Large Language Models are Inconsistent and Biased Evaluators

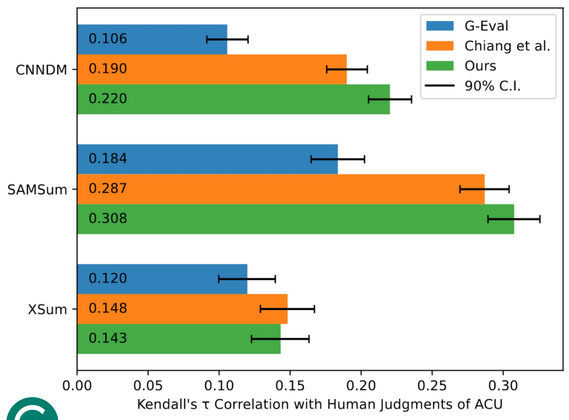
Rickard Stureborg[1,2], Dimitris Alikaniotis[2], Yoshi Suhara[3]
[1] Duke University [2] Grammarly [3] NVIDIA

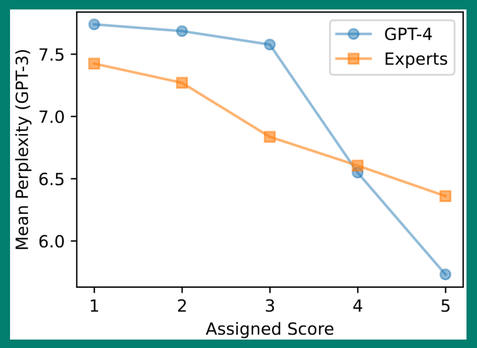
Abstract

The zero-shot capability of Large Language Models (LLMs) has enabled us to use LLMs as reference-free automatic evaluation metrics. Existing studies have shown that LLMs can be high-quality evaluators for various NLP tasks such as summarization. However, little is known about the robustness of LLM evaluators, as the existing work has focused on pursuing the best performance of LLM evaluators with respect to correlations between LLM scores and human expert scores. In this paper, we conduct a series of analysis using the SummEval dataset and report that LLM evaluators for text summarization are sensitive to prompt differences that are trivial to human understanding of text quality. This includes the rating scale itself, scores assigned to previous dimensions of analysis biasing future scores in the same generation, bias towards lower-perplexity summaries, and reliance on features that are uncorrelated the true summary quality (such as worsened performance on rating Fluency of a summary when the source document is not included). We share recipes for how we should configure LLM evaluators while clarifying the limitations, resulting in significantly better performance than G-Eval-style evaluation on the SAMSum partition of the RoSE dataset.

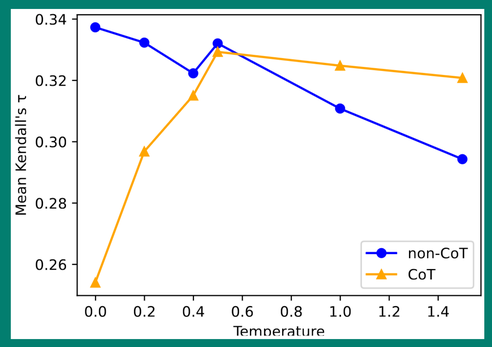
To the right we present some trends we find through our extensive analyses.

Implementing simple fixes to the inconsistencies and biases we found in this work already yields significant results. We improve significantly upon the SOTA in the SAMSum partition of the RoSE dataset. This dataset is used due to its high data quality

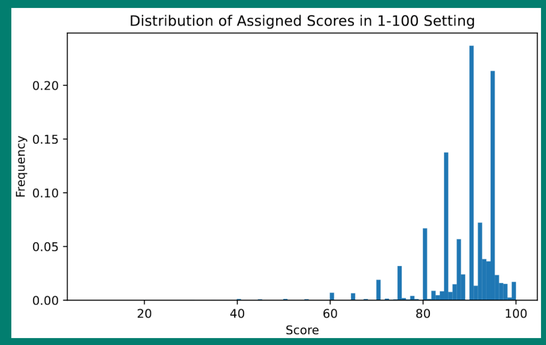




Average Perplexity associated with Automatic Evaluation Score.
Summaries are grouped by evaluation scores (as assigned either by Experts or by GPT-4), perplexities are computed with GPT-3 (text-davinci-003) on the summary text. GPT-4 is disproportionately biased towards low perplexity summaries as compared with expert annotators. This may, at least in part, explain some previous works' observations that models are biased to prefer their own generation



Performance of CoT and non-CoT prompting at varying Temperatures. Each prediction is computed by the average of 10 generations. Low temperatures are beneficial when making simple predictions, but higher temperatures (to a point) help improve performance when using Chain-of-Thought (CoT) prompting. This could be because of a more diverse set of explanations, leading to more unique features for prediction.



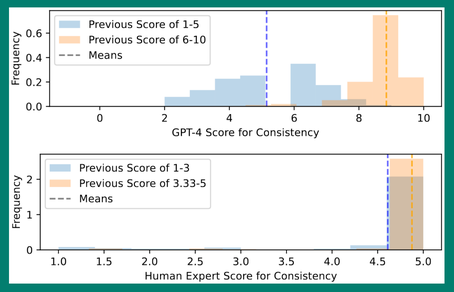
Frequencies of each possible score as found in 64,000 predictions using the 1-100 scale. Models sparsely predict scores within the range. Frequencies of some scores, such as 90 and 95, are far higher than 'odd' scores such as 92 or 19, and much of the range is almost entirely ignored (1-60). Interestingly, 1-60 is a range often largely ignored in academic grading scales. This indicates an issue within instruction-following specific to automatic evaluation.

Correlation with human judgement for GPT-4 by method for increased granularity.
"*G*" is the effective granularity (number of unique scores) possible within the given scale. Methods denoted "avg" are a 10-sample average run with temperature 1.0, while all other methods benefited from reducing temperature to 0. It seems that increasing granularity generally helps low-granularity methods, while high-granularity methods are harmed by increasing granularity. This may be due to the increase in temperature setting. Our results indicate that there may be diminishing returns of increasing scoring granularity.

Method	G	Coh	Con	Flu	Rel	Avg
1-5 star	5	.332	.362	.325	.337	.339
1-5 avg	41	.422	.370	.356	.439	.397
5 +word mod.	13	.361	.408	.345	.363	.369
5 +word (avg)	121	.394	.364	.316	.419	.373
5 +float mod.	13	.425	**.453**	**.380**	.395	.413
5 +float (avg)	121	.416	.378	.334	.438	.392
1-10 score	10	.450	.433	.366	**.462**	**.428**
1-10 avg	91	.424	.366	.332	.435	.389
1-100 score	100	**.463**	.423	.308	.339	.383
1-100 avg	991	.406	.351	.343	.414	.379

Performance of GPT-3.5-Turbo with and without Source Document.
Removing the source document (unsurprisingly) substantially reduces the performance of the automatic evaluator. However, this is also true for attributes that should not be dependent on the source document in the first place, such as Fluency. For categories such as relevance, making a prediction on the summary quality without the article should be impossible

Source Doc	Coh	Con	Flu	Rel	Avg
Included	**.346**	**.250**	**.237**	**.330**	**.291**
Excluded	.291	.167	.212	.183	.213
Δ	-.055	-.083	-.025	-.147	-.078
%Δ	-15.9	-33.2	-10.6	-44.6	-26.7



(Top) Score distribution for consistency, conditioned on the previously assigned score for coherence when predicting both within the same context. (Bottom) Human-determined scores for consistency conditioned on what range the score fell into for coherence.
Yet again we note the LLM evaluator does not make use of the full range of the scores, with no predictions of 5/10 for consistency in this experiment. Human scores are correlated by Pearson's r = 0.315, while GPT-4 scores are correlated by r = 0.979. The above figures clearly show how previous scores bias the distribution of future scores in the generation. While such biasing is natural (and in part valid), the effect here is so large it harms performance.

grammarly