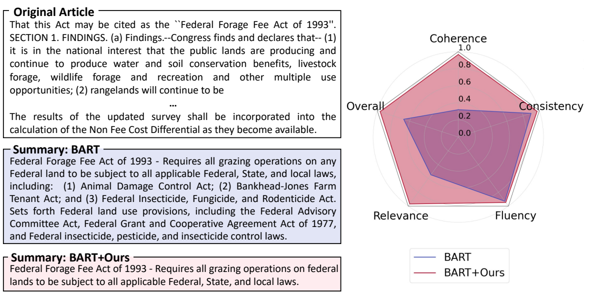
Multi-Dimensional Optimization for Text Summarization via Reinforcement Learning

Sangwon Ryu[*1], **Heejin Do**[*1], **Yunsu Kim**[3], **Gary Geunbae Lee**[1,2], **Jungseul Ok**[1,2]
[1]Graduate School of Artificial Intelligence, POSTECH, South Korea
[2]Department of Computer Science and Engineering, POSTECH, South Korea
[3]aiXplain Inc., Los Gatos, CA, USA
{ryusangwon, heejindo, gblee, jungseul}@postech.ac.kr, yunsu.kim@aixplain.com

Abstract

The evaluation of summary quality encompasses diverse dimensions such as *consistency*, *coherence*, *relevance*, and *fluency*. However, existing summarization methods often target a specific dimension, facing challenges in generating well-balanced summaries across multiple dimensions. In this paper, we propose multi-objective reinforcement learning tailored to generate balanced summaries across all four dimensions. We introduce two multi-dimensional optimization (MDO) strategies for adaptive learning: 1) MDO$_{\text{min}}$, rewarding the current lowest dimension score, and 2) MDO$_{\text{pro}}$, optimizing multiple dimensions similar to multi-task learning, resolves conflicting gradients across dimensions through gradient projection. Unlike prior ROUGE-based rewards relying on reference summaries, we use a QA-based reward model that aligns with human preferences. Further, we discover the capability to regulate the length of summaries by adjusting the discount factor, seeking the generation of concise yet informative summaries that encapsulate crucial points. Our approach achieved substantial performance gains compared to baseline models on representative summarization datasets, particularly in the overlooked dimensions.



Original Article
That this Act may be cited as the ``Federal Forage Fee Act of 1993''. SECTION 1. FINDINGS. (a) Findings.--Congress finds and declares that-- (1) it is in the national interest that the public lands are producing and continue to produce water and soil conservation benefits, livestock forage, wildlife forage and recreation and other multiple use opportunities; (2) rangelands will continue to be

...

The results of the updated survey shall be incorporated into the calculation of the Non Fee Cost Differential as they become available.

Summary: BART
Federal Forage Fee Act of 1993 - Requires all grazing operations on any Federal land to be subject to all applicable Federal, State, and local laws, including: (1) Animal Damage Control Act; (2) Bankhead-Jones Farm Tenant Act; and (3) Federal Insecticide, Fungicide, and Rodenticide Act. Sets forth Federal land use provisions, including the Federal Advisory Committee Act, Federal Grant and Cooperative Agreement Act of 1977, and Federal insecticide, pesticide, and insecticide control laws.

Summary: BART+Ours
Federal Forage Fee Act of 1993 - Requires all grazing operations on federal lands to be subject to all applicable Federal, State, and local laws.

Figure 1: While the baseline model produces an imbalanced summary (🟦), we aim to generate overall high-quality summaries (🟥). The radar chart illustrates `UniEval` scores for four dimensions.

1 Introduction

Determining a "good summary" extends beyond a single factor, generally embracing multiple dimensions such as *coherence*, *consistency*, *fluency*, and *relevance* (Kryscinski et al., 2019; Zhong et al., 2022; Liu et al., 2022b; Wang et al., 2023b; Liu et al., 2023a). Despite the remarkable advancements in abstractive summarization, challenges persist in addressing issues such as factual inconsistency, which generates inaccurate information, and irrelevance, which involves omitting crucial details.

Recently, there have been ongoing efforts to focus on such inferior dimensions (Pasunuru and Bansal, 2018; Gunasekara et al., 2021; Cao et al., 2022; Berezin and Batura, 2022; Wan et al., 2023; Liu et al., 2023b; Nan et al., 2021; Wang et al., 2023b; Chern et al., 2023), and reinforcement learning (RL) is applied as one strategy. Most existing RL approaches use a single reward of the ROUGE score (Lin, 2004), which measures the overlap with the reference summary. However, its subpar quality across various datasets has been frequently underscored (Liu et al., 2023c; Zhang et al., 2024; Goyal et al., 2023).

Pointing out the limitations of ROUGE scores in detecting hallucinations, various studies have focused on addressing this issue. Pasunuru and Bansal (2018) assigned weights to each word to overcome shortcomings of ROUGE, Roit et al. (2023) provided a reward with the natural language inference (NLI) entailment relationship between generated summary and the document, and Gunasekara et al. (2021) provided rewards via Question Answering (QA) model. However, those methods cannot capture summary-intrinsic dimensions, such as *fluency* or *coherence*. Addressing shortcomings in one dimension often leads to unintended drawbacks in other dimensions; thus, achieving a high-quality summary generation by balancing multiple dimensions remains challenging (Figure 1).

[*] Equal contribution

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5858–5871
August 11-16, 2024 ©2024 Association for Computational Linguistics

In this work, we introduce multi-objective RL, aiming to generate solid summaries that are coherent, factually consistent, fluent, and relevant. Our RL approach is based on a proximal policy optimization (PPO) (Schulman et al., 2017), and we incorporate four dimensions of a unified multi-dimensional evaluation metric, `UniEval` (Zhong et al., 2022), as multiple rewards. We suggest two strategies for optimal rewarding with multiple objectives, namely MDO_{min} and MDO_{pro}. MDO_{min} fosters adaptive learning by selecting the lowest dimension score as the reward at each iteration. Meanwhile, MDO_{pro} projects gradient onto the normal plane to handle conflicting gradients in multi-task RL, leveraging a PCGrad (Yu et al., 2020) optimizer. By effectively projecting the gradients of multiple rewards, our method can adjust the learning direction for optimal training. Both strategies aim to enhance deficient dimensions while preserving superior ones during training.

In summarization tasks, unlike typical PPO usage that rewards at each step, the score for a generated summary is obtained only at the end of the episode when the entire summary is produced. KL-penalty replaces the reward per token during episodes; hence, the discount factor can be crucial in obtaining an optimal policy (Kim et al., 2022). Consequently, we investigate how adjusting the discount factor affects the generated summaries, particularly in length.

Our MDO strategies outperform the baseline model in experiments using the representative CNN/DM and BillSum summarization datasets. Notably, our methods significantly enhance the previously inferior *relevance* dimension, supporting competitive results in other dimensions. Additional examinations, measuring whether the contents of the generated summaries are from the original articles, reveal around 90% coverage with a shorter average length. This outcome implies the capacity of the MDO to create brief yet pertinent summaries.

Our contributions are summarized as follows:

- We propose two multi-dimensional optimization methods for multi-objective RL, introducing multiple `UniEval` dimensions as rewards.
- We have empirically verified improvements in deficient dimensions while maintaining competitiveness in superior dimensions across two datasets, outperforming naive MDO methods.
- We find that adjusting a discount factor can control the generated summary length.

2 Related Work

Dimension-specific text summarization Previous studies have mostly focused on improving specific dimensions of text summarization, such as generating consistent summaries by resolving hallucinations. Wang et al. (2023b) involves a two-stage process where key entities are first extracted in the initial stage, followed by the integration of these entities to generate summaries in the second stage. Wan et al. (2023) altered the decoding strategy using a ranker and lookahead approach to produce the token with the highest faithfulness score. Their methods only considers to generate faithful summaries but overlooks other various dimensions.

RL for abstractive text summarization RL methods for text summarization have predominantly utilized the ROUGE score as a reward (Narayan et al., 2018; Chen and Bansal, 2018; Pasunuru and Bansal, 2018; Kryściński et al., 2018; Dong et al., 2018; Paulus et al., 2018; Wang et al., 2018; Parnell et al., 2022). However, recent studies emphasized that the ROUGE score fails to evaluate summaries adequately due to the revealed poor quality of reference summaries in summarization tasks (Liu et al., 2023c; Zhang et al., 2024; Goyal et al., 2023). Moreover, the ROUGE score only calculates the word overlap with the reference summary, failing to evaluate whether sentences are natural or consistent.

Therefore, some researchers have explored the application of the NLI model (Roit et al., 2023) or QA model (Gunasekara et al., 2021) as a reward, which does not solely rely on the ROUGE score. Roit et al. (2023) employs reinforcement learning with an NLI reward, aiming to maintain high *consistency* by using the entailment relationship between the summary and the document as a reward. Gunasekara et al. (2021) generate questions from both the document and the summaries using a QA model to verify the presence of answers, aiming to enhance precision and recall related to *consistency* and *relevance*. Yet, these methods do not comprehensively consider diverse quality dimensions.

Multi-objective RL RL with multiple rewards can lead to more efficient model training (Dann et al., 2023). However, multi-reward application in text summarization has not been extensively explored. Pasunuru and Bansal (2018) employ multi-

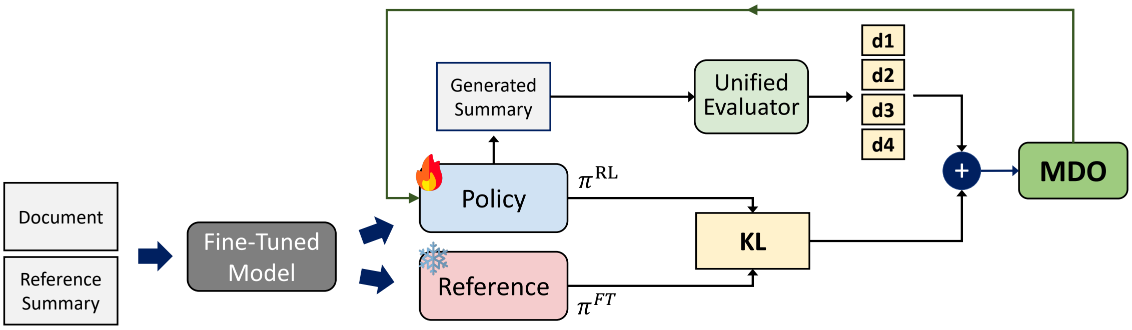


Figure 2: Entire process of Multi-dimensional Optimization (MDO). Through MDO, we optimize the scores for each dimension while training the policy. $d1$, $d2$, $d3$, and $d4$ refer to *coherence*, *consistency*, *fluency*, and *relevance*, respectively.

ple rewards such as ROUGE-L, ROUGE-Sal (which weighs vital information), and entailment, but they simply approach as multi-task learning without consideration for finer optimization. Su et al. (2023) utilize multiple RL policies to summarize multiple documents by constructing individual policy models for *importance*, *redundancy*, and *length*. They aim to concisely summarize multiple documents, preventing content overlap and including only salient information. Yet, they did not aim for a comprehensive summary of a single document, as only the *importance* feature was considered. Unlike their exclusive focus on enabling the model to capture the essential or relevant content, we explore the optimal strategies for multi-objective RL, aiming for well-balanced summarization.

3 Method Description

Throughout the RL process, it is crucial to maintain the fundamental summarization capabilities of the fine-tuned model while simultaneously improving scores across various dimensions. To achieve this goal, we employ proximal policy optimization (PPO) (Schulman et al., 2017) for RL application, utilizing a supervised, parameter-frozen reference model to guide the policy. In our pursuit of multi-objective RL in summarization, we adopt `UniEval` (Zhong et al., 2022), a metric that evaluates scores across different dimensions using a QA model. Incorporating four dimensions in the rewarding process, we introduce two optimal MDO methods to guide RL policy updates effectively. The entire process is illustrated in Figure 2.

Multi-rewards `UniEval` leverages a QA module for a unified multi-dimensional assessment in the rewarding process. The dimensions tackled by the `UniEval` closely align with human preferences, evaluating summaries based on key quality indicators. They include *coherence* (the structural coherence of the summary), *consistency* (the absence of discrepancies with the main text), *fluency* (the natural flow of sentences within the summary), and *relevance* (the inclusion of only important content from the document).

PPO PPO stands out as a well-established policy gradient model, renowned for its efficiency and stability attributed to its clipping surrogate objective. This object mitigates abrupt changes during policy updates, ensuring overall stability and avoiding divergence. Given the clipped surrogate objective, L^{CL}, the value loss, L^{VF}, and the entropy, S, the full PPO loss at timestep t is defined as follows:

$$\mathrm{L}_t(\theta) = \hat{\mathbb{E}}_t\left[\mathrm{L}_t^{\mathrm{CL}}(\theta) - c_1\mathrm{L}_t^{\mathrm{VF}}(\theta) + c_2 S[\pi_\theta](s_t)\right]$$

Unlike typical PPO applications that provide rewards at each time step, the summary can only be evaluated once when the entire sentences are generated in the summarization task. Thus, in line with the approach proposed by Stiennon et al. (2020), we employ a supervised fine-tuned summarization model as the policy π^{RL}. The value model shares parameters with π^{RL}, with an additional value head. Again, we utilize a reference model π^{FT}, which is also a fine-tuned summarization model but with frozen parameters, to maintain the summarization performance of the π^{RL}. In particular, rewards for each action, except for the generation of the last token, is the KL penalty between the policy π^{RL} and the reference model π^{FT}. This process ensures that the π^{RL} does not diverge too far from the supervised fine-tuned summarization model during the RL process. For the final action, which is the selection of the last token of the summary, a total reward is assigned by a reward model, $r(x, y)$, for the entire summary:

$$\mathrm{R}(x, y) = r(x, y) - \beta \log[\pi_\theta^{\mathrm{RL}}(y|x)/\pi^{\mathrm{FT}}(y|x)]$$

Generalized advantage estimation (GAE) (Schulman et al., 2016) is used for advantage estimation. Finely adjusting the influence of future reward in GAE is facilitated by employing the discount factor γ alongside parameter λ . x and y denote the document and summary, respectively. The state s is the current token, the action a is the selection of the next token by the π^{RL}, and the action space is the vocabulary of the π^{RL}, V.

In our multi-objective setting, the score for each dimension d_k corresponds to a reward $r_k(x, y)$. The key focus of our two MDO strategies lies in optimizing these multi-rewards to train the policy effectively. We use online learning, similar to the previous methods (Stiennon et al., 2020), which demonstrated strong performance across various domains (Fan et al., 2023).

Algorithm 1 Calculation of MDO$_{\mathrm{min}}$

```
Input: documents=$\{D_1, D_2, \dots, D_{\mathcal{N}}\}$,
policy $\pi_\theta$, model parameter $\theta$, $Evaluator$
hyperparameter $\beta$, $\lambda$, discount factor $\gamma$,
$Dims \leftarrow$ {“coh”, “con”, “flu”, “rel”}
$\mathcal{M} \leftarrow length(Dims)$
for $i = 1$ to $\mathcal{N}$ do
    $L \leftarrow 0$
    // Generate a summary
    $S_i = \pi_\theta(D_i)$
    // Calculate rewards
    for $j = 1$ to $\mathcal{M}$ do
        $r_j = Evaluator(Dims[j])$
    end for
    $r = \mathrm{argmin}_{1 \leq m \leq \mathcal{M}}\, r_m(D_i, S_i)$
    $R = r(D_i, S_i) - \beta \log\left(\frac{\pi_\theta^{RL}(S_i|D_i)}{\pi^{FT}(S_i|D_i)}\right)$
    // Estimate advantage $\hat{A}$ using GAE
    $\delta \leftarrow r_t + \gamma V(s_{t+1}) - V(s_t)$
    $\hat{A}_t \leftarrow \delta_t + \gamma\lambda\delta_{t+1} + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1}$
    $L \leftarrow$ PPO loss for $\hat{A}_t, R, \pi_\theta$
    update $\Delta\theta$
end for
```

3.1 MDO$_{\mathrm{min}}$

Focusing on the most vulnerable dimensions, we suggest MDO$_{\mathrm{min}}$, which selects a minimum dimension score as the reward, $r(x, y)$, among the evaluated four-dimensional scores. This approach intuitively aims to uplift the performance of the inferior-quality dimensions. By adopting the minimum score, the model is prompted to perform policy gradients to address the weakest dimension, achieving a balanced summary generation. The same model evaluates all four dimensions; thus, no scaling is required, and the lowest-rated dimension is directly utilized as the reward. The details of the MDO$_{\mathrm{min}}$ is explained in Algorithm 1.

Algorithm 2 Calculation of MDO$_{\mathrm{pro}}$

```
Input: documents=$\{D_1, D_2, \dots, D_{\mathcal{N}}\}$,
policy $\pi_\theta$, model parameter $\theta$, $Evaluator$
hyperparameter $\beta$, $\lambda$, discount factor $\gamma$,
$Dims \leftarrow$ {“coh”, “con”, “flu”, “rel”}
$\mathcal{M} \leftarrow$ length($Dims$)
for $i = 1$ to $\mathcal{N}$ do
    $L \leftarrow 0$
    // Generate a summary
    $S_i = \pi_\theta(D_i)$
    // Calculate rewards
    for $j = 1$ to $\mathcal{M}$ do
        $r_j = Evaluator(Dims[j])$
        $R_j = r_j(D_i, S_i) - \beta \log\left(\frac{\pi_\theta^{RL}(S_i|D_i)}{\pi^{FT}(S_i|D_i)}\right)$
        // Estimate advantage $\hat{A}$ using GAE
        $\delta \leftarrow r_t + \gamma V(s_{t+1}) - V(s_t)$
        $\hat{A}_t \leftarrow \delta_t + \gamma\lambda\delta_{t+1} + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1}$
        $L \leftarrow L +$ PPO loss for $\hat{A}_t, R_j, \pi_\theta$
    end for
    $g_m \leftarrow \nabla_\theta L(\theta)\ \forall m \in Dims$
    $g_m^{PC} \leftarrow g_m\ \forall m$
    // Project conflict gradient
    $(p, q) \leftarrow$ select $(p, q) \in Dims \times Dims$ where $p \neq q$
    if $g_p^{PC} \cdot g_q < 0$ then
        $g_p^{PC} \leftarrow g_p^{PC} - \frac{g_p^{PC} \cdot g_q}{\|g_q\|^2} g_q$
    end if
    update $\Delta\theta = g^{PC} = \sum_m g_m^{PC}$
end for
```

3.2 MDO$_{\mathrm{pro}}$

While rewards can be adaptively provided based on individual dimension scores, it may prove insufficient if there exists an inherent trade-off relationship between dimensions. For instance, attempting to improve *consistency* by including entities from the main document in the summary could potentially reduce the *fluency* between sentences within the summary. Consequently, finding a Pareto improvement becomes challenging when faced with such inherent trade-offs.

To overcome the intrinsic trade-off relationship, we suggest an MDO$_{pro}$, which projects multiple conflicting gradients onto a plane, utilizing the PCGrad optimizer (Yu et al., 2020). Treating multiple dimensions as distinct tasks, the optimizer projects each task's gradient onto the normal plane of the gradient of other tasks with conflicting gradients. In cases where gradients from multiple losses oppose each other, the learning may become ineffective. The PCGrad optimizer alleviates interference between the gradients of different dimensions by en-

		UniEval						
Model	Fine-tune	Coherence	Consistency	Fluency	Relevance	Overall	QuestEval	BERTScore
PEGASUS	SFT	0.823	0.832	0.849	0.814	0.830	0.392	0.899
BART$_{base}$	SFT	0.838	0.833	0.845	0.779	0.824	0.425	0.902
BART$_{base}$	SFT+MDO$_{\text{min}}$	**0.859**	**0.857**	**0.853**	0.806	**0.843**	**0.431**	**0.924**
BART$_{base}$	SFT+MDO$_{\text{pro}}$	0.857	0.853	0.846	**0.813**	0.842	0.428	**0.924**
BART$_{large}$	SFT	0.884	0.865	0.864	0.843	0.864	0.424	0.904
BART$_{large}$	SFT+MDO$_{\text{min}}$	0.899	0.894	**0.882**	0.869	**0.886**	**0.435**	**0.924**
BART$_{large}$	SFT+MDO$_{\text{pro}}$	**0.900**	**0.895**	0.877	**0.871**	**0.886**	0.432	0.922
T5$_{base}$	SFT	0.840	0.874	0.832	0.775	0.830	0.430	0.912
T5$_{base}$	SFT+MDO$_{\text{min}}$	0.872	0.883	0.850	0.819	0.856	0.433	0.918
T5$_{base}$	SFT+MDO$_{\text{pro}}$	**0.882**	**0.887**	**0.858**	**0.836**	**0.866**	**0.435**	**0.922**
GPT-4	-	0.973	0.843	0.831	0.971	0.904	0.443	0.851

Table 1: The results of automatic multi-dimension evaluation measured on the BillSum dataset. Within the same baseline, the bold denotes the highest score, and the underline denotes the second-highest score.

		UniEval						
Model	Fine-tune	Coherence	Consistency	Fluency	Relevance	Overall	QuestEval	BERTScore
PEGASUS	SFT	0.936	0.939	0.815	0.684	0.843	0.584	0.877
BRIO	SFT	0.951	0.931	0.826	0.776	0.871	0.619	0.883
BART$_{base}$	SFT	**0.963**	0.952	0.850	0.702	0.867	**0.594**	0.877
BART$_{base}$	SFT+MDO$_{\text{min}}$	0.955	0.958	0.894	0.734	0.885	0.555	**0.896**
BART$_{base}$	SFT+MDO$_{\text{pro}}$	0.959	**0.960**	**0.896**	**0.750**	**0.891**	0.556	**0.896**
GPT-3+CoT	-	0.948	0.870	0.948	0.910	0.919	0.574	0.874
GPT-4	-	0.967	0.840	0.945	0.934	0.921	0.597	0.864

Table 2: The results of automatic multi-dimension evaluation measured on the CNN/DailyMail (CNN/DM) dataset.

suring that the gradient of one dimension does not adversely affect the gradient of others. The detailed process is outlined in Algorithm 2.

4 Experimental Setup

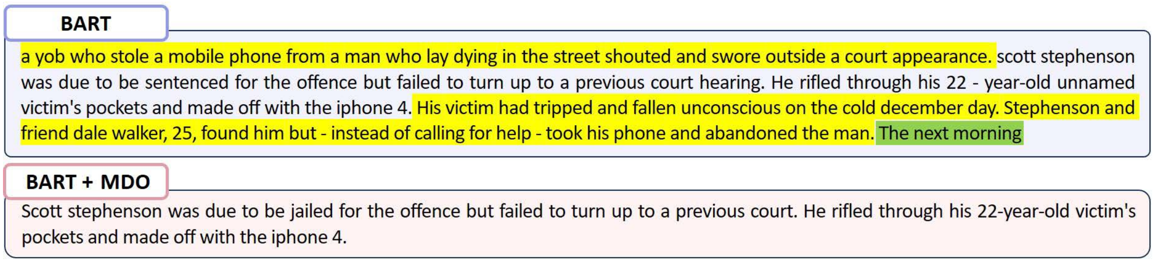
Datasets We utilize two text summarization datasets considering potential influences of source document complexity: the BillSum dataset for legislative content and the CNN/Daily Mail dataset for news summarization. BillSum comprises an 18.9K training set and a 3.2K test set, while CNN/DM has a 287K training set and an 11.5K test set. In light of studies indicating poor quality of reference summaries in the datasets (Liu et al., 2023c; Zhang et al., 2024; Goyal et al., 2023), we use an enhanced version of CNN/DM test set introduced by Wang et al. (2023b).

Baseline models As baseline models, we employ encoder-decoder models commonly used for the text summarization task, including BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). For additional comparison, we report PEGASUS (Zhang et al., 2020a) and BRIO (Liu et al., 2022a) results. To ensure comparability, we fine-tune BART-base, BART-large, and T5-base under the same hyperparameter settings: a batch size of 4, a learning rate of 5e-5, and 10 epochs. For PEGASUS and BRIO models, we utilized already fine-tuned versions on the Billsum[1] and CNN/DM[2][3].

In addition, we compare our model with LLMs, GPT-3-CoT (Wang et al., 2023b) and GPT-4 (OpenAI et al., 2024). GPT-3-CoT is a 2-stage chain-of-thought approach where the first stage extracts the core elements, and the second stage integrates them to address the issue of LLMs not sufficiently incorporating elements in generated summaries in the news datasets. We used `GPT-4-turbo` for GPT-4.

Hyperparameters for RL For RL, we use a batch size of 4, a learning rate of 1.41e-6, discount factor $\gamma = 0.9$, and randomly select only 10K samples from the training set of each dataset. We conduct experiments with three different seeds and report the average scores.

[1] https://huggingface.co/google/pegasus-billsum
[2] https://huggingface.co/google/pegasus-cnn_dailymail
[3] https://huggingface.co/Yale-LILY/brio-cnndm-cased



BART

a yob who stole a mobile phone from a man who lay dying in the street shouted and swore outside a court appearance. scott stephenson was due to be sentenced for the offence but failed to turn up to a previous court hearing. He rifled through his 22 - year-old unnamed victim's pockets and made off with the iphone 4. His victim had tripped and fallen unconscious on the cold december day. Stephenson and friend dale walker, 25, found him but - instead of calling for help - took his phone and abandoned the man. The next morning

BART + MDO

Scott stephenson was due to be jailed for the offence but failed to turn up to a previous court. He rifled through his 22-year-old victim's pockets and made off with the iphone 4.

Figure 3: Examples of the generated summaries by the baseline model and our MDO$_{pro}$ on the same document. Unimportant contents are highlighted in yellow , and unnatural or structurally disruptive ones are marked in green .

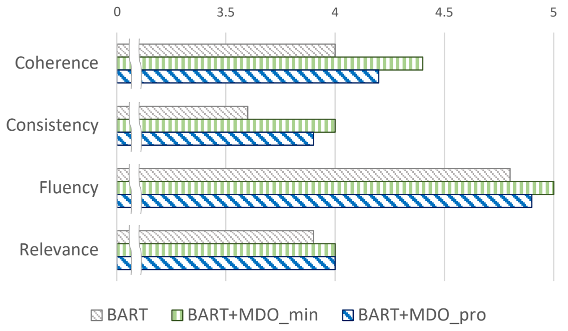


Figure 4: Multi-dimensional evaluation results with ChatGPT on the BillSum.

Metrics We use various evaluation metrics for multi-dimension assessment, such as UniEval, ChatGPT, and human evaluations. For detailed measurements on each dimension, we also use QuestEval (Scialom et al., 2021) and BERTSCore (Zhang et al., 2020b). QuestEval assesses precision by generating questions from summaries using a question generation model and checking if the answers are in the document. It generates questions from the document and verifies whether the answers are in the summaries for recall. The overall QuestEval score is an F1 score based on precision and recall. We use precision value for the BERTScore, which calculates the similarity between the token vectors in the generated summaries and those in the reference summaries based on BERT embeddings.

5 Results

Main results In Table 1, our multi-objective optimization techniques, MDO$_{min}$ and MDO$_{pro}$, have consistently demonstrate enhanced performance across all UniEval dimensions. Notably, applying to the BART-base exhibit significant advancements in the lowest-quality dimension, *relevance*, with MDO$_{min}$ and MDO$_{pro}$ showing increases of 0.027 and 0.034, respectively. Similarly, in the dimension of *consistency*, also had inferior quality, MDO$_{min}$ and MDO$_{pro}$ lead to notable improvements of 0.024 and 0.020, respectively. Our methods consistently yield modest yet discernible enhancements even in dimensions with relatively high baseline scores. The same trend is evident in the evaluation of the BART-large model, with considerable strides made in dimensions that initially exhibited lower performance, accompanied by marginal but discernible improvements in dimensions already featuring high scores. This underlines adaptive learning capabilities our methods, enabling the model to dynamically adjust its focus and balance diverse dimensions with the overall enhancements. In the assessment using alternative metrics such as QuestEval and BERTScore, the BART-large+MDO$_{min}$ model stands out. These results highlight that our generated summaries maintain competitive quality even when measured based on the original document and the reference summaries. The standard deviation is specified in Appendix A.1.

We extend our experiments to include the CNN/DM dataset. As illustrated in Table 2, training with multi-dimensional optimized methods enhances the performance on the CNN/DM dataset akin to those observed on the BillSum dataset. Notably, substantial score improvements are recorded in the dimensions of *fluency* and *relevance*, registering increases of 0.046 and 0.048, respectively, addressing areas where the quality was initially deficient. Still, the scores remained comparable or slightly lower in dimensions where the model already demonstrated high proficiency, such as *coherence* and *consistency*. Consequently, RL with MDO has resulted in well-balanced summaries across various dimensions.

As LLMs have demonstrated superior performance in summarization tasks (Zhang et al., 2024; Goyal et al., 2023; Pu et al., 2023), we compare our

	Comprehension	Attribution	Salience	Conciseness
BART	4.11	3.81	3.81	3.76
BART+MDO$_{min}$	4.73	4.17	4.36	4.74
BART+MDO$_{pro}$	4.80	4.42	4.55	4.75

Table 3: Human evaluation for the BillSum dataset. The scores are the average by three human expert.

model with the latest LLMs, GPT-3+CoT (Wang et al., 2023b) and GPT-4 (OpenAI et al., 2024). Despite the smaller model size, our method exhibits comparable performance to the larger and more expensive GPT-4 with only 0.018 differences in BillSum (Table 1). Moreover, it shows higher BERTScore in BillSum and CNN/DM (Table 2).

Figure 3 illustrates actual changes in summaries as multi-dimensional scores increase with our model. The initial models frequently incorporated irrelevant details and awkwardly constructed sentences. In contrast, our model, fine-tuned to enhance each dimension through MDO, effectively omits non-essential information and improves the natural flow of sentences. Qualitative observations suggest a positive link between improving UniEval scores and producing high-quality summaries.

ChatGPT evaluation Recent studies informed that ChatGPT's evaluation capabilities closely align with human judgments (Gao et al., 2023; Chiang and Lee, 2023; Wang et al., 2023a). To further verify with indicators other than the QA-based metrics, we include ChatGPT evaluation with four dimensions identical to those in UniEval. Inputting the document and its summaries into ChatGPT, we request evaluations for each dimension on a scale ranging from 0 to 5 (the highest) using detailed prompts. As depicted in Figure 4, the model with MDO$_{min}$ and MDO$_{pro}$ exhibits improvements across all evaluated dimensions compared to the baseline model, particularly demonstrating a noteworthy 11.1% and 8.3% increase in the lowest-rated dimension, *consistency*. The prompts are shown in Appendix B.

Human evaluation Given that the English-written BillSum dataset has congressional information, we hired three experts who are native English speakers and possess extensive experience with government documents via Upwork[4]. We follow the evaluation criteria outlined in Roit et al. (2023), which employed NLI-based RL: *comprehension*, *attribution*, *salience*, and *conciseness*. *comprehension* assesses the ease of understanding the summary, *attribution* gauges the consistency of the summary with the main document, *salience* determines whether the summary includes only the most important information, and *conciseness* evaluates the brevity of the summary. As outlined in Table 3, our model surpasses the baseline across all dimensions. Moreover, evaluators preferred summaries generated by our model over those produced by the baseline model, as depicted in Figure 5. To verify our methods, we conduct significance tests on the BillSum dataset for both human evaluation results and the UniEval *overall* score. The results of the two-tailed paired t-test, with p-values < 0.05, demonstrate statistically significant performance differences in MDO$_{min}$ and MDO$_{pro}$ compared to the baseline model, BART.

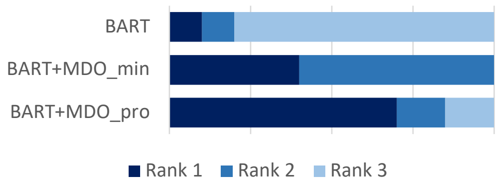


Figure 5: Human preferences for each model. Rank 1 signifies the most preferred summary among the evaluated summaries.

	ROUGE-L	Coverage	Summary Length
PEGASUS	0.431	0.902	193.073
BART	0.336	0.890	73.164
BART+MDO$_{min}$	0.284	0.907	39.464
BART+MDO$_{pro}$	0.276	0.898	37.002
T5	0.365	0.945	74.624
T5+MDO$_{min}$	0.351	0.942	63.559
T5+MDO$_{pro}$	0.340	0.939	55.957

Table 4: The Mechanical Evaluation of summarization models. Our model generates brief summaries containing only the essential information.

Mechanical analysis Recent studies (Liu et al., 2023c; Zhang et al., 2024; Goyal et al., 2023) pointed out that reference summaries generally exhibit low quality; thus, ROUGE, which solely relies on overlap with reference summaries may not accurately capture the true quality of the summaries. Nevertheless, we assess our model using traditional evaluation metrics, including ROUGE, coverage, and the average summary length. Summaries of our model show relatively lower ROUGE scores (Table 4), yet the comparative coverage, which

[4] https://www.upwork.com

		UniEval						
Model	Fine-tune	Coherence	Consistency	Fluency	Relevance	Overall	QuestEval	BERTScore
BART$_{large}$	SFT	0.884	0.865	0.864	0.843	0.864	0.424	0.904
BART$_{large}$	MDO$_{\text{sum-r}}$	0.922	0.931	0.465	0.916	0.809	0.448	0.929
BART$_{large}$	MDO$_{\text{sum-l}}$	0.892	0.887	0.872	0.861	0.878	0.431	0.924
BART$_{large}$	MDO$_{\text{min}}$	0.899	0.894	0.882	0.869	0.886	0.435	0.924
BART$_{large}$	MDO$_{\text{pro}}$	0.900	0.895	0.877	0.871	0.886	0.432	0.922

Table 5: Comparison of performance between two naive methods of summing the rewards (MDO$_{\text{sum-r}}$) or losses (MDO$_{\text{sum-l}}$) and our two optimization methods (MDO$_{\text{min, pro}}$). Our strategies show better overall performance than the former two methods and show balanced results, unlike MDO$_{\text{sum-r}}$ exhibiting a severely low score for *fluency*.

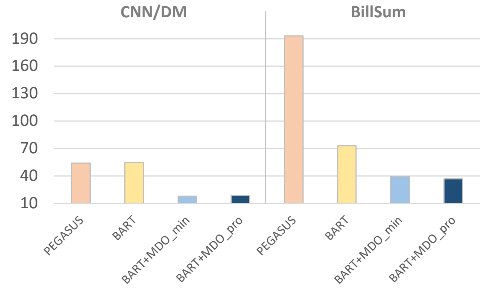


Figure 6: Comparison of summary length for each model on different datasets. Even in complex data (BillSum; right), our methods produce shorter summaries.

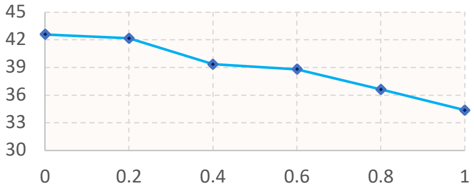


Figure 7: Averaged length of generated summaries (y-axis) according to the discount factor γ (x-axis).

calculates the proportion of tokens in the generated summary that are present in the document.

Meanwhile, models with MDO produce shorter summaries compared to those generated by base models. Comprehensive results of the substantial `coverage`, high *relevance* and *salience* scores (Table 1, 3) imply that our shorter summaries concisely encapsulate only the essential contents from the document. In contrast, summaries generated by PEGASUS average around 193 words, which is excessively long for a summary. As demonstrated by Guo and Vosoughi (2023), lengthy summaries are favorable in mechanical metrics like ROUGE. Further, Roit et al. (2023) reported a decrease in entailment percentage as the token length increases. The observed trends persist in our results, where PEGASUS, producing the longest summaries, shows the highest ROUGE scores.

6 Discussions

Summary length varies by text complexity In text summarization tasks, concisely encapsulating only the critical information is crucial. However, the optimal length of a summary depends on the document's informational content, resulting in varying ideal lengths across datasets. When a document contains rich information, its summary tends to be longer; conversely, a document with less information leads to a shorter summary. The CNN/DM news dataset includes less information, allowing for the essential contents to be sufficiently covered in a shorter length. On the other hand, the legislative dataset, BillSum, characterized by longer texts and a greater volume of information, tends to yield longer summaries for all models, as revealed in Figure 6. Remarkably, our models consistently produce short yet concise summaries for both datasets, while the PEGASUS model outputs severely lengthy summaries when the data complexity increases.

Discount factor affects summary length We investigate the impact of a discount factor γ on the length of the generated summaries. A clear pattern is found in our empirical experiments on the BillSum dataset – a larger discount factor results in shorter summaries (see Figure 7). This phenomenon can be attributed to the training process of the policy model, particularly its emphasis on the *relevance* dimension. When estimating the advantage A, a larger γ places more emphasis on future rewards. As the reward for the last token is determined using UniEval, and *relevance* often receives the lowest score among dimensions, the training focus may lean heavily towards optimizing *relevance*. Consequently, the model tends to anticipate higher scores by generating concise summaries that mostly include only the most crucial sentences, aligning with *relevance*'s evaluation criteria of containing essential information. Thus, a larger discount factor is expected to generate shorter summaries in this specific context.

Comparison with naive approaches When using RL in Language Models, careful attention should be paid to training, as models have the potential to diverge easily, and the value model may fail to converge properly. Considering the intricacy of multi-reward optimization, we conduct additional experiments, emphasizing the need for specialized optimization for multiple rewards. We explore straightforward optimization strategies, such as summing the rewards for each aspect score to formulate the final reward ($\text{MDO}_{\text{sum-r}}$) and aggregating the losses for each aspect score, akin to conducting multi-task training ($\text{MDO}_{\text{sum-l}}$). However, employing the $\text{MDO}_{\text{sum-r}}$ method amplifies the performance gap between dimensions, making the superior ones better while the inferior ones (*fluency*) worse, thereby boosting the imbalance. $\text{MDO}_{\text{sum-l}}$, a naive multi-task approach, shows improved results over the baseline but fails to outperform MDO_{min} and MDO_{pro} (Table 5). These findings highlight the importance of our adaptive optimization strategies for multi-objective RL compared to simple multi-rewarding.

7 Conclusion

This work aims to elevate the summary quality on diverse dimensions by introducing optimized multi-objective RL strategies. With the adoption of UniEval, we incorporate the assessed four-dimensional scores of summaries for rewarding. In particular, we propose two multi-dimensional optimization (MDO) strategies, aiming to learn the optimal policy during the multi-objective RL process. Our MDO strategies exhibited improved performance across all dimensions, and human-evaluated results further proved the capacity to generate balanced summaries. Comparisons with the naive summing of rewards or losses imply that our finer optimization strategies facilitates the efficacy of RL in summarization.

Limitation

In this work, we solely utilize UniEval, an open-source evaluation metric, for multi-dimensional evaluation due to its strong correlation with human judgment. However, our approach could be extended and applied if additional evaluation metrics for multiple dimensions become available. As a future work, combining multiple metrics for each single dimension can be further considered as in Wan et al. (2023). We explored the relationship between the discount factor and summary length, yet did not investigate how it practically affects performance enhancement. Observing how performance varies by adjusting the discount factor could be an intriguing topic. Also, we employ MDO on the open-source small encoder-decoder models, considering their cost-effectiveness. This choice is attributed to our main goal of showcasing the applicability of multi-objective RL in summarization tasks. However, given the model-agnostic nature of MDO, implementation with other LLMs is feasible; thus, our method can be extended in future works.

Ethical Statement

We utilized public datasets such as BillSum, CNN/DM, and CNN/DM element-aware test sets in our research. For the human evaluation conducted through Upwork, we compensated fairly for the assessments. A total of $50 was paid per person as a fixed prize for evaluating three summaries per document across ten documents, covering four dimensions and preference assessments.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00217286), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH)), and the Technology Innovation Program (20015007, Development of Digital Therapeutics of Cognitive Behavioral Therapy for treating Panic Disorder) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea).

References

Sergey Berezin and Tatiana Batura. 2022. Named entity inclusion in abstractive text summarization. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 158–162, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

I-chun Chern, Zhiruo Wang, Sanjan Das, Bhavuk Sharma, Pengfei Liu, and Graham Neubig. 2023. Improving factuality of abstractive summarization via contrastive reward learning. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 55–60, Toronto, Canada. 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*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Christoph Dann, Yishay Mansour, and Mehryar Mohri. 2023. Reinforcement learning can be more efficient with multiple rewards. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6948–6967. PMLR.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3739–3748, Brussels, Belgium. Association for Computational Linguistics.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. 2023. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pages 79858–79885. Curran Associates, Inc.

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

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3.

Chulaka Gunasekara, Guy Feigenblat, Benjamin Sznajder, Ranit Aharonov, and Sachindra Joshi. 2021. Using question answering rewards to improve abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 518–526, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaobo Guo and Soroush Vosoughi. 2023. Length does matter: Summary length can bias summarization metrics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15869–15879, Singapore. Association for Computational Linguistics.

MyeongSeop Kim, Jung-Su Kim, Myoung-Su Choi, and Jae-Han Park. 2022. Adaptive discount factor for deep reinforcement learning in continuing tasks with uncertainty. *Sensors*, 22(19):7266.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817, Brussels, Belgium. Association for Computational Linguistics.

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.

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, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed Hassan Awadallah. 2023b. On improving summarization factual consistency from natural language feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 15144–15161, Toronto, Canada. Association for Computational Linguistics.

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

Yixin Liu, Kejian Shi, Katherine S He, Longtian Ye, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2023c. On learning to summarize with large language models as references.

Yizhu Liu, Qi Jia, and Kenny Zhu. 2022b. Reference-free summarization evaluation via semantic correlation and compression ratio. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2109–2115.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

OpenAI et al. 2024. Gpt-4 technical report.

Jacob Parnell, Inigo Jauregi Unanue, and Massimo Piccardi. 2022. A multi-document coverage reward for RELAXed multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 5112–5128, Dublin, Ireland. Association for Computational Linguistics.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 646–653, New Orleans, Louisiana. Association for Computational Linguistics.

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

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos Garea, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 6252–6272, Toronto, Canada. Association for Computational Linguistics.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2016. High-dimensional continuous control using generalized advantage estimation. In *Proceedings of the International Conference on Learning Representations*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

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

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.

D. Su, D. Su, J. M. Mulvey, and H. Poor. 2023. Optimizing multidocument summarization by blending reinforcement learning policies. *IEEE Transactions on Artificial Intelligence*, 4(03):416–427.

David Wan, Mengwen Liu, Kathleen McKeown, Markus Dreyer, and Mohit Bansal. 2023. Faithfulness-aware decoding strategies for abstractive summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2864–2880, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Li Wang, Junlin Yao, Yunzhe Tao, Li Zhong, Wei Liu, and Qiang Du. 2018. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial*

Intelligence, IJCAI-18, pages 4453–4460. International Joint Conferences on Artificial Intelligence Organization.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023b. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *Proceedings of the 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.

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

A Detailed Experimental Results

A.1 Standard deviation

We evaluated the standard deviation for the experiments in Table 1 and Table 2. The standard deviation results for each dataset are reported in Table 6 and Table 7, respectively.

A.2 Performance variation according to the size of the value model

We investigated whether the size of the policy and the value models influence the performance improvement extent in MDO. The UniEval, used as our reward, is based on the T5-large with 770M parameters. Compared to the reward model, the value models of BART-base (139M) and BART-large (406M) have smaller parameters. Consequently, it might be challenging for the value model to accurately predict rewards due to its relatively smaller size than the reward model. As shown in Figure 8, the closer the value model's size to the reward model's size, the higher the performance improvement over the baseline.

A.3 Performance differences based on the base optimizer of PCGrad

In the MDO$_{\text{pro}}$, we utilized Adam as the base optimizer for PCGrad. The Adam optimizer adjusts the size of parameter updates based on the gradient magnitude, which results in significantly better performance compared to the SGD optimizer in the MDO$_{\text{pro}}$ method that involves gradient projection (Table 8).

A.4 Details of used metrics

- `UniEval` (Zhong et al., 2022): Evaluation model, which evaluates four dimensions with a single model. Each dimension is trained with questions and answers using T5. Scores for each dimension are calculated by inserting a prompt along with the summary.

- `QuestEval` (Scialom et al., 2021): Utilizes a question generation model to create questions from the document and checks if the answers to these questions are present in the summary, calculating recall. Conversely, it generates questions from the summary to check if the answers to these questions are present in the text, calculating precision.

- `BERTScore` (Zhang et al., 2020b): Calculates precision and recall through the cosine similarity between the token embeddings of the generated summary and the reference summary.

- `Coverage`: Measures whether each token of the generated summary is present in the document. Unlike exact copy, this metric is finely calculated through lemmatization and case conversion using the NLTK[5] library.

- ROUGE[6]: Counts the number of overlapping words between the generated summary and the reference summary.

[5] https://www.nltk.org

[6] https://huggingface.co/spaces/evaluate-metric/rouge

		UniEval						
Model	Fine-tune	Coherence	Consistency	Fluency	Relevance	Overall	QuestEval	BERTScore
BART$_{base}$	SFT+MDO$_{\text{min}}$	±0.011	±0.013	±0.013	±0.011	±0.007	±0.002	±0.005
BART$_{base}$	SFT+MDO$_{\text{pro}}$	±0.009	±0.010	±0.019	±0.010	±0.004	±0.001	±0.004
BART$_{large}$	SFT+MDO$_{\text{min}}$	±0.002	±0.001	±0.022	±0.003	±0.006	±0.001	±0.006
BART$_{large}$	SFT+MDO$_{\text{pro}}$	±0.007	±0.005	±0.008	±0.006	±0.003	±0.002	±0.005
T5$_{base}$	SFT+MDO$_{\text{min}}$	±0.016	±0.007	±0.016	±0.019	±0.014	±0.002	±0.004
T5$_{base}$	SFT+MDO$_{\text{pro}}$	±0.008	±0.006	±0.018	±0.008	±0.009	±0.001	±0.001

Table 6: The standard deviation for the MDO$_{\text{min}}$ and MDO$_{\text{pro}}$ models in the BillSum dataset.

		UniEval						
Model	Fine-tune	Coherence	Consistency	Fluency	Relevance	Overall	QuestEval	BERTScore
BART$_{base}$	SFT+MDO$_{\text{min}}$	±0.010	±0.008	±0.008	±0.012	±0.005	±0.003	±0.014
BART$_{base}$	SFT+MDO$_{\text{pro}}$	±0.008	±0.006	±0.008	±0.019	±0.009	±0.006	±0.028

Table 7: The standard deviation for the MDO$_{\text{min}}$ and MDO$_{\text{pro}}$ models in the CNN/DM dataset.

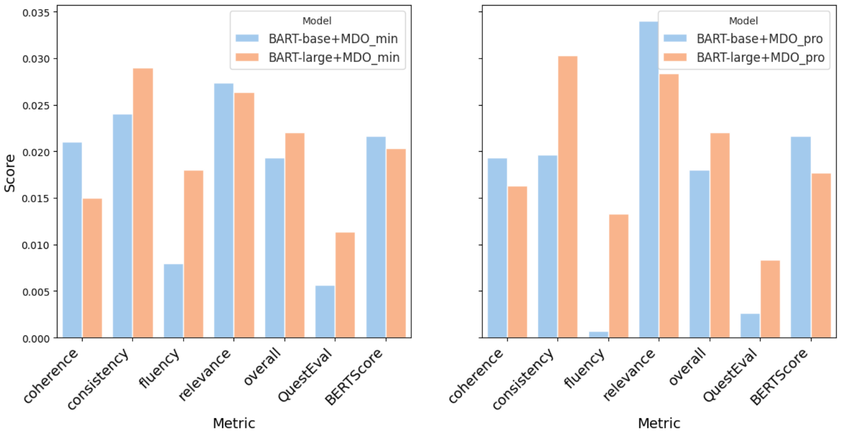


Figure 8: Performance improvement degree over the baseline model according to the value model size.

	Coherence	Consistency	Fluency	Relevance	Overall
BART	0.963	0.952	0.850	0.702	0.867
MDO$_{\text{pro-SGD}}$	0.957	0.951	0.862	0.707	0.869
MDO$_{\text{pro-Adam}}$	0.959	0.960	0.896	0.750	0.891

Table 8: In MDO$_{\text{pro}}$, the choice of the base optimizer for PCGrad leads to performance differences.

- Summary length: Counts the total word of the summary.

A.5 Hardware usage

For MDO, we used NVIDIA A100-SXM4-80GB, and for fine-tuning the baseline models on text summarization, we utilized NVIDIA RTX A5000.

B Detailed Evaluation Setup

B.1 ChatGPT evaluation

For the ChatGPT[7] evaluation, we specified how each summary should be assessed. Providing a detailed description of the dimensions enables ChatGPT to assess each dimension properly. Scores were assigned on a scale from 0 to 5 (the highest) points. When given detailed prompts to evaluate each dimension, ChatGPT provides scores for each dimension along with explanations for its evaluations. For instance, if the summary includes incorrect information, such as hallucinations, ChatGPT will measure a low consistency score and provide an explanation for this assessment. The details of prompts are in Table 9.

Description of the ChatGPT evaluation
Please evaluate the summaries. The dataset contains government and legislative data. Please evaluate three summaries per document on four aspects. The aspect required for the evaluation is as follows (score each aspect between 0 and 5, highest score of 5.0). 1. Coherence: Whether all the sentences form a coherent body. 2. Consistency: Factual alignment between the summary and the source document. 3. Fluency: The quality of individual sentences. 4. Relevance: Whether the summary contains only the important information of the source document.

Table 9

[7] https://chat.openai.com

Description of the human evaluation
Please Evaluate the summaries. The dataset contains government and legislative data. Please evaluate three summaries per document on four aspects. The aspect required for the evaluation is as follows (score each aspect between 0 and 5, highest score of 5.0) 1. Comprehension: Is that summary easy to understand? 2. Attribution: Is that summary consistent with the document? 3. Salience: Does that summary contain only important information? (There should be no unimportant content) 4. Conciseness: Is that summary short enough as a summary? 5. Overall: The overall score of the summary (in your preferences).

Table 10

B.2 Human evaluation

For our human evaluation, we hired three English-native experts through Upwork. We provided detailed scripts on how each dimension should be evaluated. Instead of using the dimensions of *coherence*, *consistency*, *fluency*, and *relevance* measured by `UniEval`, which we used as rewards, we followed the human evaluation dimensions used by Roit et al. (2023). As the four dimensions used for our rewards are core elements in assessing the summary quality, we assumed that optimizing all four core elements would likely lead to positive evaluations in other unused dimensions as well. The detailed description we provided for human evaluation is illustrated in Table 10.