly well, though the moderately sized LLMs may not always select the best response (with respect to labels).

5.5 Subset of Comparisons

Due to $O(N^2)$ number of comparisons required for the full comparison matrix, it might be practical to only consider a subset of comparisons. Figure 4 shows the downstream Spearman correlation for SummEval coherency, when averaged over 50 runs, for different comparison selection strategies. Of the three schemes, we observe that for small R (i.e. less than half the total number of comparisons) selecting comparisons with no repeats leads to a marginal improvement over random selection. Further, by using the symmetric selection scheme, despite the number of comparisons being half that of no-repeat (although each comparison is done twice, once in each permutation), interestingly there is only a performance difference of 1 in terms of Spearman. Finally, we observe that debiasing can be very effective in efficient set-ups, and leads to larger benefits when the number of comparisons is small. Equivalent plots for other tasks/scores can be found in Appendix A.1.

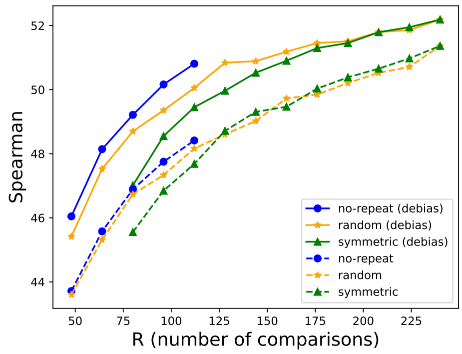


Figure 4: FlanT5-3B performance for SummEval COH when a subset of the comparisons are selected by either random, no-repeat or symmetric (as described in §3.4). For no-repeat, each pair is compared once, hence has a smaller maximum R.

6 Conclusions

This paper investigates LLM comparative assessment, a simple zero-shot approach to NLG evaluation. We demonstrate that for moderately sized LLMs, comparative assessment outperforms absolute scoring, and is an effective automatic assessment, achieving near state-of-the-art performance for a range of NLG evaluation tasks. Furthermore, we show that LLMs are prone to have positional bias that could impact their decisions, however, we introduce a simple debiasing approach that leads to performance boosts, especially for biased systems.

Limitations

Computational Cost. The comparative assessment framework with the full set of comparisons uses $N \cdot (N-1)$ comparisons, which for large N can be computationally prohibitive. This paper investigated datasets with at most 16 candidates, and may not scale when more candidates are required.

Base LLMs. The empirical findings are for LLMs of up to 13B parameters. By using larger models (with 100B+ parameters) one may expect further performance improvements. However, due to API costs and the O(N^2) number of comparisons, results are limited to open-source LLMs.

Selection of the subset of comparisons. For our comparison selection scheme, this work only considered static selection schemes. Future work may investigate dynamic selection schemes, either by considering sorting algorithms or ELO competition schemes, and methods similar to those studied in information retrieval by Qin et al. (2023).

Ethics Statement

For some tasks/datasets, comparative assessment could be ineffective and have poor generalisation over the task. Deploying machine learning classifiers in real-world classification settings has many associated risks, and careful analysis should be made before deploying such systems. Misuse/overconfidence in the approach may lead to mistrust of users towards LLM solutions.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

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*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

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

Alice Lai and Joel Tetreault. 2018. Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223, Melbourne, Australia. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lucian Lita, Monica Rogati, and Alon Lavie. 2005. BLANC: Learning evaluation metrics for MT. In

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 740–747, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. 2023. Mitigating word bias in zero-shot prompt-based classifiers.

Potsawee Manakul and Mark JF Gales. 2022. Podcast summary assessment: A resource for evaluating summary assessment methods. *arXiv preprint arXiv:2208.13265*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023a. Mqag: Multiple-choice question answering and generation for assessing information consistency in summarization. *arXiv preprint arXiv:2301.12307*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023b. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Louis L Thurstone. 1927. A law of comparative judgment. *Psychological review*, 34(4):273.

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*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

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

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

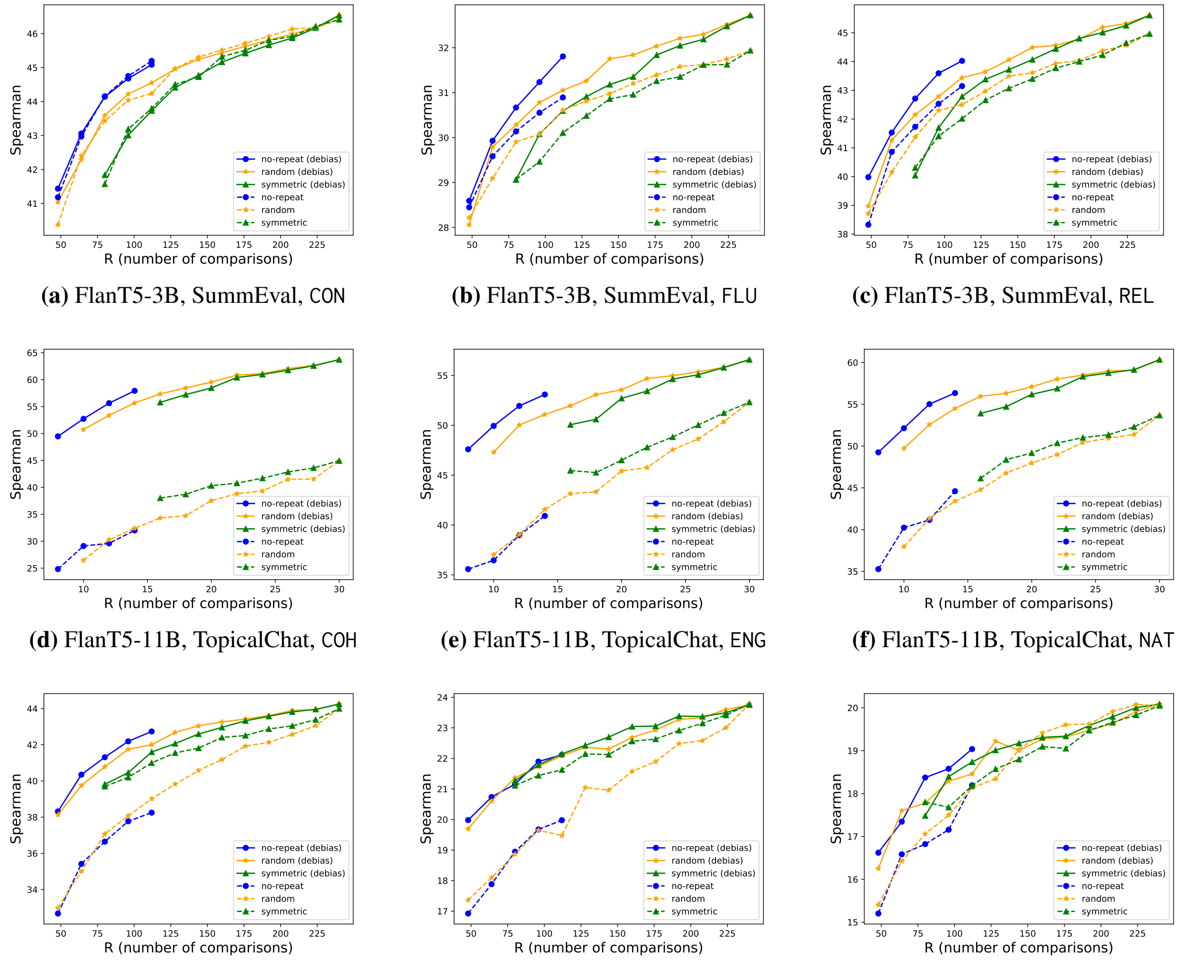
Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Additional Results

A.1 Partial Comparison Curves



(a) FlanT5-3B, SummEval, `CON` **(b)** FlanT5-3B, SummEval, `FLU` **(c)** FlanT5-3B, SummEval, `REL`

(d) FlanT5-11B, TopicalChat, `COH` **(e)** FlanT5-11B, TopicalChat, `ENG` **(f)** FlanT5-11B, TopicalChat, `NAT`

(g) FlanT5-11B, SummEval, `REL` **(h)** Llama-chat-7B, SummEval, `CON` **(i)** Llama-chat-13B, SummEval, `FLU`

Figure 5: Assessment Performance when only a subset of comparisons are considered (extending the results of Figure 4). Multiple different base LLMs, datasets and scores and displayed.

A.2 Positional Bias

System	prompt	SummEval COH	CON	FLU	REL	TopicalChat COH	CNT	ENG	NAT	WebNLG FLU	GRA	SEM	Podcast
FlanT5 3B	1	0.37	0.46	0.41	0.42	0.47	0.44	0.50	0.49	0.46	0.41	0.89	-
	2	0.43	0.47	0.42	0.44	0.46	0.44	0.47	0.47	0.38	0.36	0.85	-
FlanT5 11B	1	0.18	0.25	0.16	0.23	0.25	0.17	0.27	0.26	0.15	0.19	0.56	-
	2	0.24	0.29	0.19	0.26	0.27	0.13	0.29	0.31	0.19	0.21	0.42	-
Llama2-chat 7B	1	0.41	0.21	0.28	0.18	0.57	0.26	0.25	0.36	0.36	0.53	0.98	0.33
	2	0.68	0.57	0.50	0.45	0.56	0.37	0.22	0.35	0.37	0.48	0.90	0.24
Llama2-chat 13B	1	0.31	0.43	0.20	0.32	0.69	0.73	0.67	0.74	0.23	0.38	0.50	0.22
	2	0.29	0.37	0.22	0.26	0.65	0.65	0.62	0.68	0.28	0.40	0.29	0.40

Table 8: Fraction of comparisons where the candidate in the first position was selected by the LLM when using the full (symmetric) set of comparisons. The bias is presented for both prompts, over all datasets and scores, extending the results in Table 5.

A.3 Accuracy of Pairwise Comparisons

System	debias	SummEval				TopicalChat				WebNLG			Podcast
		COH	CON	FLU	REL	COH	CNT	ENG	NAT	FLU	GRA	SEM	
FlanT5	✗	68.6	82.0	68.2	67.2	75.3	71.0	65.6	70.3	66.2	65.5	51.8	-
3B	✓	69.8	82.1	68.8	67.8	75.4	72.2	65.6	69.9	66.7	66.6	51.3	-
FlanT5	✗	61.6	70.3	60.3	63.3	70.0	60.5	68.0	68.9	60.8	62.7	69.6	-
11B	✓	66.2	76.7	65.9	67.4	76.6	74.2	74.4	74.7	67.6	67.3	69.9	-
Llama2-chat	✗	59.6	63.8	59.6	61.0	64.0	62.0	61.0	60.4	56.6	61.1	48.3	63.4
7B	✓	60.3	65.7	60.4	63.1	64.0	64.3	65.9	61.6	57.1	61.1	50.2	-
Llama2-chat	✗	62.6	75.4	61.1	65.4	64.5	66.8	72.0	62.3	64.7	67.6	67.3	70.3
13B	✓	65.8	76.9	67.2	68.5	65.9	69.4	73.8	65.2	66.7	67.4	68.9	-

Table 9: Accuracy of pairwise comparisons of all candidates which differ in true value. Accuracies are shown for all datasets and scores, extending the results of Table 6.

B Alternate Ranking Strategies

In the main paper, we only consider the win ratio as an approach of converting comparisons to ranks, due to win-ratio being simple and intuitive. However alternate ranking strategies are possible; a well-motivated decoding approach is to select the ranks with the highest probability given the observed comparisons. By Bayes' theorem, this is equivalent to finding the ranks that maximise the likelihood of the observations.

$$\hat{r}_{1:N} = \underset{r_{1:N}}{\text{argmax}}\ P(\mathcal{C}|r_{1:N}) \tag{11}$$

For a set of ranks $r_{1:N}$, let $z_{ij} = \mathbb{1}(r_i < r_j) \in \{0, 1\}$, i.e. whether the ranks imply x_i is better than x_j. Given the probability of each comparison, the **likelihood** of the ranks can be defined as

$$P(\mathcal{C}|r_{1:N}) = \prod_{(i,j)\in\mathcal{C}} \left(p_{ij}^{z_{ij}} + (1 - p_{ij})^{1-z_{ij}}\right) \tag{12}$$

If only hard decisions are available (i.e. the probabilities are not), then one can instead approximate the likelihood and find the ranks that maximise the **approximate-likelihood**.

$$P(\mathcal{C}|r_{1:N}) = \prod_{(i,j)\in\mathcal{C}} P(\hat{y}_{ij}|z_{ij}) \tag{13}$$

Since $\hat{y}_{ij} \in \{0, 1\}$ and $z_{ij} \in \{0, 1\}$, there are 4 conditional probabilities $P(\hat{y}_{ij}|z_{ij})$. Setting one probability will set the other 3, which can be estimated with the system's comparative statistics.

B.1 Initial Results

Table 10 presents initial results for FlanT5-3B on Summeval, comparing the maximum likelihood ranking to the win ratio approach. The initial finding was that performance was similar between the two conversion schemes. However, it's worth noting that minimizing the objective function poses intractability challenges, necessitating an approximate greedy search. For the sake of simplicity, our main paper focused on the win-ratio method, while future research may explore more advanced conversion strategies.

	SummEval			
	COH	CON	FLU	REL
win-loss	51.4	46.4	31.9	45.0
likelihood	51.7	46.0	31.5	44.7

Table 10: Spearman correlation when the comparisons are converted using either win-ratio or maximum likelihood, for FlanT5-3B on SummEval.