ContrastScore: Towards Higher Quality, Less Biased, More Efficient Evaluation Metrics with Contrastive Evaluation

Xiao Wang[1], Daniil Larionov[2,5], Siwei Wu[1], Yiqi Liu[1],
Steffen Eger[3,5], Nafise Sadat Moosavi[4], Chenghua Lin[1]
[1]The University of Manchester, [2]University of Mannheim,
[3]University of Technology Nuremberg, [4]The University of Sheffield, [5]NLLG
{xiao.wang-26,siwei.wu-2,yiqi.liu-6}@postgrad.manchester.ac.uk
daniil.larionov@uni-mannheim.de, steffen.eger@utn.de
n.s.moosavi@sheffield.ac.uk, chenghua.lin@manchester.ac.uk

Abstract

Recent advances in automatic evaluation of natural language generation have increasingly relied on large language models as general-purpose metrics. While effective, these approaches often require high-capacity models, which introduce substantial computational costs, and remain susceptible to known evaluation pathologies, such as over-reliance on likelihood. We introduce ContrastScore, a contrastive evaluation paradigm that builds on the widely used BARTScore formulation by comparing token-level probabilities between a stronger and a weaker model. Instead of relying on single-model likelihoods or prompt-based judgments, ContrastScore captures disagreement between models to better reflect confidence and uncertainty in generation quality. Empirical results on summarization and machine translation benchmarks show that ContrastScore, instantiated with paired moderate-scale models across both Qwen and LLaMA families, consistently outperforms larger alternatives, such as Qwen 7B and LLaMA 8B, in correlation with human ratings. In addition to improving evaluation quality, ContrastScore significantly reduces susceptibility to likelihood bias, offering a more robust and cost-effective alternative to larger LLM-based evaluation methods. [1]

1 Introduction

Evaluating the quality of automatically generated text remains a fundamental challenge in natural language processing (NLP) and, in some cases, is nearly as difficult as generating the text itself. Traditional evaluation methods primarily rely on reference-based metrics such as BLEU, ROUGE, and METEOR (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005; Shen et al., 2023; Pan et al., 2024; Kalyan, 2024), which are inadequate as they compare generated text to human-written references based on surface-level lexical overlap, often failing to capture semantic adequacy and fluency. Embedding-based metrics like BERTScore (Zhang et al., 2020) improve upon lexical approaches by leveraging contextualized representations for better semantic alignment. However, they remain sensitive to domain shifts, depend on high-quality references, and exhibit relatively weak correlations with human judgments (Zhao et al., 2023). To address these limitations, recent research has shifted towards source-based evaluation methods, leveraging large language models (LLMs). Metrics such as BARTScore (Yuan et al., 2021) and GPTScore (Fu et al., 2024) assess generation quality using a single model's probability distribution, while prompt-based LLM evaluators process structured prompts to evaluate text based on predefined criteria (Que et al., 2024) or conduct comparative judgments without an explicit rubric (Liusie et al., 2024). Although these methods show promise, their effectiveness is inherently constrained by the underlying model's capacity. Larger models generally perform better but are computationally expensive and incur high API costs (Larionov and Eger, 2024). Conversely, smaller models often have reduced capacity, resulting in unreliable evaluations. Furthermore, existing LLM-based metrics are susceptible to biases (Deutsch et al., 2022; Sun et al., 2022; Lu et al., 2023; Liu et al., 2024; Hong et al., 2025), including over-reliance on likelihood, which may not align with human evaluation criteria. These challenges highlight the need for novel evaluation methods that are both efficient and reliable, balancing model capacity, accessibility, and evaluation robustness (Chen and Eger, 2023; Zhao et al., 2024).

Introduced by Li et al. (2023), Contrastive Decoding enhances both diversity and factual accuracy by leveraging the disparity between a stronger *expert model* and a weaker *amateur model*, pri-

[1]Source code is available at https://github.com/sandywangxiao/ContrastScore.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3045–3060

oritizing tokens with the largest probability gap. Beyond decoding, contrastive principles have also been applied to model architectures, such as the Self-Contrast Mixture-of-Experts (SCMoE), where different expert pathways within a model act as internal contrastive filters to enhance reasoning (Shi et al., 2024). These approaches highlight the potential of contrastive mechanisms in improving model outputs by using weaker models to generate contrastive signals, enabling stronger models to refine their predictions and achieve a balance between fluency and diversity.

Inspired by contrastive principles, we introduce **ContrastScore**, an evaluation metric that leverages structured disagreement between two models of differing capacities. Unlike conventional LLM-based evaluation methods that rely solely on a single model's likelihood estimates, ContrastScore incorporates a weaker auxiliary model as a contrastive signal, dynamically adjusting probability scores to improve alignment with human judgements. Specifically, ContrastScore is built on a discrepancy-based probability formulation that measures the absolute difference between the probabilities assigned by two models. By utilizing discrepancies between a stronger (expert) model and a weaker (amateur) model, it generates more calibrated and robust evaluation scores. We conduct extensive experiments on two widely employed text generation tasks, namely, machine translation and summarization, to evaluate the effectiveness of ContrastScore. Our evaluation compares ContrastScore against established metrics and various baseline models, including single-model and ensemble-based approaches. Experimental results show that ContrastScore achieves a higher correlation with human judgments, outperforming both single-model and ensemble-based methods. Notably, ContrastScore using smaller models (Qwen 3B, Qwen 0.5B) even surpasses Qwen7B with 7.0% of improvement in summarization despite having only half the parameters, demonstrating its efficiency. Furthermore, ContrastScore based on smaller models substantially enhances evaluation speed, providing at least a 1.5-fold increase in processing speed compared to a single larger model across both the Qwen and LLaMA families. Additionally, it effectively mitigates the likelihood biase, enhancing robustness in automatic evaluation. By incorporating contrastive principles into evaluation, ContrastScore paves the way for a new paradigm of more robust and efficient text evaluation.

The contributions of our paper are four-fold:

- We propose a simple yet highly effective difference-based formulation for contrastive evaluation, leveraging structured model discrepancies to produce more calibrated and reliable evaluation scores.
- We conduct extensive experiments across multiple generation tasks, diverse datasets, and various model families to rigorously assess the effectiveness of our approach.
- Contrastive evaluation strongly correlates with human judgments, outperforming single-model and ensemble methods while effectively addressing likelihood bias, which are prevalent in automatic evaluation.
- ContrastScore achieves significantly faster evaluation compared to larger single-model approaches, improving inference speed by at least 1.5 times while using only half the parameters, while maintaining comparable or even superior performance.

2 Related Work

Automatic Evaluation of Text Generation. Automatic evaluation metrics for text generation can broadly be categorized into *task-specific* and *general-purpose* approaches. Early task-specific metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and CHRF (Popović, 2015) rely on surface-level n-gram overlap with a reference. While originally designed for tasks like machine translation and summarization, these metrics often fail to reflect actual semantic similarity due to the variability of valid natural language expressions. To overcome these limitations, embedding-based metrics, such as BERTScore (Zhang et al., 2020) and MoverScore (Zhao et al., 2019), were introduced to compute token-level or document-level similarity using contextualized embeddings . Further improvements came from task-specific metrics fine-tuned on human ratings, such as COMET (Rei et al., 2020), COMET-KIWI (Rei et al., 2022), BLEURT (Sellam et al., 2020), and Prism (Thompson and Post, 2020). These models learn to directly predict human preferences but often generalize poorly across domains and tasks. General-purpose evaluation has increasingly shifted toward leveraging LLMs as evaluators. Prompt-based approaches such as G-Eval (Liu et al., 2023), ChatEval (Chan et al., 2024; Hong et al., 2025), and GPT-based judge systems treat LLMs as reference-

free annotators that directly assess text quality via carefully crafted prompts. While these methods often achieve high correlation with human judgments, they typically rely on proprietary models like GPT-4, making them costly, and sensitive to prompt phrasing (Leiter and Eger, 2024). An alternative line of work explores probability-based evaluation metrics, including BARTScore (Yuan et al., 2021) and GPTScore (Fu et al., 2024). BARTScore and GPTScore share a common formulation: they compute the conditional log-likelihood of a generation given context, using a pretrained language model as the scoring function. However, these evaluation methods inherit the biases and limitations of the underlying model (Ohi et al., 2024).

Contrastive Paradigm. Contrastive methods have recently gained attention in text generation, where the differences between models of varying capacities are leveraged to improve output quality (Su et al., 2022; O'Brien and Lewis, 2023; Li et al., 2023). Rather than relying on ensemble methods or majority voting, contrastive decoding explicitly uses a weaker model as a contrastive signal to refine or steer the output of a stronger model. This approach has demonstrated success in mitigating common generation issues such as repetition, overconfidence, and incoherence. Moreover, contrastive principles have been extended to model architecture optimization. Shi et al. (2024) propose the Self-Contrast Mixture-of-Experts (SC-MoE) framework, where different expert pathways within the same model act as internal contrastive filters, which improves reasoning capabilities and overall model performance. Note that contrastive decoding is also related to language model arithmetic in which arithmetic combinations of LLMs are considered for adjusted generation (Dekoninck et al., 2024). This includes subtraction, which leads to output favored by one LLM but disfavored by the other, which could be leveraged in adversarial settings (Zhang and Eger, 2024).

Inspired by these advancements, as described in §3.3, we extend contrastive principles beyond generation to the domain of evaluation. Instead of refining model outputs, we propose ContrastScore—a novel metric with a different formulation from prior contrastive decoding work.

3 Methodology

The primary goal of this study is to overcome limitations of single-model evaluators, such as capacity bias and over-reliance on model-specific likelihoods, by capturing where strong and weak models disagree in token-level probabilities. This section first reviews generative likelihood-based evaluation and contrastive decoding, then presents the design and formulation of ContrastScore.

3.1 Generative Evaluation

Our evaluation framework builds on probability-based text assessment methods, such as BARTScore, which estimate quality through log-likelihood computations under a pretrained generative model. Given a generated hypothesis **h**, its quality is evaluated as:

$$\text{Score}(\mathbf{h}|d,\mathcal{S}) = \sum_{t=1}^{m} w_t \log P(h_t|\mathbf{h}_{<t},\mathcal{S},\theta) \quad (1)$$

where $\mathcal{S}$ represents the supplementary text, which may consist of the source text **s** (in a source-based setting) or the reference output **r** (in a reference-based setting). d denotes the evaluation dimension (e.g., fluency), w_t [2] refers to the token-level weight, and θ is the model parameters.

3.2 Contrastive Decoding

Assume an expert and amateur model assigning probabilities

$$p^t_{\text{EXP}} = p(h_t \mid h_{<t},\mathcal{S},\theta_{\text{EXP}}) \quad (2)$$

$$p^t_{\text{AMA}} = p(h_t \mid h_{<t},\mathcal{S},\theta_{\text{AMA}}) \quad (3)$$

to the next token h_t, where p^t_{EXP} and p^t_{AMA} represent the probabilities of the expert and amateur models, respectively, for h_t and where θ_{EXP} and θ_{AMA} refer to the parameters of the expert and amateur model. Li et al. (2023) propose the *contrastive decoding* objective

$$\text{CD-Score} = \begin{cases} \log \frac{p^t_{\text{EXP}}}{p^t_{\text{AMA}}} & \text{if } h_t \in V_{\text{head}}(h_{<t}) \\ -\infty & \text{else} \end{cases} \quad (4)$$

Here, $V_{\text{head}}(h_{<t})$ selects the top most likely tokens x_t under the expert model given history $h_{<t}$. During next token prediction, CD-Score effectively only considers the most likely tokens of the expert model as candidates, and then assigns modified logits $\log \frac{p^t_{\text{EXP}}}{p^t_{\text{AMA}}}$ to them. The intuition behind CD-Score is to consider the most likely expert tokens as continuation for generation but then choose those

[2] Following prior studies (Yuan et al., 2021; Qin et al., 2023), w_t is weighed equally for all tokens in this work.

tokens among them which the amateur might not favor (e.g., repetitive tokens).

Unfortunately, CD-Score is unsuitable for evaluation because every generated sequence $\mathbf{h}$ containing a token h_t not in $V_{\text{head}}(h_{<t})$ would receive a score of $-\infty$, essentially being ruled out as a candidate, even though it should have been considered as part of the sequence under evaluation. As an alternative, one might consider the division score $\log \frac{p^t_{\text{EXP}}}{p^t_{\text{AMA}}}$ itself as an objective for evaluation. However, as pointed out by Li et al. (2023), this approach suffers from penalizing many standard text sequences as unlikely, for which both expert and amateur would hold the same probabilities. For example, if both models assign a probability of 0.9 or 0.1 to a token, the score remains the same, making it unable to distinguish between confidently correct and incorrect predictions. As a result, this approach fails to highlight important distinctions between tokens that should matter in evaluation. The division score also becomes instable especially when the amateur probability is close to zero.

3.3 ContrastScore

To address this issue *for evaluation*, we propose a subtraction-based contrastive formulation with a scaling factor $\gamma \in [0, 1]$ that downweights the amateur model. In this way, we naturally retain all tokens in the input sequence for evaluation, as well as a property that favors generations preferred by the expert model, similar to CD-Score. Importantly, the subtraction-based formulation ensures that the scores of individual tokens remain comparable across positions—unlike the division-based approach, where small variations in the amateur model's likelihood can lead to disproportionately large and unstable scores. By using a small γ, we control the influence of the amateur model's probability and mitigate such instability, yielding more reliable evaluation signals. Empirically, we choose a formula for contrastive evaluation that leverages the absolute value distance between the expert and the scaled down amateur:

$$\text{ContrastScore} = \sum_{t=1}^{m} w_t \log\left(\left|p^t_{\text{EXP}} - \gamma p^t_{\text{AMA}}\right|\right) \tag{5}$$

At each time step t, ContrastScore rewards tokens for which expert and (downweighted) amateur generation probabilities are maximally distinct (as the absolute value is a distance function between two distributions). This would ensure that ContrastScore leverages the strengths of the expert model but removes the limitations of the amateur model; in practice, it might for example lean toward generations that are less repetitive, which the amateur model tends to favor. Note, however, that there are probability ranges where the scaled down amateur would be preferred over the expert—e.g., when $p^t_{\text{EXP}} = 0$, then our formula would still assign the scaled probability of the amateur to h_t. With $\gamma = 1$, ContrastScore would be indifferent between the expert and the amateur but indiscriminately favor disagreement between them, an undesirable behavior similar to the division score above.

4 Experimental Setup

4.1 Datasets and Tasks

For summarization, we use the **SummEval** dataset (Fabbri et al., 2021), which contains model-generated summaries of CNN/DailyMail articles, as well as **QAGS-XSUM** (Wang et al., 2020), which includes 239 system outputs on the XSUM dataset. For machine translation, we use the **MQM22** and **MQM23** datasets from the WMT22 (Freitag et al., 2022) and WMT23 (Freitag et al., 2023) Metrics Shared Tasks. The language pairs involved represent a diverse range of translation scenarios, including high-resource (EN-DE), typologically distant (ZH-EN and EN-RU), and low-resource (HE-EN) settings, thereby enabling a comprehensive evaluation of our approach. See AppendixA.1 for more details.

4.2 Baseline Metrics

We consider the following baseline metrics for comparison: **BLEU** (Papineni et al., 2002), **CHRF** (Papineni et al., 2002), **ROUGE 1**, **ROUGE 2**, **ROUGE L** (Lin, 2004), **COMET** (Rei et al., 2020), **COMET-KIWI** (Rei et al., 2022), **BERTScore** (Zhang et al., 2020), **MoverScore** (Zhao et al., 2019), and **BARTScore** (Yuan et al., 2021). The details of these metrics can be found in Appendix A.2. In addition, we compare it against **Single-model** evaluators as well as an **Ensemble** setting, where the probabilities of the expert and amateur models are averaged per token, using models from the LLaMA and Qwen families to ensure generalizability. The ensemble probability at token t, denoted as p^t_{Ens}, is computed as:

$$p^t_{\text{Ens}} = \frac{1}{2}\left(p^t_{\text{EXP}} + p^t_{\text{AMA}}\right) \tag{6}$$

4.3 Meta-Evaluation

We assess the evaluation metrics in terms of quality, biases, and efficiency.

Correlation with Human Scores (Quality). The effectiveness of an evaluator is measured using the Pearson correlation between its scores and human scores, where a higher correlation indicates stronger alignment with human judgments. For translation, Pearson correlation assesses the evaluator's ability to capture differences between candidate translations based on the source texts, comparing them to human judgment scores derived from fine-grained MQM annotations (Freitag et al., 2021). For summarization, we compute Pearson correlation with human judgments across different quality aspects: coherence, fluency, consistency, relevance, and factuality.

Bias Evaluation. We measure biases in likelihood bias. Recent studies have shown that LLMs tend to favor sentences they deem more likely, often assigning higher evaluation scores and performing better on such sentences, regardless of the specific task or evaluation criteria (Ohi et al., 2024; McCoy et al., 2024a,b). Ohi et al. (2024) find that LLM-based evaluators systematically overrate high-likelihood sentences and underrate low-likelihood ones compared to human scores. To quantify this likelihood bias, they propose BiasScore defined as:

$$\text{BiasScore} = \rho(LS, US) \tag{7}$$

where ρ is the Spearman correlation coefficient, *LS* (Likelihood Score) represents the probability P assigned by the LLM, and *US* (Unfairness Score) captures the discrepancy between evaluator scores and human scores. Likelihood bias measures the extent to which an evaluation metric is influenced by the model's inherent probability estimates. A *lower BiasScore* indicates that the metric is less dependent on the model's likelihood biases.

Efficiency. It is measured by the number of samples processed per second, with a higher processing speed indicating greater efficiency.

4.4 Hyperparameters and Environment

We use models from the same family but of different sizes, where a larger model serves as the expert and a smaller model as the amateur. Specifically, we use LLaMA3.2-Instruct (1B, 3B) and LLaMA3.1-Instruct (8B), as the LLaMA 3.2 version does not include an 8B model. Additionally, we evaluate Qwen2.5-Instruct (0.5B, 3B, and 7B).[3] In all our experiments[4], the scaling factor for the amateur model, γ, is set to 0.1.[5] All experiments were run on a GPU cluster under same conditions and node configuration to ensure fair comparisons. Each node of the cluster has 4×H100 GPUs.

5 Results

We evaluate ContrastScore in terms of its correlation with human scores, bias, and efficiency, compared to single-model and ensemble methods.

5.1 Correlation with Human Scores

Summarization. Table 1 presents the Pearson correlation between different evaluators and human scores for summarization. Our results indicate that ContrastScore provides best results overall. Furthermore, ContrastScore consistently surpasses single models while utilizing much fewer parameters. Specifically, ContrastScore with Qwen(3B, 0.5B) outperforms the single Qwen 7B by 7.0%, and ContrastScore with LLaMA(3B, 1B) exceeds single LLaMA 8B by 6.2%. These results reinforce the observation that ContrastScore can efficiently enhance model evaluation by leveraging smaller models. Besides improving overall correlation with human scores, ContrastScore demonstrates notable gains across key summarization dimensions, particularly coherence, fluency, relevance and factuality. When applied to LLaMA(8B, 3B), it enhances coherence by 25.5%, fluency by 6.7%, relevance by 11.1%, and factuality by 57.6%, compared to the single LLaMA 8B. However, its impact on consistency differs across model families.

Machine Translation. Table 2 presents the Pearson correlation between various evaluators and human scores. The results show that, overall, ContrastScore outperforms both single-model and ensemble approaches. For example, ContrastScore based on Qwen(7B, 3B) improves correlation by 7.1% over the single Qwen 7B model and by 4.6% over the ensemble of Qwen(7B, 0.5B). In contrast, ensembling does not always improve performance: While ensemble methods are generally expected to refine predictions, they can sometimes reduce cor-

[3] See the discussion in the Limitations section on why we use models from the same family.

[4] Following Li et al. (2023), we set the decoding temperatures for the expert and amateur to 0.5 and 1.5, respectively.

[5] This setting is based on our pilot investigation, which found that $\gamma = 0.1$ gives the best overall performance empirically for ZH-EN in MT (Figure 2 in Appendix).

		Evaluators	QAGS-XSUM	SummEval				AVG
			Factuality	Coherence	Consistency	Fluency	Relevance	
Baseline		ROUGE-1	0.055	0.238	0.079	0.071	0.330	0.154
		ROUGE-2	0.121	0.164	0.073	0.066	0.229	0.130
		ROUGE-L	0.075	0.207	0.071	0.078	0.303	0.147
		MoverScore	0.028	0.129	0.141	0.133	0.280	0.142
		BERTScore	0.024	0.342	0.115	0.156	0.372	0.202
		BARTScore	0.099	0.450	0.339	0.343	**0.428**	0.332
LLaMA	Single	1B	0.064	0.417	0.549	0.542	0.344	0.383
		3B	0.157	0.388	0.575	0.530	0.338	0.398
		8B	0.205	0.368	0.588	0.539	0.332	0.406
	Ensemble	(3B,1B)	0.170	0.427	0.564	0.542	0.344	0.409
		(8B,1B)	0.212	0.429	0.581	0.556	0.357	0.427
		(8B,3B)	0.241	0.414	0.587	0.544	0.351	0.428
	Contrast	(3B,1B)	0.262	0.456	0.554	0.551	0.334	0.431
		(8B,1B)	**0.324**	0.461	0.584	**0.575**	0.368	0.462
		(8B,3B)	0.323	**0.462**	0.587	**0.575**	0.369	**0.463**
Qwen	Single	0.5B	0.107	0.386	0.536	0.520	0.315	0.373
		3B	0.112	0.370	0.544	0.492	0.319	0.367
		7B	0.091	0.371	0.560	0.510	0.327	0.372
	Ensemble	(3B,0.5B)	0.081	0.394	0.568	0.536	0.331	0.382
		(7B,0.5B)	0.095	0.386	0.577	0.538	0.333	0.386
		(7B,3B)	0.103	0.378	0.584	0.521	0.330	0.383
	Contrast	(3B,0.5B)	0.128	0.397	0.594	0.546	0.323	0.398
		(7B,0.5B)	0.122	0.376	**0.608**	0.554	0.322	0.396
		(7B,3B)	0.134	0.384	0.606	0.557	0.334	0.403

Table 1: Pearson correlation of evaluators with human scores in summarization. **boldface** represents best overall scores, while underline represents best scores within each model group (Baseline, LLaMA, Qwen). Overall, ContratScore outperforms single and ensemble methods as well as baseline metrics for both LLaMA and Qwen families.

relation with human scores. For instance, ensembling LLaMA (8B, 1B) leads to a 3.3% decrease in correlation compared to the LLaMA 8B model. Similar trends can be observed in pairwise accuracy (Deutsch et al., 2023), as shown in Table 10 of the Appendix.

Beyond the cases when the amateur model is too weak to provide a meaningful correction signal—such as the low correlation scores observed in the HE-EN pair for both LLaMA 1B and 3B (0.32 and 0.29, respectively), which fail to offer reliable contrastive feedback against LLaMA 8B—our results demonstrate that ContrastScore can achieve correlation scores comparable to, or even exceeding, those of the largest model using only two smaller models. Specifically, ContratScore of Qwen(3B, 0.5B) surpasses single Qwen 7B by 3.1%, improving from 0.449 to 0.463 with only half the parameters.

5.2 Bias Analysis

Table 3 presents the likelihood bias scores across different evaluators of the LLaMA family. The results clearly demonstrate that ContrastScore substantially mitigates likelihood bias compared to both single-model and ensemble approaches. Specifically, ContrastScore achieves lower likelihood bias than any of the individual models employed. For instance, using LLaMA (8B, 3B), it reduces likelihood bias by 18.0% and 64.3% compared to the single LLaMA 8B model on QAGS-XSUM and MQM22, respectively.

ContrastScore is also effective in mitigating likelihood bias for the Qwen model family, although the improvements are somewhat smaller than those observed with LLaMA. Notably, ContrastScore consistently outperforms all individual Qwen models in reducing likelihood bias, with the exception of the 0.5B model on MQM23. These findings suggest that ContrastScore is highly effective in diminishing the reliance on the inherent likelihood of individual models. See also Tables 8 and 9 in Appendix for more details and breakdown results.

5.3 Efficiency Analysis

We report processing speeds for both machine translation and summarization evaluation tasks (see Table 7 in Appendix C). Recall that in Tables 1 and 2, our ContrastScore employing two smaller models (for both LLaMA and Qwen families) achieves higher correlation with human judgments com-

		Evaluators	MQM22			MQM23			AVG
			EN-DE	ZH-EN	EN-RU	EN-DE	ZH-EN	HE-EN	
Baseline		BLEU	0.194	0.156	0.142	0.163	0.085	0.208	0.158
		CHRF	0.236	0.156	0.169	0.232	0.063	0.244	0.183
		BERTScore	0.263	0.311	0.197	0.325	0.236	0.336	0.278
		BARTScore	0.254	0.287	0.201	0.201	0.182	0.317	0.240
		COMET	**0.476**	0.403	**0.417**	0.432	0.396	0.417	0.424
		COMET-KIWI	0.392	0.367	0.354	0.475	0.442	0.395	0.404
LLaMA	Single	1B	0.255	0.371	0.286	0.530	0.518	0.320	0.380
		3B	0.284	0.366	0.293	0.578	0.526	0.293	0.390
		8B	0.363	0.376	0.356	0.632	0.574	**0.491**	0.465
	Ensemble	(3B,1B)	0.287	0.371	0.307	0.568	0.538	0.306	0.396
		(8B,1B)	0.325	0.382	0.354	0.605	0.575	0.462	0.450
		(8B,3B)	0.337	0.379	0.348	0.619	0.574	0.453	0.452
	Contrast	(3B,1B)	0.338	0.382	0.347	0.599	0.562	0.284	0.419
		(8B,1B)	0.392	0.391	0.406	**0.641**	0.590	0.484	**0.484**
		(8B,3B)	0.383	0.393	0.409	0.639	0.595	0.482	0.483
Qwen	Single	0.5B	0.222	0.394	0.294	0.487	0.557	0.347	0.383
		3B	0.306	0.413	0.299	0.573	0.594	0.415	0.433
		7B	0.326	0.419	0.330	0.600	0.574	0.445	0.449
	Ensemble	(3B,0.5B)	0.290	0.408	0.315	0.548	0.599	0.421	0.430
		(7B,0.5B)	0.301	0.412	0.331	0.567	0.598	0.446	0.443
		(7B,3B)	0.333	0.424	0.341	0.599	0.608	0.456	0.460
	Contrast	(3B,0.5B)	0.342	0.432	0.351	0.590	0.628	0.431	0.463
		(7B,0.5B)	0.359	0.435	0.382	0.611	0.628	0.465	0.480
		(7B,3B)	0.362	**0.439**	0.384	0.605	**0.629**	0.466	0.481

Table 2: Pearson correlation of evaluators with human scores in machine translation. **bold** represents best overall scores, while underline represents best scores within each model group (Baseline, LLaMA, Qwen). Overall, ContrastScore outperforms single and ensemble methods, as well as baseline metrics for both LLaMA and Qwen families.

Settings		Machine Translation		Summarization	
		MQM22	MQM23	Q-XSUM	SummEval
Single	1B	0.342	0.212	0.382	0.348
	3B	0.323	0.245	0.289	0.385
	8B	0.297	0.352	0.267	0.381
Ensemble	(3B,1B)	0.215	0.123	0.249	0.308
	(8B,1B)	0.180	0.152	0.225	0.326
	(8B,3B)	0.222	0.229	0.242	0.359
Contrast	(3B,1B)	**0.058**	**0.026**	0.233	**0.183**
	(8B,1B)	0.104	0.134	0.220	0.262
	(8B,3B)	0.106	0.137	**0.219**	0.240

Table 3: Likelihood bias scores for machine translation and summarization tasks across LLaMA model family. The lowest overall bias score is boldfaced.

Settings		Machine Translation		Summarization	
		MQM22	MQM23	Q-XSUM	SummEval
Single	0.5B	0.341	**0.252**	0.347	0.376
	3B	0.442	0.451	0.349	0.392
	7B	0.441	0.463	0.373	0.398
Ensemble	(3B,0.5B)	0.331	0.327	**0.282**	0.379
	(7B,0.5B)	0.286	0.294	0.289	0.393
	(7B,3B)	0.358	0.381	0.345	0.369
Contrast	(3B,0.5B)	0.307	0.314	0.294	**0.236**
	(7B,0.5B)	0.287	0.302	0.318	0.296
	(7B,3B)	**0.272**	0.287	0.353	0.329

Table 4: Likelihood bias scores for machine translation and summarization tasks across Qwen model family. The lowest overall bias score is boldfaced.

pared to using a single larger model in summarization, and similar trends are observed in machine translation within the Qwen family. To assess efficiency, we compare the processing speeds of ContrastScore using two smaller models with those of a single larger model. In the LLaMA family, the single large model LLaMA 8B processes 26.43 samples/s for summarization, whereas ContrastScore with LLaMA 3B and 1B achieves 39.12 samples/s, respectively—approximately 1.5× faster in both cases. Similarly, in the Qwen family, ContrastScore using Qwen 3B and 0.5B processes approximately 1.7× faster on summarization and 1.5× faster on machine translation than Qwen 7B. Figure 1 further illustrates that the ContrastScore framework with two smaller models delivers improved efficiency and better evaluation quality in summarization.

6 Further Analysis

Weighted Ensemble. We investigate whether the averaged ensemble baseline used in our experiment is a sufficiently strong setup, or if a weighted

Source Text		产品防伪标识在哪里									
	Tokens:	Where	is	the	anti	-	fe	iting	logo	**Log(P)**	**Rank**
Hypothesis 1 Human Rank: 1	Expert	0.2471	0.7305	0.9922	0.02881	0.9805	0.7969	1.000	0.000457	-0.717	2
	Amateur	2.265e-06	0.5625	0.9922	0.000335	1.000	1.000	1.000	0.06592	-1.319	1
	Contrast	0.2471	0.6758	0.8945	0.02881	0.8789	0.6953	0.8984	0.006134	-0.605	1
	Tokens:	Where	is	the	anti	-	fe	iting	product		
Hypothesis 2 Human Rank:2	Expert	0.2471	0.7305	0.9922	0.02881	0.9805	0.7969	1.000	0.002808	-0.618	1
	Amateur	2.265e-06	0.5625	0.9922	0.000335	1.000	1.000	1.000	5.841e-05	-1.701	2
	Contrast	0.2471	0.6758	0.8945	0.02881	0.8789	0.6953	0.8984	0.002808	-0.647	2
	Tokens:	What	Is	The	National	Debt	Limit				
Hypothesis 3 Human Rank:3	Expert	0.005951	0.000168	0.2910	4.268e-05	0.001602	0.004944			-2.668	3
	Amateur	0.000572	6.845e-08	0.1543	6.482e-07	3.123e-05	0.000140			-4.294	3
	Contrast	0.005951	0.000168	0.2754	4.268e-05	0.001602	0.004944			-2.672	3

Table 5: Case Study: Comparison of ContrastScore with Qwen 3B as an expert and Qwen 0.5B as an amateur model for Chinese-to-English (**ZH-EN**) on MQM23. **Log(P)** denotes mean log-probability across all target tokens.

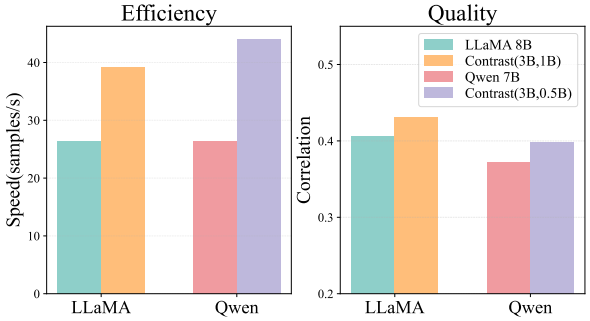

Figure 1: Efficiency and quality of ContrastScore with smaller models compared to single larger model on summarization task.

combination of model probabilities could yield further improvements. Specifically, we evaluate a weighted ensemble approach that linearly combines the output probabilities of a larger and a smaller model, allowing us to examine whether tuning the weight parameter provides a meaningful advantage over uniform averaging.

$$p^t_{\text{Ens}} = \gamma \cdot p^t_{\text{EXP}} + (1-\gamma) \cdot p^t_{\text{AMA}} \qquad (8)$$

where $\gamma \in [0, 1]$ is the weight factor that controls the contribution of each model. When $\gamma = 0.5$, this formulation corresponds to that reported in our main results. To analyze the impact of γ, we conduct a parameter sweep on the summarization task using LLaMA (3B, 1B), varying γ from 0 to 1 in increments of 0.05. The results, presented in Figure 3, show that the weighted ensemble achieves its highest performance at $\gamma = 0.55$, which is nearly identical to the performance at $\gamma = 0.5$ (i.e., the averaged ensemble baseline). This indicates that the simple averaged ensemble is already a strong and competitive configuration, and that further weighting does not lead to meaningful improvement.

Case Study. To provide a qualitative analysis, Table 5 presents a case study comparing ContrastScore with a single expert model (Qwen 3B) and an amateur model (Qwen 0.5B) for Chinese-to-English (ZH-EN) machine translation. It can be observed that the expert model misranks *Hypothesis 1* as the second-best translation, whereas human evaluators rank it as the best (Rank 1). This misranking is largely influenced by the expert model assigning a very low probability (0.000457) to *"logo"*, a key token in the translation. In contrast, the amateur model assigns a significantly higher probability (0.06592) to *"logo"*, demonstrating greater confidence in its relevance. ContrastScore leverages this probability discrepancy to adjust the probability of *"logo"*, thereby increasing the overall score of *Hypothesis 1* and correctly ranking it as the best translation. This case illustrates how ContrastScore mitigates expert model underestimation of critical tokens, leading to improved evaluation alignment with human judgment.

In the case of *Hypothesis 3*, all methods (i.e., expert, amateur, and ContrastScore) correctly rank it as the worst translation (Rank 3). This consistency occurs because both the expert and amateur models assign low probabilities across all tokens, indicating poor translation quality. Unlike in *Hypothesis 1*, where significant probability discrepancies require adjustment, ContrastScore makes minimal modifications since the expert and amateur models are already in agreement in their evaluation. This demonstrates that ContrastScore does not introduce unnecessary changes when the models agree, ensuring that it only refines scores when meaningful corrections are needed.

7 Conclusion

In this paper, we introduce ContrastScore, an evaluation metric that leverages structured disagreement between two language models of differing capacities. Unlike conventional LLM-based evaluation methods that rely solely on a single model's likelihood estimates, ContrastScore incorporates a weaker auxiliary model as a contrastive signal to adjust probability scores for better alignment with human judgments. Extensive experiments demonstrate that ContrastScore consistently achieves stronger correlation with human evaluations than both single-model and ensemble-based baselines. Furthermore, ContrastScore is computationally-efficient and can effectively mitigate likelihood bias, resulting in a more robust evaluation.

Limitations

This work presents the following potential limitations: (1) Our investigation is limited to the LLaMA and Qwen families of models. While ContrastScore demonstrates effectiveness within these two model families, further evaluation across a broader range of large language model architectures is necessary to establish its generalizability. (2) This work relies on token-level probabilities and therefore cannot combine models from different families due to differences in their subword tokenization. In future work, word-level probability could be explored to enable mixing across model families. (3) The metrics explored in this paper are not competitive to the state-of-the-art (SOTA) metrics developed for MT and summarization, such as MetricX (Juraska et al., 2024) or XCOMET-XXL (Guerreiro et al., 2024), which may use much larger models and/or fine-tuning on human annotations. We do not claim in this paper to beat SOTA metrics, but that *contrastive* principles, involving two models, are superior to non-contrastive principles for evaluation metric design within one evaluation paradigm (BARTScore, in our case). To design competitive metrics, future work should consider involving (considerably) larger models than we did in this work, beyond 8B parameters.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback and helpful suggestions. Xiao Wang is supported by a studentship funded by the Department of Computer Science, University of Manchester. The NLLG Lab gratefully acknowledges support from the Federal Ministry of Education and Research (BMBF) via the research grant "Metrics4NLG" and the German Research Foundation (DFG) via the Heisenberg Grant EG 375/5-1.

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.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Yanran Chen and Steffen Eger. 2023. MENLI: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.

Jasper Dekoninck, Marc Fischer, Luca Beurer-Kellner, and Martin Vechev. 2024. Controlled text generation via language model arithmetic. In *The Twelfth International Conference on Learning Representations*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. 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.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, and 1 others. 2023. Results of wmt23 metrics

shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Hanhua Hong, Chenghao Xiao, Yang Wang, Yiqi Liu, Wenge Rong, and Chenghua Lin. 2025. Beyond one-size-fits-all: Inversion learning for highly effective nlg evaluation prompts. *arXiv preprint arXiv:2504.21117*.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Katikapalli Subramanyam Kalyan. 2024. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, 6:100048.

Daniil Larionov and Steffen Eger. 2024. Promptoptme: Error-aware prompt compression for llm-based mt evaluation metrics. *Preprint*, arXiv:2412.16120.

Christoph Leiter and Steffen Eger. 2024. PrExMe! large scale prompt exploration of open source LLMs for machine translation and summarization evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11481–11506, Miami, Florida, USA. Association for Computational Linguistics.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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.

Yiqi Liu, Nafise Moosavi, and Chenghua Lin. 2024. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12688–12701, Bangkok, Thailand. Association for Computational Linguistics.

Adian Liusie, Potsawee Manakul, and Mark JF Gales. 2024. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models. In *EACL (1)*.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Qingyu Lu, Liang Ding, Liping Xie, Kanjian Zhang, Derek F. Wong, and Dacheng Tao. 2023. Toward human-like evaluation for natural language generation with error analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5892–5907, Toronto, Canada. Association for Computational Linguistics.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024a. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D Hardy, and Thomas L Griffiths. 2024b. When a language model is optimized for reasoning, does it still show embers of autoregression? an analysis of openai o1. *arXiv preprint arXiv:2410.01792*.

Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *arXiv preprint arXiv:2309.09117*.

Masanari Ohi, Masahiro Kaneko, Ryuto Koike, Mengsay Loem, and Naoaki Okazaki. 2024. Likelihood-based mitigation of evaluation bias in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3237–3245, Bangkok, Thailand. Association for Computational Linguistics.

Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. 2024. G-DIG: Towards gradient-based DIverse and hiGh-quality instruction data selection for machine translation. In

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15395–15406, Bangkok, Thailand. 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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Yiwei Qin, Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2023. T5Score: Discriminative fine-tuning of generative evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15185–15202, Singapore. Association for Computational Linguistics.

Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Zekun Moore Wang, Jian Yang, Ge Zhang, and 1 others. 2024. Hellobench: Evaluating long text generation capabilities of large language models. *arXiv preprint arXiv:2409.16191*.

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.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte M Alves, Alon Lavie, and 1 others. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. *arXiv preprint arXiv:2209.06243*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Xiangqing Shen, Siwei Wu, and Rui Xia. 2023. Dense-atomic: Towards densely-connected atomic with high knowledge coverage and massive multi-hop paths. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13292–13305.

Chufan Shi, Cheng Yang, Xinyu Zhu, Jiahao Wang, Taiqiang Wu, Siheng Li, Deng Cai, Yujiu Yang, and Yu Meng. 2024. Unchosen experts can contribute too: Unleashing moe models' power by self-contrast. In *Advances in Neural Information Processing Systems*, volume 37, pages 136897–136921. Curran Associates, Inc.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. *Advances in Neural Information Processing Systems*, 35:21548–21561.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is unfair: On social bias in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

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.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Ran Zhang and Steffen Eger. 2024. Llm-based multi-agent poetry generation in non-cooperative environments. *ArXiv*, abs/2409.03659.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Kun Zhao, Bohao Yang, Chenghua Lin, Wenge Rong, Aline Villavicencio, and Xiaohui Cui. 2023. Evaluating open-domain dialogues in latent space with next sentence prediction and mutual information. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–574, Toronto, Canada. Association for Computational Linguistics.

Kun Zhao, Bohao Yang, Chen Tang, Chenghua Lin, and Liang Zhan. 2024. SLIDE: A framework integrating small and large language models for open-domain dialogues evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15421–15435, Bangkok, Thailand. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings*

of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Datasets, Metrics and Prompts

A.1 Datasets

Machine Translation. **MQM22** and **MQM23** datasets are annotated by professional translators from the WMT22 (Freitag et al., 2022) and WMT23 (Freitag et al., 2023) Metrics Shared Tasks, based on the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014). **MQM22** covers 3 language pairs: English-German (EN-DE), Chinese-English (ZH-EN) and English-Russian (EN-RU), and comprises 1315, 1875, and 1315 segments for language pair EN-DE, ZH-EN, EN-RU respectively, with every segments including 15 system outputs. **MQM23** covers 3 language pairs: English-German (EN-DE), Chinese-English (ZH-EN) and Hebrew-English (HE-EN), including 460, 1177, and 820 segments and 12, 15, and 12 systems for EN-DE, ZH-EN, HE-EN respectively.

Summarization. **SummEval** (Fabbri et al., 2021) contains 1600 model-generated summaries of CNN/DailyMail articles, each annotated by 3 experts and 5 crowd workers. It covers aspects of coherence, consistency, fluency, and relevance. **QAGS-XSUM** (Wang et al., 2020) includes 239 system outputs from a fine-tuned BART on XSUM dataset, focusing on factuality aspect.

A.2 Baseline Metrics

BLEU (Papineni et al., 2002) is based on the precision of n-grams between the MT output and its reference weighted by a brevity penalty.

CHRF (Popović, 2015) uses character n-grams instead of word n-grams to compare the MT output with the reference.

ROUGE (Lin, 2004) measures the lexical overlap between the hypothesis and reference. We consider 3 variants ROUGE-1, ROUGE-2, and ROUGE-L.

COMET (Rei et al., 2020) is a learnt metric that is fine-tuned to produce evaluation scores for a given translation by comparing its representation to source and reference embeddings.

COMET-KIWI (Rei et al., 2022) is a reference-free variant of COMET for machine translation evaluation. It uses a multilingual encoder to take only the source and system output as input and predicts a continuous quality score.

BERTScore (Zhang et al., 2020) leverages contextual embeddings from BERT to compare words in candidate and reference sentences using cosine similarity.

MoverScore (Zhao et al., 2019) measures semantic similarity between a candidate and a reference by combining contextualized embeddings with Earth Mover's Distance.

BARTScore (Yuan et al., 2021) is a generative metric that uses BART to evaluate the generated text by calculating the probabilities of the tokens.

A.3 Prompts

The prompts for the summarization and machine translation tasks are presented in Table 6.

Tasks	Prompts
machine translation	Translate the following sentence to {target language}:
summarization	Write an accurate, relevant, and coherent summary of the following texts:\n {Article}\n Summary:\n

Table 6: Prompts for tasks description

A.4 Additional Details

Artifact Licenses We use the following artifacts:

- Qwen 2.5 - Qwen License Agreement
- LLaMA 3.1 and 3.2 - LLaMA 3.1 and 3.2 Community License Agreement
- COMET22 - Apache 2.0 License
- COMET-Kiwi - CC BY NC SA 4.0
- SummEval - MIT License
- WMT23 - MIT License

Our use of those artifacts complies with license terms and applicable intended use policies.

PII information in data We did not specifically check whether the datasets used contain PII information. However, we have noticed that WMT23 authors made an effort to mask PII information with special tags. We do not disseminate any new dataset; therefore, PII protection is out of the scope of our work.

Package Versions and Hardware We use 'unbabel-comet' version 2.2.0 to run baselines for MT evaluation. Experiments were conducted on the university SLURM computing cluster with various available GPUs. For benchmarking purposes, we used H100 GPUs with 96GB of VRAM.

B Analysis

B.1 The impact of hyperparameter gamma

We study how sensitive our method is to γ in Figure 2.

We test ZH-EN language pair of machine translation task, based on LLaMA family. Figure 2 shows that $\gamma \in [0.08, 0.2]$ leads to good performance, robust in different model size settings. Furthermore, $\gamma = 0.1$ produces the best performance, ensuring a small, controlled refinement.

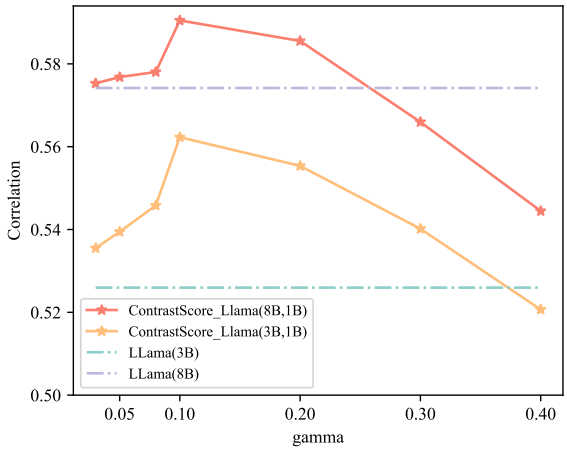

Figure 2: Exploration of the impacts of γ. Testing correlation between evaluator score and human score in ZH-EN language pair on MQM23.

B.2 Weighted ensemble

To explore whether the advantages of ContrastScore can be replicated by a simpler method based on probability combination, we evaluate a weighted ensemble approach that linearly merges the output probabilities of the larger and smaller models. The results are presented in the Figure 3.

C Results

C.1 Likelihood Bias

The detailed likelihood bias scores for every datasets of machine translation and summarization tasks are shown in Table 8 and Table 9. ContrastScore demonstrates consistent effectiveness in reducing likelihood bias across both machine translation and summarization tasks, particularly in challenging language pairs and critical evaluation aspects. In machine translation in Table 8, ContrastScore yields substantial improvements for the LLaMA family across diverse language pairs, especially for low-resource or morphologically complex pairs such as EN-RU and HE-EN. In summarization in Table 9, ContrastScore offers the most pronounced improvements in factuality and coherence, which are often the most difficult dimensions for language models. Overall, ContrastScore effectively mitigates likelihood bias compared to both single-model and ensemble baselines in both the LLaMA and Qwen families.

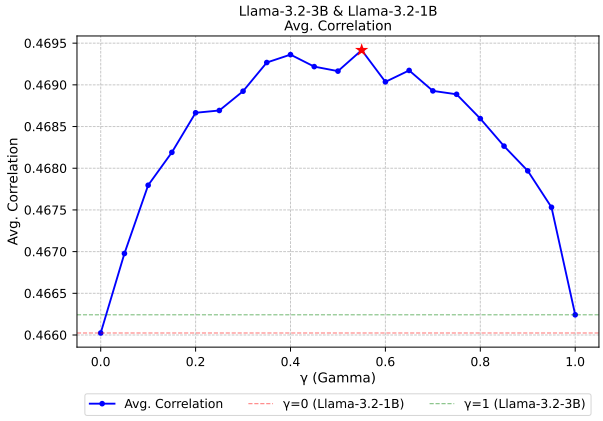

Figure 3: Exploration of weighted ensemble parameter γ for LLaMA-3.2 (3B, 1B) on summarization. Best quality occurs at $\gamma = 0.55$ (star), but remains below ContrastScore using the same models. Horizontal lines show individual model performance.

C.2 Efficiency

We use processing speed as an indicator of efficiency, with detailed results for both machine translation and summarization evaluation tasks presented in Table 7. Processing more samples per second indicates higher evaluation efficiency. These results demonstrate that ContrastScore offers substantial gains in evaluation efficiency, especially when leveraging smaller models.

	LLaMA	SUM	MT	Qwen	SUM	MT
Single	1B	63.29	339.04	0.5B	141.96	552.68
	3B	47.61	153.25	3B	48.21	194.22
	8B	26.43	72.17	7B	26.43	99.84
Ensemble	(3B,1B)	39.51	110.63	(3B,0.5B)	44.94	154.5
	(8B,1B)	23.81	66.8	(7B,0.5B)	26.11	93.92
	(8B,3B)	19.81	53.95	(7B,3B)	20.35	71.04
Contrast	(3B,1B)	39.12	109.31	(3B,0.5B)	44.06	153.99
	(8B,1B)	23.33	66.11	(7B,0.5B)	25.71	93.41
	(8B,3B)	19.05	53.45	(7B,3B)	19.61	70.97

Table 7: Processing Speed for Machine Translation and summarization evaluation tasks. Measured in Samples Per Second with batch size of 16 on single H100 GPU.

	Settings		MQM22				MQM23			
			EN-DE	ZH-EN	EN-RU	AVG	EN-DE	ZH-EN	HE-EN	AVG
LLaMA	Single	1B	0.437	0.312	0.276	0.342	0.142	0.237	0.257	0.212
		3B	0.426	0.290	0.253	0.323	0.146	0.284	0.304	0.245
		8B	0.367	0.287	0.238	0.297	0.181	0.309	0.565	0.352
	Ensemble	(3B,1B)	0.324	0.219	0.101	0.215	-0.002	0.169	0.202	0.123
		(8B,1B)	0.278	0.189	0.074	0.180	0.006	0.148	0.302	0.152
		(8B,3B)	0.298	0.227	0.140	0.222	0.079	0.223	0.385	0.229
	Contrast	(3B,1B)	0.195	0.143	-0.165	0.058	-0.106	0.096	0.088	0.026
		(8B,1B)	0.164	0.131	0.016	0.104	-0.044	0.093	0.353	0.134
		(8B,3B)	0.187	0.141	-0.011	0.106	-0.051	0.110	0.350	0.137
Qwen	Single	0.5B	0.493	0.233	0.296	0.341	0.184	0.214	0.358	0.252
		3B	0.580	0.367	0.379	0.442	0.279	0.476	0.598	0.451
		7B	0.559	0.383	0.383	0.441	0.333	0.438	0.616	0.463
	Ensemble	(3B,0.5B)	0.463	0.275	0.256	0.331	0.115	0.370	0.498	0.327
		(7B,0.5B)	0.421	0.236	0.202	0.286	0.095	0.304	0.483	0.294
		(7B,3B)	0.490	0.303	0.280	0.358	0.206	0.376	0.562	0.381
	Contrast	(3B,0.5B)	0.432	0.257	0.233	0.307	0.087	0.350	0.505	0.314
		(7B,0.5B)	0.415	0.239	0.206	0.287	0.092	0.296	0.518	0.302
		(7B,3B)	0.395	0.235	0.187	0.272	0.091	0.292	0.499	0.287

Table 8: Likelihood bias of machine translation task for the LLaMA and Qwen family. ContrastScore can effectively mitigate the likelihood bias compared to both single and ensemble methods on MQM22 and MQM23 datasets.

	Settings		Q-XUM	SummEval				
			Factuality	Coherence	Consistency	Fluency	Relevance	AVG
LLaMA	Single	1B	0.382	0.104	0.471	0.539	0.279	0.348
		3B	0.289	0.169	0.477	0.573	0.321	0.385
		8B	0.267	0.161	0.477	0.575	0.311	0.381
	Ensemble	(3B,1B)	0.249	0.095	0.406	0.488	0.240	0.308
		(8B,1B)	0.225	0.119	0.417	0.507	0.261	0.326
		(8B,3B)	0.242	0.152	0.450	0.538	0.295	0.359
	Contrast	(3B,1B)	0.233	-0.008	0.279	0.339	0.122	0.183
		(8B,1B)	0.220	0.074	0.337	0.424	0.214	0.262
		(8B,3B)	0.219	0.048	0.326	0.400	0.185	0.240
Qwen	Single	0.5B	0.347	0.139	0.489	0.567	0.308	0.376
		3B	0.349	0.173	0.492	0.570	0.332	0.392
		7B	0.373	0.185	0.489	0.584	0.334	0.398
	Ensemble	(3B,0.5B)	0.282	0.170	0.461	0.567	0.318	0.379
		(7B,0.5B)	0.289	0.194	0.464	0.582	0.332	0.393
		(7B,3B)	0.345	0.151	0.472	0.560	0.295	0.369
	Contrast	(3B,0.5B)	0.294	0.017	0.364	0.410	0.152	0.236
		(7B,0.5B)	0.318	0.077	0.411	0.488	0.208	0.296
		(7B,3B)	0.353	0.112	0.430	0.526	0.249	0.329

Table 9: Likelihood bias of summarization task for the LLaMA and Qwen family. ContrastScore can effectively mitigate the likelihood bias compared to both single and ensemble methods on QAGS-XSUM and SummEval datasets.

		Evaluators	MQM22			MQM23			AVG
			EN-DE	ZH-EN	EN-RU	EN-DE	ZH-EN	HE-EN	
LLaMA	Single	1B	0.407	0.493	0.495	0.655	0.625	0.440	0.519
		3B	0.415	0.497	0.498	0.661	0.629	0.441	0.524
		8B	0.439	0.503	0.515	0.676	0.648	0.511	0.549
	Ensemble	(3B,1B)	0.416	0.498	0.503	0.665	0.633	0.442	0.526
		(8B,1B)	0.426	0.504	0.518	0.674	0.646	0.489	0.543
		(8B,3B)	0.430	0.504	0.515	0.675	0.647	0.490	0.544
	Contrast	(3B,1B)	0.435	0.502	0.509	0.674	0.639	0.438	0.533
		(8B,1B)	0.449	0.507	0.528	0.682	0.649	0.496	0.552
		(8B,3B)	0.445	0.509	0.530	0.683	0.652	0.496	<u>0.553</u>
Qwen	Single	0.5B	0.395	0.500	0.505	0.650	0.637	0.451	0.523
		3B	0.417	0.513	0.502	0.665	0.659	0.477	0.523
		7B	0.427	0.516	0.517	0.665	0.655	0.494	0.539
	Ensemble	(3B,0.5B)	0.414	0.510	0.511	0.663	0.656	0.479	0.546
		(7B,0.5B)	0.418	0.512	0.518	0.665	0.655	0.491	0.539
		(7B,3B)	0.425	0.518	0.518	0.671	0.663	0.497	0.543
	Contrast	(3B,0.5B)	0.428	0.517	0.519	0.675	0.666	0.482	0.549
		(7B,0.5B)	0.435	0.519	0.530	0.676	0.666	0.499	0.548
		(7B,3B)	0.435	0.521	0.531	0.676	0.666	0.499	**0.554**

Table 10: Pairwise Accuracy of evaluators with human scores in machine translation. **bold** represents best overall scores, while <u>underline</u> represents best scores within each model group (Baseline, LLaMA, Qwen). Overall, ContrastScore outperforms single and ensemble methods, as well as baseline metrics for both LLaMA and Qwen families.