On Learning to Summarize with Large Language Models as References

Yixin Liu[1] **Kejian Shi**[1] **Katherine S He**[1] **Longtian Ye**[1]
Alexander R. Fabbri[2] **Pengfei Liu**[3] **Dragomir Radev**[1] **Arman Cohan**[1,4]
[1]Yale University [2]Salesforce AI [3]Shanghai Jiao Tong University [4]Allen Institute for AI
{yixin.liu,arman.cohan}@yale.edu

Abstract

Recent studies have found that summaries generated by large language models (LLMs) are favored by human annotators over the original reference summaries in commonly used summarization datasets. Therefore, we study an LLM-as-reference learning setting for smaller text summarization models to investigate whether their performance can be substantially improved. To this end, we use LLMs as both oracle summary generators for standard supervised fine-tuning and oracle summary evaluators for efficient contrastive learning that leverages the LLMs' supervision signals. We conduct comprehensive experiments with source news articles and find that (1) summarization models trained under the LLM-as-reference setting achieve significant performance improvement in both LLM and human evaluations; (2) contrastive learning outperforms standard supervised fine-tuning under both low and high resource settings. Our experimental results also enable a meta-analysis of LLMs' summary evaluation capacities under a challenging setting, showing that LLMs are not well-aligned with human evaluators. Particularly, our expert human evaluation reveals remaining nuanced performance gaps between LLMs and our fine-tuned models, which LLMs fail to capture. Thus, we call for further studies into both the potential and challenges of using LLMs in summarization model development.

1 Introduction

Recent studies (Liu et al., 2023b; Zhang et al., 2024; Pu et al., 2023) have discovered that large language models (LLMs), like GPT-3.5 (Ouyang et al., 2022), can generate summaries that are preferred by human annotators when compared to ***reference summaries*** from widely used datasets, such as CNN/DailyMail (Nallapati et al., 2016) and XSum (Narayan et al., 2018), in a *reference-free* human evaluation setting. This quality issue of existing reference summaries effectively puts an upper bound on the performance of summarization models trained on them, which likely contributes to the performance gap between supervised summarization models, e.g., BART (Lewis et al., 2020), and LLMs as observed by related work (Goyal et al., 2022; Liang et al., 2023; Liu et al., 2023b; Zhang et al., 2024).

Therefore, we aim to investigate **whether smaller summarization models can be substantially improved with better references**. To this end, we study an ***LLM-as-reference*** distillation setting, where the LLMs are considered the reference or the gold-standard oracle for the summarization task. Specifically, we employ LLMs in the training of smaller text summarization models in two manners: (1) LLMs as the gold summary generator, where the model is trained with the LLM summary as the reference under the standard supervised fine-tuning; (2) LLMs as the gold summary evaluator, where LLM-based automatic evaluation methods (Fu et al., 2023; Liu et al., 2023a) are used as supervision signals for training techniques such as contrastive learning (Liu et al., 2022; Zhao et al., 2023b) and reinforcement learning (Paulus et al., 2018; Stiennon et al., 2020).

Using the source articles in the CNN/DailyMail dataset, we conduct comprehensive experiments for this LLM-as-reference setting, across proprietary and open-source LLMs under low and high resource conditions. The experimental results demonstrate that (1) LLM-generated summaries are better references for the smaller models than the original reference summaries, and (2) contrastive learning with LLMs as evaluators outperforms standard supervised fine-tuning. In particular, our best-performing fine-tuned BART checkpoint can outperform GPT-3.5 under GPT-4's evaluation (OpenAI, 2023). Meanwhile, under expert human evaluation, it can achieve similar or superior overall performance to GPT-3.5 in 50% of cases, while the original fine-tuned BART has a comparable rate of

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8644–8661

success in only 4% of cases.

To have a more comprehensive understanding of this LLM-as-reference setting, we then conduct a meta-analysis of the LLM-based evaluation methods by assessing their alignment level to the human evaluation. While these LLM-based methods achieve strong performance in existing meta-evaluation datasets consisting of summaries generated by supervised summarization systems (Fu et al., 2023; Liu et al., 2023a), we find that they do not correlate well with human evaluation when comparing close-performing systems in our setting. Particularly, although our fine-tuned model achieves better performance than LLMs under LLM-based evaluation, human evaluation reveals remaining nuanced performance gaps between our model and LLMs.

Our main contributions are two-fold: (1) We empirically demonstrate that the performance of smaller models can be substantially improved when trained using better references (LLMs) and learning methods (contrastive learning). (2) We perform a meta-analysis of LLM-based evaluation under a challenging scenario enabled by our task setting where LLMs need to compare summarization systems with close performance, which indicates that LLMs fail to align with human evaluation and capture nuanced performance differences.[1]

2 Methods

2.1 Preliminary

A neural abstractive summarization model g aims to generate a text sequence S that summarizes the information of a source document D: $S \leftarrow g(D)$. When g is an *auto-regressive* text generation model, it factorizes the probability of a candidate summary S given the source document D as

$$p_g(S|D) = \prod_{i=1}^{l_S} p_g(s_i|S_{<i}, D), \quad (1)$$

where s_i is the i-th token in S and s_0 is a special begin-of-sequence (BOS) token, $S_{<i}$ is the prefix-string of S before s_i, l_S is the length of S (without the BOS token), and p_g is a probability distribution parameterized by the summarization model g.

The standard training algorithm for g is Maximum Likelihood Estimation (MLE) with a single reference (gold standard) summary S^*. With Eq. 1, the MLE optimization on this example is equivalent to minimizing the following cross-entropy loss:

$$\mathcal{L}_{xent}(\theta) = -\log p_g(S^*|D;\theta), \quad (2)$$

where θ are the learnable parameters of g.

2.2 Large Language Models as References

Similar to Eq. 1, an auto-regressive LLM h defines a target distribution for text summarization:

$$p_h(S|D) = \prod_{i=1}^{l_S} p_h(s_i|S_{<i}, D), \quad (3)$$

which is different from the point-mass distribution defined by a single reference summary (Eq. 2). Consequently, the cross-entropy loss becomes

$$\mathcal{L}_{xent}(\theta;h) = -\sum_{S\in\mathcal{S}} p_h(S|D)\log p_g(S|D;\theta), \quad (4)$$

where $\mathcal{S}$ is the set of possible outputs (candidate summaries). This setting is coined ***sequence-level knowledge distillation*** by Kim and Rush (2016). In practice, computing Eq. 4 is intractable because $\mathcal{S}$ is infinite. Thus, we explore various approaches to approximate this learning objective.

2.2.1 LLMs as Gold Summary Generators

MLE Fine-tuning Our baseline method treats the greedy decoding results of the LLM h as the reference summaries and optimizes the summarization model g using MLE. The loss function becomes

$$\hat{\mathcal{L}}_{xent}(\theta;h) = -\log p_g(\hat{S}|D;\theta), \quad (5)$$

where $\hat{S}$ is the greedy decoding result of h:

$$\hat{s}_i = \arg\max_s p_h(s|\hat{S}_{<i}, D), \quad (6)$$

where s denotes a token in the vocabulary.

Contrastive Learning To improve the performance beyond MLE, we adopt a contrastive learning method, BRIO (Liu et al., 2022), for *reference-based* model training, which sets the following training objective: given two candidate summaries S_i, S_j, if S_i is better than S_j, the summarization model g should assign S_i a higher probability (Eq. 1). In more detail, this loss is defined with a set of candidate summaries $\mathcal{S}_c$, which is *descendingly sorted* by their ***similarity with the reference summary***, as measured by an automatic metric such as ROUGE (Lin, 2004). The summarization model g

[1]We release the scripts, training data, and model outputs at https://github.com/yixinL7/SumLLM.

is tasked with assigning a probability that is at least twice[2] as large to a better candidate:

$$\frac{p_g(S_i|D)}{p_g(S_j|D)} > 2(j-i), \forall i, j, i < j, \tag{7}$$

which corresponds to the following margin loss:

$$\begin{aligned}\mathcal{L}_{ctr}(\theta) = \sum_{S_i, S_j \in \mathcal{S}_c, i<j} & \max(0, \log p_g(S_j|D;\theta) \\ & - \log p_g(S_i|D;\theta) + \log 2(j-i)).\end{aligned} \tag{8}$$

Following Liu et al. (2022), we combine the cross-entropy loss (Eq. 5) with the contrastive loss as a multi-task loss:

$$\mathcal{L}_{mul}(\theta) = \hat{\mathcal{L}}_{xent}(\theta; h) + \alpha \mathcal{L}_{ctr}(\theta), \tag{9}$$

where α is the weight of the contrastive loss.

2.2.2 LLMs as Gold Summary Evaluators

Apart from the reference summaries, LLMs can also provide *reference-free* supervision signals for model training since they can be used to evaluate the quality of any candidate summary. As these LLM-based evaluation methods have shown superior performance than traditional metrics such as ROUGE (Fu et al., 2023; Liu et al., 2023a), we hypothesize that they can provide more accurate supervision that enables efficient training. Consequently, we expand the contrastive learning approach (Eq. 8) by using LLM-based evaluation to provide the gold ranking of the candidate summaries. We focus on two recent LLM-based evaluation methods: GPTScore (Fu et al., 2023) and an extended version of G-Eval (Liu et al., 2023a), which we coin GPTRank.

GPTScore for Summary Quality Evaluation The contrastive learning objective (Eq. 8) requires access to ground-truth candidate summary quality scores from the reference LLM. Therefore, we first adopt GPTScore (Fu et al., 2023) for the summary quality evaluation. Specifically, GPTScore interprets the length-normalized conditional log-probability of a candidate summary predicted by the reference LLM h as its quality score, i.e.,

$$\bar{p}_h(S|D) = \frac{\sum_{i=1}^{l_S} \log p_h(s_i|S_{<i}, D)}{l_S}. \tag{10}$$

Consequently, the set of candidate summaries $\mathcal{S}_c$ used in Eq. 8 is sorted based on the (normalized) target distribution (Eq. 3), such that for any $S_i, S_j \in \mathcal{S}_c, i < j, \bar{p}_h(S_i|D) > \bar{p}_h(S_j|D)$.

[2]We found that in practice the model training is insensitive to this value so we set it to a constant value for simplicity.

GPTRank for Summary Quality Evaluation Instead of leveraging the LLM predicted probability, recent work, e.g., G-Eval (Liu et al., 2023a), formulates the automatic evaluation as a text completion or infilling task for the LLMs, requiring them to provide a numerical quality score for an evaluation task. We extend this evaluation method, which we coin **GPTRank**, by requiring the LLM to provide a ***quality ranking*** to a list of different candidate summaries for the same source article. Moreover, since recent work (Liu et al., 2023a) has found that language models can benefit from a self-explaining stage for an evaluation task, we prompt the LLM to first generate an ***explanation*** before providing the actual ranking. The ranking is then used in contrastive learning (Eq. 8).

3 Learning with LLMs as References

We conduct experiments with both proprietary and open-source LLMs in the LLM-as-reference learning setting of smaller summarization models across low and high resource conditions and compare different training methods.

3.1 Learning under Low Resource Settings

Proprietary LLMs, such as GPT-4, can be more capable than open-source LLMs but are less cost-efficient. Therefore, we focus on a cost-effective low-resource setting using contrastive learning where the LLMs are used as both summary generators and evaluators.

3.1.1 Experimental Setting

Data Source We use mainly source articles from the CNN/DailyMail (CNNDM) dataset for our experiments, and 100 test examples are sampled for LLM-based and human evaluation. The LLMs are prompted to generate three-sentence summaries to approximate the original summary style with a 0 sampling temperate to approximate the greedy decoding process (Eq. 6).[3]

Training Details We choose BART as the smaller summarization model for fine-tuning because it is widely used and is relatively small with around 350 million parameters.[4] The fine-tuning process involves an MLE warm-up stage with around 10K GPT-3.5 summaries[5] and further

[3]Further information regarding the prompts and the process of generating LLM summaries can be found in Appendix A.1.

[4]https://huggingface.co/facebook/bart-large

[5]We used the checkpoint `gpt-3.5-turbo-0301` at https://platform.openai.com/docs/models/gpt-3-5.

System	LP	GS	R1	R2	Len.
GPT3D3	-22.62	-0.271	100.0	100.0	85.4
BART	-59.55	-0.789	46.85	24.38	79.0
GPT3D2	-41.21	-0.547	55.40	33.72	78.7
Alpaca	-44.82	-0.567	51.53	30.18	81.8
GPT-3.5	-45.12	-0.498	58.14	37.46	92.0
BART.GPT3D3	-36.13	-0.420	**59.50**	**40.70**	85.6
BRIO.GPT3D3	**-26.20**	**-0.318**	56.21	36.47	83.7

Table 1: Results with GPTScore. **LP** is the log-probability predicted by **GPT3D3**. **GS** is the GPTScore based on GPT3D3. **R1** and **R2** are the ROUGE1/2 F1 scores respectively. **Len.** is the average summary length. **BART.GPT3D3** is fine-tuned with MLE training while **BRIO.GPT3D3** is fine-tuned with contrastive learning.

MLE training or contrastive learning using a reference LLM with around 100 - 1000 training examples. During the experiments, we compare the model performance trained under MLE and contrastive learning, while making sure that the LLM API cost is similar under the two settings for fair comparison.

For contrastive learning, 8 candidate summaries are used on each training example, generated by MLE-finetuned model checkpoints. Further experimental details can be found in A.2.

Automatic Evaluation For *reference-based* evaluation, we report the ROUGE-1/2 F1 scores between the system outputs and the reference summaries generated by the reference LLM. For *reference-free* evaluation, we use either GPTScore (Fu et al., 2023) or GPTRank (§2.2.2). In particular, for GPTScore we report both the un-normalized and normalized sum of log-probability.

Baseline Models The following model's performance is compared: (1) GPT3D3, (2) the BART checkpoint fine-tuned on the original CNNDM dataset, (3) GPT3D2 (OpenAI's `text-davinci-002`), (4) a 7B Alpaca checkpoint,[6] (5) GPT-3.5 (OpenAI's `gpt-3.5-turbo-0301`).

3.1.2 Learning with GPTScore

For GPTScore, the reference LLM we choose is OpenAI's `text-davinci-003` (**GPT3D3**), since its API provides access to the predicted log-probability.[7] We report the model performance in Table 1, with the following observations:

(1) Compared with the original BART checkpoint, MLE training on reference summaries from LLMs can effectively improve the model performance as measured by either GPTScore or ROUGE.
(2) The model trained with contrastive learning (BRIO.GPT3D3) can achieve significantly better GPTScore than the model fine-tuned with MLE training (BART.GPT3D3), demonstrating the effectiveness of contrastive learning for approximating the target distribution of the reference LLM.
(3) BRIO.GPT3D3 can already achieve a similar GPTScore as the reference LLM (GPT3D3) itself while only being trained on 100 examples with contrastive learning, showing a promising path to further close the performance gap.

3.1.3 Learning with GPTRank

We now conduct experiments using GPTRank for model training and evaluation. The reference LLMs we choose are GPT-3.5 and GPT-4 (OpenAI, 2023) since they have shown state-of-the-art performance on summarization evaluation (Liu et al., 2023a).[8] To enable a more accurate evaluation, we choose GPT-3.5 as the baseline model and use the LLMs to conduct a *pairwise* comparison between different systems and GPT-3.5. To reduce the positional bias in LLM evaluation results as noted by Wang et al. (2023b), we evaluate each summary pair in *both* directions and take the average of results. In addition, we allow the LLMs to predict a tie between two summaries.[9]

In Figure 1, we report the pairwise comparison results of different models against GPT-3.5 under both GPT-3.5 and GPT-4's evaluation. We note:
(1) As in §3.1.2, using better references and contrastive learning helps the model to achieve better LLM-based evaluation results.
(2) Interestingly, GPT-3.5 prefers both BRIO.GPT-3.5 and BRIO.GPT-4 over its own outputs in the pairwise comparison, suggesting that contrastive learning can efficiently optimize the summarization model for a specific evaluation metric.
(3) LLM-based evaluation results vary across different LLMs. For example, while GPT-3.5 prefers BRIO.GPT-4 over itself, GPT-4 prefers GPT-3.5.
(4) BRIO.GPT-3.5 can outperform BART.GPT-4 despite the fact that BRIO.GPT-3.5 is trained with a reference LLM that is supposedly weaker, indicating the advantage of contrastive learning.

[6] https://github.com/tatsu-lab/stanford_alpaca
[7] We note that the more recent OpenAI models, such as GPT-4, do not provide log-probability of the input tokens.
[8] We use the `GPT-4-0314` version: https://platform.openai.com/docs/models/GPT-4.
[9] The prompt templates are shown in Appendix A.3.

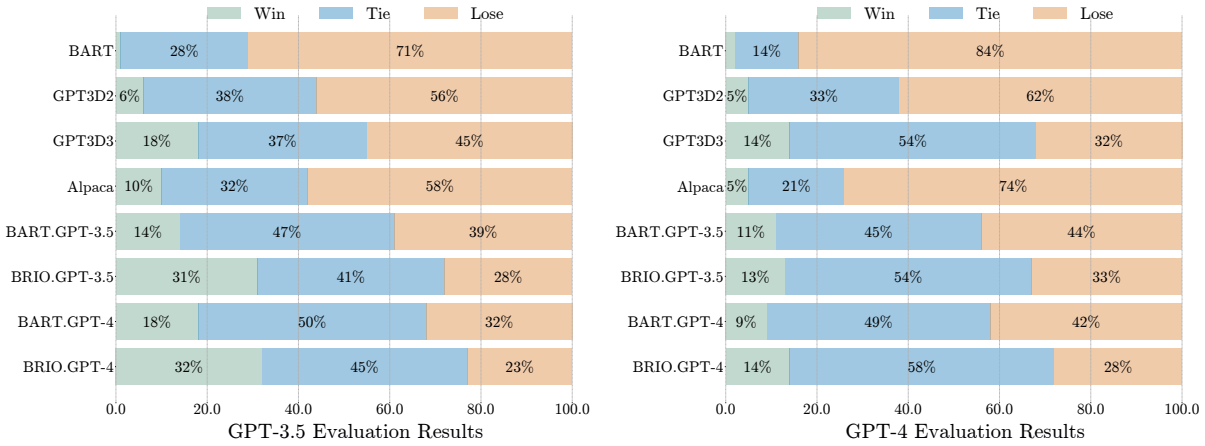

Figure 1: Pairwise comparison (GPTRank) results of different models *against GPT-3.5* under GPT-3.5's evaluation (left) and GPT-4's evaluation (right). BART.GPT-3.5 and BART.GPT-4 are fine-tuned with MLE training and GPT-3.5/GPT-4 as the reference, BRIO.GPT-3.5 and BRIO.GPT-4 are fine-tuned with contrastive learning.

System	GPT-3.5		GPT-4		Length
	R1	R2	R1	R2	
GPT-3.5	63.43	44.09	100.0	100.0	92.0
GPT-4	100.0	100.0	63.43	44.09	90.0
BART	50.83	29.47	50.54	29.31	79.0
GPT3D2	55.17	33.23	55.34	33.31	78.7
GPT3D3	56.12	34.72	58.14	37.46	85.4
Alpaca	54.77	33.23	53.41	31.48	81.8
BART.ChatGPT	59.52	40.45	62.04	**43.76**	94.1
BRIO.ChatGPT	57.56	35.74	61.40	40.74	93.1
BART.GPT4	**63.22**	**44.70**	62.08	43.55	91.8
BRIO.GPT4	58.65	37.57	**62.79**	43.65	92.8

Table 2: Reference-based evaluation results of GPTRank-based training. GPT-3.5 and GPT-4's summaries are used as the references. **R1** and **R2** are the ROUGE1/2 F1 scores respectively. **Len.** is the average summary length.

The reference-based evaluation results can be found in Table 2.

3.1.4 Comparative Study

We investigate the generalization ability of our training method regarding the choice of the backbone model and the data format.

Experiments with FLAN-T5 We repeat the experiment in §3.1.3 but use a three billion FLAN-T5 (Chung et al., 2022) model[10] as the backbone model. Results in Figure 2 suggest that the choice of training algorithms can be more important than the model size for model performance, as BRIO.GPT-4 can outperform T5.GPT-4. The FLAN-T5 checkpoint trained with contrastive learning, T5BRIO.GPT-4, achieves a strong performance. However, we note that its summaries are significantly longer than those of other systems, which makes the result more difficult to interpret as recent work has found a strong correlation between the summary rating and length in both human and LLM-based summarization evaluation (Liu et al., 2023b; Rajani et al., 2023). Further discussion is in Appendix A.4.

[10] https://huggingface.co/google/flan-t5-xl

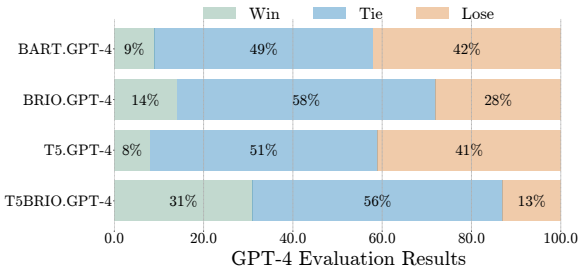

Figure 2: Results of T5 and BART models compared against GPT-3.5 under GPT-4's evaluation. BART.GPT-4 and T5.GPT-4 are MLE fine-tuned, BRIO.GPT-4 and T5BRIO.GPT-4 are fine-tuned with contrastive learning.

Experiments on XSum We conduct experiments on XSum (Narayan et al., 2018), another commonly used dataset. We follow the original XSum data format by having the models generate one-sentence summaries. The experimental settings are similar to those in §3.1.1 & §3.1.3 and more details are in Appendix A.5. The results in Figure 3 show a similar trend in that training with better references helps to improve model performance.

3.2 Learning under High Resource Settings

Open-source LLMs provide easier access than proprietary LLMs, however, their performance can

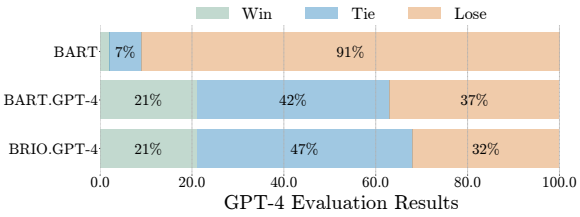

Figure 3: Results on XSum dataset. Different models are compared against GPT-3.5 under GPT-4's evaluation. BART.GPT-4 is fine-tuned with MLE training while BRIO.GPT-4 is fine-tuned with contrastive learning.

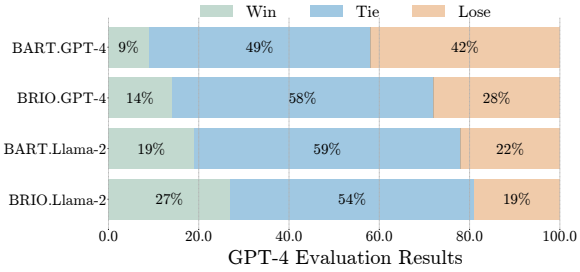

Figure 4: Model performance under low and high resource settings. Models are compared against GPT-3.5 under GPT-4's evaluation. The models trained with Llama-2 are under high resource settings and the models trained with GPT-4 are under low resource settings.

be worse, especially for complicated evaluation tasks (Liu et al., 2024). Therefore, we now investigate a high resource setting, where open-source LLMs are used as summary *generators* only to obtain a large number of reference summaries.

3.2.1 Experimental Setting

We use the Llama-2 7B Chat model (Touvron et al., 2023) to generate around 280K reference summaries for the model training. We fine-tune the BART model using both MLE training and *reference-based* contrastive learning, where the candidate summaries in contrastive learning are ranked by a reference-based automatic evaluation metric. For contrastive learning, the candidate summaries are ranked based on their content similarity to the Llama-2 summaries, rather than using an LLM-based evaluation method. A recently introduced metric, A3CU (Liu et al., 2023c), is used to measure content similarity, which has better performance than traditional metrics like ROUGE.

3.2.2 Results

The main evaluation results are reported in Figure 4, where the models are compared against GPT-3.5 under GPT-4's evaluation. Reference-based evaluation results are in Table 3. We found that:

System	ROUGE-1	ROUGE-2	A3CU	Length
Llama-2	100.00	100.00	95.54	93.9
BART.Llama-2	58.96	**37.39**	50.71	92.1
BRIO.Llama-2	**60.55**	37.38	**53.25**	92.5

Table 3: Reference-based evaluation results under the high resource setting on CNNDM. **BART.Llama-2** is fine-tuned with MLE training while **BRIO.Llama-2** is fine-tuned with contrastive learning.

(1) Models trained in high resource settings can outperform those in low resource settings, highlighting the benefit of abundant training data.
(2) The model trained with contrastive learning, BRIO.Llama-2, outperforms GPT-3.5 under GPT-4's evaluation, indicating that a smaller summarization model has the capacity to reach LLM-level performance *under LLM-based evaluation*.

4 Human Evaluation and Meta-Analysis

In §3 we have demonstrated that smaller summarization models that are trained with better references can achieve on-par or even better performance than LLMs under ***LLM-based evaluation***. However, the alignment between LLM and human evaluation still requires examination. Therefore, we first conduct a human evaluation comparing the performance of models in §3, then perform a meta-analysis regarding the LLM-based evaluation.

4.1 Human Evaluation Collection

Evaluation Design We formulate the human evaluation as a summary pairwise comparison task.[11] The summary pairs are compared on three aspects: (1) salience, (2) coherence, and (3) overall preference/quality, where the annotators are required to choose which summary is better (ties are allowed). The detailed aspect definitions are in Appendix B.1.

Crowd-Annotation Collection We use Amazon Mechanical Turk[12] (MTurk) for the crowd-annotation collection. Each data example is annotated by three annotators who are given two minutes for one task and compensated accordingly. The participated crowd-annotators need to pass related qualification tests and have previous experience in evaluating summary quality. We choose three system pairs for the collection on 100 test examples, where GPT-3.5 is the baseline

[11] The summary pairs are randomly shuffled.
[12] https://www.mturk.com/

Group	System	Salience	Coherence	Overall
1	GPT-3.5	83	84	87
	BART	26	34	20
2	GPT-3.5	68	68	62
	BART.GPT-4	45	63	41
3	GPT-3.5	60	65	61
	BRIO.GPT-4	50	56	39

Table 4: Crowd-annotations conducted on 3 groups of system pairs on 100 examples. The count of *wins* for each system is reported, including ties as dual wins.

LLM, and three BART checkpoints from §3.1.3 are compared against GPT-3.5. To check the inter-annotator agreement, we calculate the Krippendorff's alpha (Krippendorff, 2011) with MASI distance (Passonneau, 2006) following Goyal et al. (2022). We found the average agreement to be 0.064, close to the agreement (0.05) reported by Goyal et al. (2022) for similar evaluation settings.

Expert Evaluation The low agreement of crowd-annotation raises concerns about annotation quality. Therefore, we (the co-authors) conducted a careful expert evaluation to better understand this phenomenon and provide more trustworthy evaluation results. We select 50 test examples to perform a pairwise comparison on three crowd-evaluated system groups and four additional groups. We found the average agreement to be 0.044 among the expert annotators after a careful annotation, which re-confirms the hypotheses made in the related work (Goyal et al., 2022; Zhang et al., 2024) regarding the inherent subjectivity of summarization evaluation especially when comparing summaries with similar quality. Besides, the experts agree with each other 55% of the time, similar to the agreement level (65%) in recent work (Rafailov et al., 2023). We provide further analyses in Appendix B.2, which shows two main scenarios: (1) cases where the annotators unanimously favor LLM summaries; (2) cases where both LLM and smaller LM have good performance, resulting in different annotator preferences. While higher agreement might be achieved with a more constrained evaluation protocol, we believe such a higher agreement can be "artificial" and cannot reflect the diverse distribution of human preferences.

4.2 Result Analysis

The crowd-annotation and expert-evaluation results are in Table 4 & 5 respectively. We note:

Group	System	Salience	Coherence	Overall
1	GPT-3.5	44	49	49
	BART	10	4	2
2	GPT-3.5	40	35	35
	BART.GPT-4	22	24	18
3	GPT-3.5	*32*	39	33
	BRIO.GPT-4	*29*	24	21
4	BART.GPT-4	22	26	17
	BRIO.GPT-4	41	36	39
5	GPT-3.5	36	38	38
	BART.Llama-2	19	28	18
6	GPT-3.5	34	33	*28*
	BRIO.Llama-2	22	33	*25*
7	BART.Llama-2	25	28	22
	BRIO.Llama-2	31	36	31

Table 5: Expert evaluation conducted on 7 groups of system pairs on 50 examples. The count of *wins* for each system is reported, including ties as dual wins.

(1) The models trained with the LLMs as references can outperform the BART checkpoint trained on the original CNNDM dataset by a large margin, showing the importance of better references.
(2) When under a direct comparison in expert evaluation, BRIO.GPT-4/Llama-2 can outperform BART.GPT-4/Llama-2 on three aspects, demonstrating the effectiveness of contrastive learning.
(3) While the smaller models cannot yet outperform GPT-3.5, the performance gap is smaller, with BRIO.GPT-4 achieving similar salience scores and BRIO.Llama-2 achieving similar overall scores.

4.3 Meta-Analysis of LLM-based Evaluation

BRIO.Llama-2 cannot outperform GPT-3.5 under human evaluation, even though they are favored by the evaluation methods based on GPT-4 (Figure 4). Therefore, we further investigate this discrepancy between human and LLM-based evaluation.

Human-LLM Alignment We use the expert evaluation results to evaluate the performance of LLM-based evaluation as well as the crowd-annotation, by computing their agreements with the majority vote of expert evaluation on evaluation group 2 and 3 in Table 4 & 5. Apart from GPTScore and GPTRank, we also compare the performance of G-Eval (Liu et al., 2023a). The prompts used for GPTRank and G-Eval are aspect-specific. More details are in Appendix B.3. The agreements are reported in Table 6, showing the following trends:
(1) LLM-based evaluation methods vary in perfor-

	Salience	Coherence	Overall
Crowd-Individual	0.189	0.061	0.062
Crowd-Major-Voting	0.241	0.116	0.166
G-EVAL-3.5	-0.214	-0.168	-0.114
G-EVAL-4	-0.082	-0.143	-0.019
GPTScore	-0.115	-0.021	-0.029
GPT-3.5Rank	0.036	-0.034	0.018
GPT-4Rank	0.191	0.051	0.105

Table 6: Performance comparison of LLM-based evaluation and crowd-annotation in terms of their agreements with expert evaluation. G-EVAL-3.5 and G-Eval-4 are G-Eval scores based on GPT-3.5 and GPT-4 respectively. GPTScore is based on GPT3D3. GPT-3.5Rank and GPT-4Rank are two versions of the GPTRank.

mance, and GPT-4 outperforms GPT-3.5.
(2) GPT-4 with GPTRank can already outperform the performance of individual crowd-workers, while majority voting from crowd-workers still achieves the highest agreement.

LLM Positional Bias and Self-Inconsistency The GPTRank evaluation protocol performs pairwise comparisons, with which the LLMs can have a positional bias favoring either the first output or the second in the comparison (Wang et al., 2023b). We observe that both GPT-3.5 and GPT-4 have similar positional biases in our study, which lead to a self-inconsistency – the LLMs can favor different outputs when the output order is flipped in the pairwise comparison. In Table 7, we highlight this positional bias and the self-inconsistency rate of the LLMs. Both LLMs have a bias toward the second output, and GPT-4 gave inconsistency decisions around 50% of the time when the order of two outputs is flipped.

Discussion Our meta-analysis reveals that LLMs are *noisy* summary evaluators because of the relatively low alignment level with human evaluation and the self-inconsistency. We note that they are still better references for smaller model training compared with the original reference summaries (§4.2). However, we advise against using *only* LLMs for system evaluation, particularly when comparing closely performing systems.

5 Related Work

Training Methods of Text Generation Models The standard MLE training of text generation models has two major limitations: (1) a discrepancy between the training objective, i.e., the cross-entropy loss, and the evaluation criteria (e.g., ROUGE); (2) a discrepancy between the teacher-forcing (Williams and Zipser, 1989) training manner and auto-regressive generation behavior during evaluation, which is known as the *exposure bias* (Bengio et al., 2015; Ranzato et al., 2016). As a result, training methods beyond MLE have been proposed to address these two limitations. Among them a family of methods is based on *reinforcement learning* (RL), which can optimize the text generation model toward a specific reward function (Ranzato et al., 2016; Bahdanau et al., 2017; Li et al., 2016; Paulus et al., 2018; Li et al., 2019; Stiennon et al., 2020; Pang and He, 2021). Apart from RL, training methods based on supervised learning have also been developed, such as Minimum Risk Training (Shen et al., 2016; Wieting et al., 2019), targeting a sequence-level optimization with various reward signals (Wiseman and Rush, 2016; Edunov et al., 2018). More recently, *contrastive learning* (Hadsell et al., 2006) has also been adopted, which enhances the model ability by requiring the model to differentiate positive (good) and negative (bad) examples (Yang et al., 2019; Pan et al., 2021; Cao and Wang, 2021; Liu and Liu, 2021; Sun and Li, 2021; Zhao et al., 2023b; Zhang et al., 2022). The latest work along this path has explored using contrastive learning to align LLMs with human feedback (Yuan et al., 2023; Zhao et al., 2023a; Rafailov et al., 2023), an alternative to reinforcement learning with human feedback (Stiennon et al., 2020; Ouyang et al., 2022).

	Output 1	Output 2	Inconsistency
GPT-3.5	33.70%	66.30%	45.16%
GPT-4	26.44%	73.56%	50.96%

Table 7: Positional bias and self-inconsistency rate of GPTRank with GPT-3.5 and GPT-4 as the backbone models (ties are ignored). Both LLMs have a bias toward the second output in pairwise comparisons.

LLM-based Automatic Evaluation Recent work has explored using LLMs for automatic NLP evaluation. GPTScore (Fu et al., 2023) leverages the LLM-predicted probability of text sequences as the quality score. On the other hand, a line of work (Chiang and Lee, 2023; Gao et al., 2023; Chen et al., 2023; Wang et al., 2023a; Luo et al., 2023), e.g., G-Eval (Liu et al., 2023a), proposes evaluation methods that use LLMs to perform text completion tasks, such as predicting the answer of a Likert scale evaluation or pairwise comparison.

Notably, several of these studies (Fu et al., 2023; Liu et al., 2023a; Zheng et al., 2023; Dubois et al., 2023; Gao et al., 2023; Chen et al., 2023; Wang et al., 2023a) all evaluate the LLM-based evaluation methods on SummEval (Fabbri et al., 2021), a summarization human evaluation benchmark, and found that LLM-based evaluation has a higher correlation with human judgments than previous methods such as ROUGE or BERTScore (Zhang* et al., 2020). Apart from summarization evaluation, LLM-based evaluation has also been used in text classification tasks (Gilardi et al., 2023) and for reward design for RL agents (Kwon et al., 2023).

LLM Distillation and LLM-based Data Augmentation To improve the performance of smaller NLP models, related work has proposed methods of distilling LLMs and using LLMs for data augmentation (Wang et al., 2021; Ding et al., 2023; Kang et al., 2023). Specifically, a line of work (Shridhar et al., 2023; Li et al., 2022; Hsieh et al., 2023) uses LLMs to generate both final answers and task-related descriptions for training smaller models on reasoning tasks, and Orca (Mukherjee et al., 2023) extends this method for LLM distillation by training smaller models on the LLM-generated explanations. Regarding text summarization, Wang et al. (2021) introduces using GPT-3 (Brown et al., 2020) to generate reference summaries while Gekhman et al. (2023) proposes using LLMs to annotate the summary factual consistency (Maynez et al., 2020) for the training of smaller evaluation models.

6 Conclusion

In this work, we study a learning setting of text summarization models where the LLMs are set to be the reference. For this setting, we leverage the LLM-based evaluation methods to guide the model training through contrastive learning and empirically demonstrate the efficiency and effectiveness of our methods. Furthermore, we conduct human evaluation and meta-analysis regarding the reliability of LLM-based evaluation, which reveals its benefits as better training references and its limitations in terms of the alignment with human evaluation. We believe our findings shed light on the direction of reliably applying the LLMs to the entire development loop (i.e., training-validation-evaluation) of smaller, task-specific NLP models, which has the potential to provide a balance between model performance and computational cost.

7 Limitations

The LLM-based evaluation results we reported are from OpenAI's APIs, which are subject to change. Therefore, the reproducibility of our experiments is limited. To mitigate this problem, we will release the training data, model outputs, and LLM and human evaluation results to facilitate future work.

Both the LLM-based and human evaluations we conducted can be resource-intensive, requiring substantial time and budget. As a result, we try to find a balance between the reliability of the evaluation result and the constraints of time and budget when selecting the sample size we used for evaluation. An evaluation at a larger scale is likely to yield more reliable results, which we leave for more dedicated future work in this direction. The resource constraints also led us to use news summarization as a case study, leaving other summarization task scenarios for future work.

We chose not to include summary factual consistency as an individual quality aspect in human evaluation and the meta-analysis of LLM-based evaluation. Related work (Tang et al., 2023; Zhang et al., 2024) has found that the factual error rate is low on CNNDM dataset, especially for LLM summaries. During our expert evaluation, the authors also did not observe significant flaws in factual consistency. As a result, it would require a much larger sample size for an evaluation of factual consistency in order to understand the error patterns, which is out of the scope of this work. However, we believe that such an evaluation is important for better understanding the summary quality of LLMs and LLM-supervised models, and we hope that the outcome of this work (e.g., the system outputs) can be a helpful resource for future work on this topic.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. We are grateful for the compute support provided by the Google TRC program.

References

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. An actor-critic algorithm for sequence prediction. In *International Conference on Learning Representations*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence

prediction with recurrent neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1171–1179, Cambridge, MA, USA. MIT Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.

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.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Classical structured prediction losses for sequence to sequence learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 355–364, New Orleans, Louisiana. 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*, abs/2302.04166.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *ArXiv*, abs/2304.02554.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina McMillan-Major, Anna Shvets, Ashish Upadhyay, Bingsheng Yao, Bryan Wilie, Chandra Bhagavatula, Chaobin You, Craig

Thomson, Cristina Garbacea, Dakuo Wang, Daniel Deutsch, Deyi Xiong, Di Jin, Dimitra Gkatzia, Dragomir R. Radev, Elizabeth Clark, Esin Durmus, Faisal Ladhak, Filip Ginter, Genta Indra Winata, Hendrik Strobelt, Hiroaki Hayashi, Jekaterina Novikova, Jenna Kanerva, Jenny Chim, Jiawei Zhou, Jordan Clive, Joshua Maynez, João Sedoc, Juraj Juraska, Kaustubh D. Dhole, Khyathi Raghavi Chandu, Leonardo F. R. Ribeiro, Lewis Tunstall, Li Zhang, Mahima Pushkarna, Mathias Creutz, Michael White, Mihir Kale, Moussa Kamal Eddine, Nico Daheim, Nishant Subramani, Ondrej Dusek, Paul Pu Liang, Pawan Sasanka Ammanamanchi, Qinqin Zhu, Ratish Puduppully, Reno Kriz, Rifat Shahriyar, Ronald Cardenas, Saad Mahamood, Salomey Osei, Samuel Cahyawijaya, Sanja vStajner, Sébastien Montella, Shailza, Shailza Jolly, Simon Mille, Tahmid Hasan, Tianhao Shen, Tosin P. Adewumi, Vikas Raunak, Vipul Raheja, Vitaly Nikolaev, Vivian Tsai, Yacine Jernite, Yi Xu, Yisi Sang, Yixin Liu, and Yufang Hou. 2022. Gemv2: Multilingual nlg benchmarking in a single line of code. In *Conference on Empirical Methods in Natural Language Processing*.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *ArXiv*, abs/2303.15056.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *ArXiv*, abs/2209.12356.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, page 1735–1742, USA. IEEE Computer Society.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Junmo Kang, Wei Xu, and Alan Ritter. 2023. Distill or annotate? cost-efficient fine-tuning of compact models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11100–11119, Toronto, Canada. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. 2023. Reward design with language models. In *The Eleventh International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jingu Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022. Explanations from large language models make small reasoners better. *ArXiv*, abs/2210.06726.

Siyao Li, Deren Lei, Pengda Qin, and William Yang Wang. 2019. Deep reinforcement learning with distributional semantic rewards for abstractive summarization. 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 6038–6044, Hong Kong, China. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar,

Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. 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.

Yixin Liu, Alexander R Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq R. Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir R. Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*.

Yixin Liu, Alexander R Fabbri, Yilun Zhao, Pengfei Liu, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023c. Towards interpretable and efficient automatic reference-based summarization evaluation. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. ChatGPT as a factual inconsistency evaluator for text summarization. *ArXiv*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online. Association for Computational Linguistics.

Richard Yuanzhe Pang and He He. 2021. Text generation by learning from demonstrations. In *International Conference on Learning Representations*.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *ArXiv*, abs/2309.09558.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nazneen Rajani, Nathan Lambert, Sheon Han, Jean Wang, Osvald Nitski, Edward Beeching, and Lewis Tunstall. 2023. Can foundation models label data like humans? *Hugging Face Blog*. Https://huggingface.co/blog/llm-v-human-data.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Shichao Sun and Wenjie Li. 2021. Alleviating exposure bias via contrastive learning for abstractive text summarization. *CoRR*, abs/2108.11846.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

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

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

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

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Ronald J. Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.*, 1(2):270–280.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, Austin, Texas. Association for Computational Linguistics.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196, Florence, Italy. Association for Computational Linguistics.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. RRHF: Rank responses to align language models with human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Xingxing Zhang, Yiran Liu, Xun Wang, Pengcheng He, Yang Yu, Si-Qing Chen, Wayne Xiong, and Furu Wei.

2022. Momentum calibration for text generation. *ArXiv*, abs/2212.04257.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. 2023a. SLiC-HF: Sequence likelihood calibration with human feedback. *ArXiv*.

Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. 2023b. Calibrating sequence likelihood improves conditional language generation. In *The Eleventh 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, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Experimental Details

A.1 LLM Summary Generation

In §3, we use the following prompt to generate the LLM summaries:

> Article: {{Article}}
>
> Summarize the above article in three sentences.
>
> Summary:

Since text summarization is a conditional generation task that requires high accuracy, we set the sampling temperature to 0 to ensure a more accurate and deterministic behavior of the LLMs.

A.2 Additional Experimental Details

For the experiments we conducted in §3, the specific settings can be found in the training and configuration scripts we released. All the experiments are performed using 1 - 4 NVIDIA A6000 GPUs with 48GB memory. The experiments in low-resource settings take around 4 hours to converge, while the ones in high-resource settings take around 30 hours.

Implementation Details of Contrastive Learning In practice, we observe that the magnitude of the log-probability in Eq. 8 is highly dependent on the length of the candidate summaries. Therefore, we introduce a modification to Eq. 8 based on the length-normalized log-probability $\bar{p}_g$:

$$\bar{p}_g(S|D) = \frac{\sum_{i=1}^{l_S} \log p_g(s_i|S_{<i}, D)}{l_S}, \tag{11}$$

and Eq. 8 is changed to

$$\hat{\mathcal{L}}_{ctr}(\theta) = \sum_{S_i, S_j \in \mathcal{S}_c, i<j} \max(0, \bar{p}_g(S_j|D;\theta) - \bar{p}_g(S_i|D;\theta) + \frac{1}{\lambda}\log 2(j-i)), \tag{12}$$

where λ is a scaling factor approximating the average summary length, which is set to the average length of the reference summaries. As for the weight of contrastive loss (α) in Eq. 9, we performed a grid search to find the correct configuration, which is set to 100 for the low resource setting and 10 for the high resource setting.

Candidate Generation for Contrastive Learning The contrastive learning (Eq. 12) requires a list of candidate summaries. To generate the summaries, we use the LLMs fine-tuned with MLE training and leverage diverse beam search as the sampling algorithm. For training with GPTScore (§3.1.2), we set 8 beam groups and 4 beams in each group, and pick one candidate from each group as the final candidate. As for training with GPTRank (§3.1.3), we choose a larger search space with 32 beam groups, and pick 8 candidate outputs for the resulting 32 initial candidates by minimizing the similarity between them. This is to ensure the diverse quality of candidate summaries used with GPTRank. For the high resource training setting (§3.2), we follow a similar approach but use nucleus sampling (Holtzman et al., 2020) instead of beam search for candidate generation to ensure score diversity from reference-based evaluation.

A.3 Prompt Templates for GPTRank

In §3.1.3, we use the following prompt template for GPTRank with *list-wise* comparison that is used for contrastive learning:

> You will be given a news article along with a list of summaries numbered as follows: 1. Summary 1, 2. Summary 2, and so on. Please evaluate and rank the summaries in descending order of their quality. First you will give an explanation of your ranking, then you will provide the ranking itself. Please refer to the example below for the format of your response.
>
> Example Response:
>
> Explanation: "Your explanation of the ranking"

System	ROUGE-1	ROUGE-2	Length
GPT-4	100.00	100.00	90.0
BART.GPT4	63.22	44.70	91.8
BRIO.GPT4	58.65	37.57	92.8
T5.GPT4	62.99	44.31	93.9
T5BRIO.GPT4	58.44	36.69	108.4

Table 8: Performance comparison of FLAN-T5 and BART as the fine-tuned backbone model on CNNDM. **GPT-4** is the reference LLM. **BART.GPT-4** and **T5.GPT-4** are fine-tuned with MLE training while **BRIO.GPT-4** and **T5BRIO.GPT-4** are fine-tuned with contrastive learning.

> Ranking: "The ranking, e.g., 4, 2, 7, 3, 5, 6, 8, 1"
>
> Here are the actual article and summaries:
>
> Article: {{Article}}
>
> Summaries:
>
> 1. {{Summary 1}}
>
> 2. {{Summary 2}}
>
> 3. {{Summary 3}}
>
> 4. {{Summary 4}}

For *pairwise* comparison that is used for model evaluation, the prompt template is as follows:

> You will be given a news article along with two summaries. Please compare the quality of these two summaries and pick the one that is better (there can be a tie). First you will give an explanation of your decision then you will provide your decision in the format of 1 or 2 or tie.
>
> Response format:
>
> Explanation: "Your explanation here".
>
> Decision: 1 or 2 or tie.
>
> Here's the article:
>
> {{Article}}
>
> Summary 1:
>
> {{Summary 1}}
>
> Summary 2:
>
> {{Summary 2}}

System	ROUGE-1	ROUGE-2	Length
GPT-4	100.00	100.00	42.8
BART	31.90	12.35	21.8
BART.GPT4	56.45	35.82	42.6
BRIO.GPT4	57.08	36.55	42.9

Table 9: Reference-based evaluation results on XSum. **GPT-4** is the reference LLM. **BART.GPT-4** is fine-tuned with MLE training while **BRIO.GPT-4** is fine-tuned with contrastive learning.

A.4 Analysis of Experiments with FLAN-T5

The reference-based evaluation results of the FLAN-T5 fine-tuning (§3.1.4) are reported in Table 8. We found that the FLAN-T5 checkpoint fine-tuned with contrastive learning, T5BRIO.GPT-4, tends to generate longer summaries. We tried to control the summary length by adjusting the length penalty used during beam search, but found that the length difference was still present. On the other hand, we are able to control the summary length of BRIO.GPT-4. We hypothesize this is because FLAN-T5 can learn the preference of LLM-based evaluation more efficiently, which exhibits a preference for longer outputs (Rajani et al., 2023). However, we note that the length preference is not the only factor affecting the LLM-based evaluation, since we only found a moderate Spearman's correlation (0.2366) between the summary length and the ranking of GPTRank. Moreover, out of 20 summary pairs where the GPT-3.5 summary is longer than the T5BRIO.GPT-4 summary, T5BRIO.GPT-4 still wins 9 times as evaluated by GPTRank based on GPT-4.

A.5 Experimental Details on XSum

The experimental setting on XSum (§3.1.4) is similar to the setting on CNNDM (§3.1.1). Specifically, at the warm-start stage we generate around 10K summaries using GPT-3.5 to fine-tune the BART checkpoint pre-trained on the original XSum dataset (`https://huggingface.co/facebook/bart-large-xsum`). Then, we generate 1K summaries using GPT-4 and continue fine-tuning the checkpoint with MLE training, resulting in the checkpoint named BART.GPT-4. As for contrastive learning, we use GPTRank with GPT-4 to generate 500 examples, and the checkpoint from the warm-start stage is fine-tuned to a new checkpoint, BRIO.GPT-4. In Table 9, we report the reference-based evaluation results.

B Human and Meta Evaluation Details

B.1 Definition of Summary Quality Aspects

We adopt the definition of the different quality aspects in §4.1 from the previous work (Fabbri et al., 2021; Gehrmann et al., 2021, 2022) as follows:
(1) Salience: "This rating measures how well the summary captures the key points of the news article. Consider whether all and only the important information are included in the summary."
(2) Coherence: "This rating measures whether the summary is presented in a clear, well-structured, logical, and meaningful way."
(3) Overall Preference/Quality: "This rating measures how much you like the summary."

B.2 Expert Evaluation Examples

We present expert-annotated examples (§4.2) in Table 10, with two main scenarios: (1) cases where the annotators unanimously favor LLM summaries; (2) cases where both LLM and smaller LM have good performance, resulting in different annotator preferences. For those examples on which the annotators have different preferences for the *overall* summary quality, we provide their explanations written *after* the evaluation below, as a case study of the inherent subjectivity of summarization human evaluation.

Example 3

Annotator 1: I selected the BRIO.GPT-4 summary because it conveys the same information as the GPT-3.5 summary more concisely. In the sentence about the nation being split on whether Charles should become king, it felt a little repetitive for the GPT-3.5 summary to use "become king" and "ascend to the throne" in the same sentence.

Annotator 2: The summaries are nearly identical. Both summaries capture almost the same level of important information. However, I prefer GPT-3.5's summary because it reiterates the fact that public opinion is expressed through a poll, which adds grounding and enhances objectivity to the statements.

Annotator 3: These two summaries essentially convey the same information and are almost equivalent in clarity and brevity. I chose the first summary because I personally preferred the way it started with ("A poll conducted by . . . ") which gives me the source of information that I value more.

Example 5

Annotator 1: I selected the GPT-3.5 summary because I found it slightly easier to follow along. The first two sentences both start with statements from Sheriff Hodgson, creating a clear structure and line of reasoning. The last sentence of the BRIO.GPT-4 summary ends with "Hodgson said", which makes sense but does not contextualize the statement until the very end of the summary.

Annotator 2: Both summaries are of good quality, making it a difficult decision for me. Despite its lower coherence and fluency, I lean towards preferring the summary generated by BRIO.GPT-4 due to its conciseness. The summary from GPT-3.5 includes additional details such as the mention of "maximum-security Souza-Baranowski state prison" and provides extra descriptions regarding Hernandez's charm, which I personally find redundant.

Annotator 3: Both summaries are of good quality, in terms of salience and coherence. The first one provides additional context regarding the final outcome of Aaron Hernandez's sentence, which I found to be more informative than the second summary.

Example 6

Annotator 1: I selected the GPT-3.5 summary because its first and last sentences were slightly more cohesive. The first sentence mentions that Nike has "faced criticism" and the last sentence mentions that Nike's vice president "defended the decision" in a statement – a direct response to the criticism. On the other hand, the BRIO.GPT-4 summary starts by stating that Nike has "defended their new kits" but does not include any comments or defense from Nike.

Annotator 2: I find it challenging to determine a clear winner between the two summaries as they both possess merits and weaknesses. The summary generated by GPT-3.5 mentions the key figure, Vice President Charlie Brooks, who defended the design of the kits, but it overlooks any feedback from the team. On the other hand, the summary generated by BRIO.GPT-4 fails to mention Charlie Brooks but includes the players' reactions, although it does so in a slightly redundant manner by quoting the midfielder Tobin Heath. In my opinion, the advantages and disadvantages of each summary are relatively balanced, leading me to consider them on equal footing.

Annotator 3: The second summary contains more balanced perspectives from both the critics and the national team itself. It also follows an organized structure from introducing the criticism to the reaction of the team. However, the first summary appears to be more straightforward and neutral,

without individual responses and words such as "proud" which could create certain ambiguity for me. As such, I chose tie because they both match well with the purpose of a summary.

Example 7

Annotator 1: I selected the BRIO.GPT-4 summary because the last sentence provides specific, key information about Liana Barrientos' legal charges that are not mentioned in the GPT-3.5 summary, which provides important context for the case's current status.

Annotator 2: The quality of both summaries is high. GPT-3.5 mentions that all of Barrientos's marriages took place in New York State starting from 1999, which is a detail not mentioned in BRIO.GPT-4. While BRIO.GPT-4 does mention the crucial fact that some of Barrientos's partners could potentially pose threats to homeland security, I found the last sentence to be somewhat grammatically awkward. Therefore, I ultimately gave the edge to GPT-3.5.

Annotator 3: I prefer the second summary because it provides more essential details about the charges that Liana faces, including "filing a false instrument" and "faces two counts of felony fraud charges". Compared with the first summary, which essentially reiterates that Liana is a "serial bride", the second summary gives more emphasis to the legal aspect and the potential implications of her case.

B.3 LLM-based Evaluation Setting

In §4.2, we compare the performance of different LLM-based evaluation methods. Specifically, for G-Eval (Liu et al., 2023a) and GPTRank, we use different prompts for different quality aspects as we defined in Appendix B.1. The prompt templates we used for GPTRank are similar to the one shown in Appendix A.3, with specific quality aspect definitions. As for G-Eval, the prompt is as follows (using the overall quality aspect as an example):

> You will be given one summary written for a news article.
>
> Your task is to rate the overall quality of the summary with a score from 1 to 5, where 1 is the lowest and 5 is the highest.
>
> Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
>
> Evaluation Steps:
>
> 1. Read the news article carefully and identify the main topic and key points.
>
> 2. Read the summary and compare it to the news article. Check if the summary covers the main topic and key points of the news article, and if it presents them in a clear and logical order.
>
> 3. Assign a score for the summary quality on a scale of 1 to 5, where 1 is the lowest and 5 is the highest.
>
> Input:
>
> News Article:
>
> {{Article}}
>
> Summary:
>
> {{Summary}}
>
> Evaluation Form (scores ONLY):
>
> - Overall Quality (1-5):

For both GPTRank and G-Eval, we set the sampling temperature to 0 to enable more deterministic behaviors. We note G-Eval tends to generate the same scores for different summaries as discussed in Liu et al. (2023a), which likely results in its low agreement with human evaluation.

ID	GPT-3.5	BRIO.GPT-4	Annotator 1	Annotator 2	Annotator 3
1	A giant turnip weighing 33lbs and measuring 4ft long has been grown in China's Yunnan Province. The vegetable was grown naturally without the help of any fertilizers and has become a big attraction, with pictures of it going viral. This is not the first time such outsized vegetables have turned up in China, but usually they are grown by farmers using commercial fertilizers.	A giant turnip weighing 33lbs and measuring 4ft long has been grown in China's Yunnan Province. The monster vegetable, which weighs more than a three-year-old child, has been nicknamed the 'fat little girl' in the village where it was grown and pictures of it have gone viral. The specimen was grown naturally without the help of any fertilizers and was so large it hangs over the flower bed.	GPT-3.5	GPT-3.5	GPT-3.5
2	Wales midfielder Aaron Ramsey has warned his Arsenal team-mates that Wales could overtake England in the FIFA rankings. Wales are currently in 22nd place, their highest-ever position, and are unbeaten in Euro 2016 qualifying. Ramsey admits that there is already banter between himself and his English team-mates at Arsenal, and that Wales are catching up in the world rankings.	Arsenal midfielder Aaron Ramsey has warned his English team-mates to beware of Wales overtaking them in the FIFA rankings, as Wales climbed to their highest-ever position in football's world order in the April rankings to move within eight places of England. Chris Coleman's side are unbeaten in Euro 2016 qualifying and would be within touching distance of the finals in France should they beat Belgium in June, and Ramsey admits the banter with the likes of Theo Walcott, Jack Wilshere and Danny Welbeck is already flying on the Arsenal training ground.	GPT-3.5	GPT-3.5	GPT-3.5
3	A poll conducted by the Daily Mail has revealed that the majority of the British public do not want Camilla to become queen if Prince Charles succeeds to the throne. The poll also showed that the nation is split on whether Charles should become king at all, with only 43% believing he should ascend to the throne compared to an almost identical number saying he should stand aside in favour of his elder son, Prince William. William and Prince Harry are the most popular members of the Royal Family, closely followed by the Queen and the Duchess of Cambridge.	A majority of the British public do not want Camilla to become queen if Prince Charles succeeds to the throne, according to a poll for the Daily Mail. The nation is completely split on whether Charles should become king at all, with only 43% believing he should, compared to an almost identical number saying he should stand aside in favour of his elder son Prince William. William and Prince Harry are the most popular members of the Royal Family, closely followed by the Queen and the Duchess of Cambridge.	BRIO.GPT-4	GPT-3.5	GPT-3.5
4	A Russian warship and two bombers were intercepted by the UK's Royal Navy and RAF after they strayed close to UK territory, prompting suspicions that Russia was using access to a Nato war games exercise in Scotland to test UK reaction times. The incidents are the latest in a series of apparent tests by Moscow as tensions between Britain and Russia rise. Sir Michael Graydon, former head of the RAF, called for higher defence spending to defeat a "resurgent" Russia, saying that the Kremlin was using Cold War tactics to "snoop" on the exercise.	Russian warships with missiles, anti-aircraft guns and torpedoes entered the English Channel, forcing the UK to send a heavily-armed frigate to monitor the vessels. Two Russian long-range Bear bombers, capable of carrying nuclear missiles, hugged British airspace, prompting the RAF to deploy two Typhoon fighter jets, raising suspicions that Vladimir Putin was using access to the site to test UK reaction times. The incursions are the latest in a long line of apparent tests by Moscow as tensions between Britain and Russia rise.	GPT-3.5	GPT-3.5	GPT-3.5
5	Bristol County Sheriff Thomas Hodgson has said that Aaron Hernandez, the former New England Patriots player who was sentenced to life for murder, is a master manipulator and will probably do fine in prison. Hodgson said Hernandez knows how to use his charm and manipulate better than anyone he has ever seen, adding that the former football star is generally affable and polite and would try to use those qualities to get what he wanted at the Bristol County House of Corrections. Hernandez will ultimately end up at the maximum-security Souza-Baranowski state prison.	Former New England Patriots player Aaron Hernandez has been sentenced to life in prison for the 2013 killing of Odin Lloyd, who was dating the sister of Hernandez's fiancee. Bristol County Sheriff Thomas Hodgson, who had Hernandez in custody for over 18 months, said that Hernandez is a master manipulator and will probably do fine in prison. Hernandez is generally affable and polite and would try to use those qualities to get what he wanted at the Bristol County House of Corrections, Hodgson said.	GPT-3.5	BRIO.GPT-4	GPT-3.5
6	Nike has faced criticism for the new kits they have designed for the US women's national team, which do not feature the American flag. The kit is all white with hints of black and yellow, and the aqua colour cleats are a tribute to Canada, who are hosting the tournament. Nike's vice president of North America communications, Charlie Brooks, has defended the decision, stating that not all national teams have to represent the flag.	Nike has defended their new kits for the US women's national team following criticism that they do not represent the American flag. The latest design features an all white strip with hints of black and yellow, while the aqua colour cleats are a tribute to Canada who will be hosting the tournament. While patriotic supporters are up in arms, feedback from the team itself has been more positive, with midfielder Tobin Heath saying she is "proud" of the new designs.	GPT-3.5	Tie	Tie
7	Liana Barrientos, an alleged "serial bride" accused of running a wife-for-hire immigration scheme, was arrested for evading a subway fare just after leaving court on Friday. Barrientos pleaded not guilty to charges that she married 10 men over 11 years and charged a fee for her "services". She has been accused of accepting money in at least one of the marriages and all of her marriages took place in New York state, allegedly starting in 1999.	Liana Barrientos, a woman accused of running a wife-for-hire immigration scheme, was arrested for evading the fare at a Bronx subway station after leaving court. She is accused of marrying 10 men over 11 years and charging a fee for her services, some of whom could pose a threat to American safety, according to investigators. She pleaded not guilty to two felony charges of filing a false instrument, involving marriage licences, and faces two counts of felony fraud charges.	BRIO.GPT-4	GPT-3.5	BRIO.GPT-4

Table 10: Expert annotation examples of the pairwise comparison between GPT-3.5 and BRIO.GPT-4. We show the three expert annotators' ratings regarding the *overall* summary quality.