Dynamic Reward Adjustment in Multi-Reward Reinforcement Learning for Counselor Reflection Generation

Do June Min[1], Verónica Pérez-Rosas[1], Kenneth Resnicow[2], Rada Mihalcea[1]
[1]Department of Electrical Engineering and Computer Science, [2]School of Public Health
University of Michigan, Ann Arbor, MI, USA
{dojmin, vrncapr, kresnic, mihalcea}@umich.edu

Abstract

In this paper, we study the problem of multi-reward reinforcement learning to jointly optimize for multiple text properties for natural language generation. We focus on the task of counselor reflection generation, where we optimize the generators to simultaneously improve the fluency, coherence, and reflection quality of generated counselor responses. We introduce two novel bandit methods, DynaOpt and C-DynaOpt, which rely on the broad strategy of combining rewards into a single value and optimizing them simultaneously. Specifically, we employ non-contextual and contextual multi-arm bandits to dynamically adjust multiple reward weights during training. Through automatic and manual evaluations, we show that our proposed techniques, DynaOpt and C-DynaOpt, outperform existing naive and bandit baselines, demonstrating their potential for enhancing language models.

Keywords: multi-reward optimization, multi-armed bandits, reinforcement learning, linguistic rewards, policy optimization, reflection generation

1. Introduction

The field of natural language processing (NLP) has witnessed remarkable advancements in recent years, with reinforcement learning (RL) emerging as a powerful approach for optimizing language models (Li et al., 2016b; Snell et al., 2023; Ramamurthy* et al., 2023). This paradigm has enabled practitioners to train models to align with diverse text properties and constraints, such as safety, helpfulness, or harmlessness (Touvron et al., 2023; Bai et al., 2022). Central to this progress is the optimization of linguistic and behavioral constraints, which serve as guiding signals during the RL training phase, shaping the model's behavior toward desired objectives. However, as NLP tasks grow in complexity and diversity, a critical challenge arises when multiple linguistic rewards must be integrated into the training process.

Researchers have explored different strategies to tackle the challenge of incorporating multiple rewards into the optimization of language models. Two prominent classes of methods have emerged in this context: (1) alternating between optimizing individual metrics at different points in the training process (*Alternate*) (Pasunuru and Bansal, 2018; Zhou et al., 2019), and (2) optimizing language models by simultaneously considering multiple metrics and combining their associated rewards into a unified objective (*Combine*) (Sharma et al., 2021; Yadav et al., 2021; Deng et al., 2022). *Alternate* involves training language models by focusing on individual metrics at different stages of the training process, which can help address the challenge of incorporating multiple rewards by prioritizing each metric separately. *Combine* aims to optimize language models by simultaneously combining multiple metrics into a single unified objective, hence offering a more integrated approach when optimizing for various criteria.

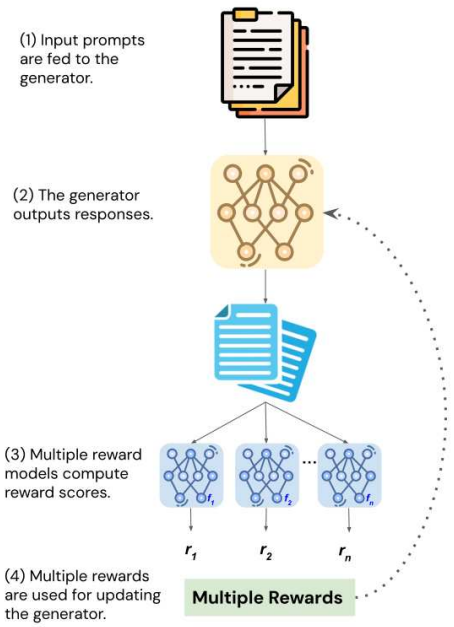

Figure 1: Workflow of RL training of language models with multiple rewards. The main question is how to simultaneously use the multiple rewards to update the LM (Step 4).

Previous methodologies employed in these approaches have predominantly relied on static configurations, wherein the alternating order or the combining ratio remains fixed and unchanging throughout the training process. To address this

LREC-COLING 2024, pages 5437–5449

limitation, Pasunuru et al. (2020) introduced the DORB extension within the *Alternate* class of methods, which harnesses multi-armed bandit (MAB) algorithms to dynamically select the reward function to optimize at each stage of training. However, notably absent from their work is an exploration of how MABs can be leveraged to enhance and adapt the *Combine* class of multi-reward optimization methods, where rewards are jointly considered.

In this paper, we extend the *Alternate* approach for multi-reward optimization by incorporating the dynamic control and adjustment of the mixing ratio of multiple rewards using MABs. Furthermore, we use contextual multi-armed bandits to address the absence of contextual information that could further aid in the optimization process. We evaluate our novel approaches against various baseline methods from both the *Alternate* and *Combine* classes on counselor reflection generation, based on Motivational Interviewing counseling exchanges. Through our experiments, (1) we find that previous naive and bandit-based multi-reward optimization methods fall short of consistently improving training reward metrics, and (2) we show that our proposed methods, DynaOpt and C-DynaOpt, offer a comparative advantage in optimizing multiple rewards in the counselor response generation task, as shown in both automated and human evaluations.

We release our code at https://github.com/michiganNLP/dynaopt.

2. Related Work

Our work relates to several main research areas at the intersection of Machine Learning and NLP.

Reinforcement Learning. RL has been successfully used to improve various NLP systems, including task-oriented dialogue systems, news article summarizers, and empathetic response generators (Singh et al., 2002; Laban et al., 2021; Sharma et al., 2021). These systems have applied various RL techniques to go beyond supervised learning with ground truth data by implementing reward models that provide learning signals to steer the behavior of language models (Ramamurthy* et al., 2023). RL strategies for this purpose include proximal policy optimization (PPO) (Schulman et al., 2017), self-critical sequence training (SCST) (Rennie et al., 2017), implicit language Q-learning (Snell et al., 2023), or Quark (Lu et al., 2022) to name a few. In our work, we chose the k-SCST algorithm for its simplicity and efficiency (Laban et al., 2021), but our approach is flexible enough to be used in tandem with RL optimization techniques.

Multi-reward Optimization. The inherent complexity of NLP tasks has motivated the use of multi-reward optimization in NLP methods (Garbacea and Mei, 2022; Dann et al., 2023). Particularly, in cases where defining a desired behavior for language models requires using multiple reward metrics, as single metrics often fall short of capturing the intricacies of model performance. Some examples can be found in (Li et al., 2016a) which uses answering, coherence, and information flow, as multiple rewards for training dialogue agents, or in Bai et al. (2022); Touvron et al. (2023), where safety and helpfulness preference models serve as guiding rewards for the reinforcement learning with human feedback (RLHF) strategy. Despite the widespread use of multiple rewards, the challenge of effectively combining them has received comparatively less attention in the literature. Notably, Pasunuru et al. (2020) tackle this issue by dynamically selecting one reward for optimization during training.

Multi-armed Bandits. MABs offer a framework for dynamically selecting among multiple actions and offer a way to navigate the exploration-exploitation trade-off (Audibert et al., 2009; Auer et al., 2002a,b; Bubeck and Cesa-Bianchi, 2012; Burtini et al., 2015a,b). This makes them suitable in NLP applications where dynamic control over multiple parameters, or input, selection is preferred over static control (Sokolov et al., 2016). MABs have been successfully used for tasks such as news article recommendation (Li et al., 2010), data selection in neural machine translation (Kreutzer et al., 2021), model selection from a pool of multiple NLP systems (Haffari et al., 2017), and crowdsourced worker selection for annotation (Wang et al., 2023). In this work, we apply MABs to the problem of multi-reward optimization.

NLP and Psychotherapy. Through our application testbed, our research is also closely connected to recent developments in NLP designed to support counselors in their practice and ongoing training within the counseling domain. Reflection, a critical element in counseling strategies like Motivational Interviewing, has been the subject of previous investigations to evaluate counseling. For instance, practitioners have used the frequency and quality of reflections as a proxy for the quality of overall counseling (Flemotomos et al., 2021; Ardulov et al., 2022). Furthermore, research has been conducted on the generation of reflections, where retrieval of human-written examples or relevant knowledge has been found to improve reflection generation performance (Shen et al., 2020, 2022; Welivita and Pu, 2023). In our work, we focus on exploring and comparing different multi-

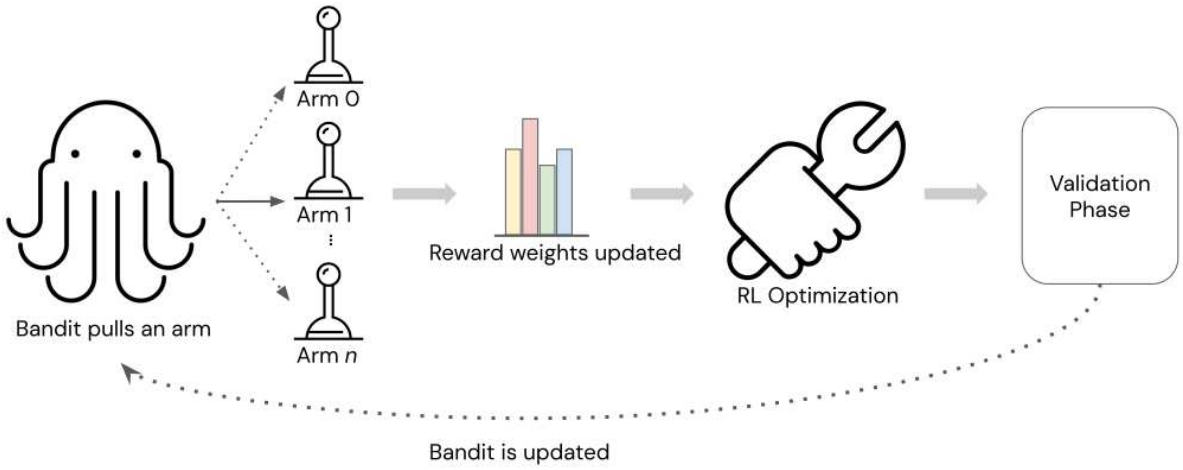

Figure 2: An overview of the DynaOpt workflow. At each bandit step, the bandit pulls an arm, which updates the reward weights. Then, the RL optimization phase uses the summed weights and updates the language model. In the bandit update phase, the LM generations are scored by the reward models, and the scores are used to update the bandit model.

reward optimization techniques for RL training of counselor reflection generation models.

3. DynaOpt: Dynamically Adjusting Rewards for Multiple Rewards Reinforcement Learning

We introduce DynaOpt, a bandit method that enables *Combine* multi-reward optimization inspired by the Dynamically Optimizing Multiple Rewards with Bandits (DORB) framework (Pasunuru et al., 2020). DynaOpt leverages the bandit framework to dynamically adjust the weights assigned to multiple rewards as shown in Figure 2.

3.1. Multi-reward Optimization with Multi-armed Bandits

The DORB framework employs the Exponential-weight algorithm for Exploration and Exploitation (Exp3) algorithm, which tackles the adversarial bandit problem (Auer et al., 2002a,b), to dynamically select from a pool of reward functions during training stages. However, it only uses the *Alternate* paradigm during the selection process. In contrast, DynaOpt enables the use of the *Combine* strategy. When the Exp3 bandit chooses a reward, its corresponding weight is incremented. In addition, while the DORB framework always chooses one reward at any given stage of training, DynaOpt allows for a "Do Nothing" option, which does not update the reward weights. Below, we describe the different steps followed by DynaOpt to enable these strategies. The pseudocode for DynaOpt is shown in Algorithm 1.

Choosing an Action. Given a reward function f_i with $i \in 1, 2, .., N$ where N is the number of rewards, the probability of selecting f_i at time t is given as:

$$p_t(i) = (1-\gamma)\frac{a_{t,i}}{\sum_{j=1}^{N} a_{t,j}} + \frac{\gamma}{N+1} \quad (1)$$

where $N+1$ indexes the choice of not updating reward weights, γ is a mixing parameter for smoothing the probability with a uniform probability over the rewards. Importantly, the arm weights $a_{t,j}$ are distinct from the reward weights ($w_{t,j}$) themselves, and are only used to sample an action, which then updates the weights.

Updating Reward Weights. Our key idea is that given a bandit's reward selection, we increment its weight instead of optimizing for that reward only. Specifically, when a bandit B chooses reward i at time t, we adjust the weight of the reward function $W_{t+1,i}$ (Lines 5 & 17, Algorithm 1), using Equations 2,3, with r_t^B set as 1:

$$\hat{r}_{t,j}^{B^W} = \begin{cases} \frac{r_t^{B^W}}{p_t(i)} & \text{if } j = i \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$w_{t+1,i} = w_{t,i} \exp\left(\frac{\gamma \hat{r}_{t,i}^{B^W}}{K}\right) \quad (3)$$

Here, we reuse the Exp3 method of updating the bandit arm weights to update our reward weights. The final reward weights W can be obtained by normalizing over $w_{t+1,i}$. Moreover, we introduce an additional weight W_{N+1}, with the $N+1$st arm representing the "Do Nothing" action, with $W_{t+1} = W_t$.

Although we reuse the weight update equation of the Exp3 algorithm, we note that different weight adjustment schemes can be employed in DynaOpt, allowing flexibility in adapting to various reward optimization scenarios.

Bandit Reward Computation. After optimizing the LM with the updated weights W_t, a reward must be computed to update the bandit B. This can be obtained by measuring the performance of the updated LM over a validation set. Departing from DORB, we use a different reward computation function for updating bandits (see line 4, Algorithm 1). Instead of using scaled rewards, we define the bandit reward as the sum of the average improvement of reward i:

$$\hat{r}^t = \sum_i {r^t}_i$$
$$ {r^t}_i = \text{Mean}(R_{t,i}) - \text{Mean}(R_{t-1,i}) \quad (4)$$

where $R_{t,i}$ = is the history of unscaled rewards for function i at time t.

Updating the Bandit. Given a reward for the bandit, we update the bandit arm weights using the Exp3 algorithm (Auer et al., 2002a; Pasunuru et al., 2020) (Line 15, Algorithm 1). Updating the bandit arm weights $a_{t,i}$ uses the same formulas used to update the reward weights (Equations 2,3).

3.2. C-DynaOpt: Reward Update with Contextual Multi-armed Bandits.

Contextual multi-armed bandits (CMABs) are a class of decision-making algorithms that incorporate contextual information to optimize action selection. In contrast to traditional MABs, which rely solely on historical performance, contextual MABs utilize additional context to make more informed and adaptive choices in various scenarios (Cortes, 2018; Agarwal et al., 2014).

We extend the DynaOpt algorithm by using *contextual* MABs (Burtini et al., 2015a; Bietti et al., 2021) instead of non-contextual MABs. We use Vowpal Wabbit's algorithm[1] to replace the Exp3 bandit and provide the current reward weights W_t and average RL dev set reward for each reward function as context to the bandit algorithm.

4. Datasets

4.1. Datasets

We use two counselor reflection datasets during our experiments.

[1] Vowpal Wabbit https://vowpalwabbit.org

Algorithm 1 DynaOpt Optimization

Input: # of rewards N, # of train steps n_{train}, # of RL validation steps $\text{round}_{\text{bandit}}$ Initial policy p_0, Initial Distribution of reward weights, W (uniform distribution over N)

1: Make a copy p_θ of initial policy p_0.
2: Initialize Exp3 bandit B with $N+1$ arms.
3: Initialize weights $w_{0,i}$ over $i \in [1, 2, \cdots, N+1]$ as uniform distribution.
4: $a \leftarrow$ chooseArm(B) ▷ Eqn 1
5: $W \leftarrow$ UpdateRewardWeight($W, B_w, a, 1$) ▷ Eqns 2,3
6: $i \leftarrow 0$
7: **while** $i < n_{\text{train}}$ **do**
8: train_responses $\leftarrow$ Sample(p_θ, train_data)
9: $r_{\text{train}} \leftarrow$ ComputeReward(train_responses, W)
10: Optimize p_θ with R_{train}, p_0 ▷ Eqn 5
11: **if** i % $\text{round}_{\text{bandit}} == 0$ **then**
12: dev_responses $\leftarrow$ Sample(p_θ, dev_data)
13: $r_{\text{bandit}} \leftarrow$ ComputeReward(dev_responses, uniform weights)
14: $r \leftarrow$ ComputeBanditReward(r_{bandit}) ▷ Eqn 4
15: UpdateBandit(B, a, r) ▷ Eqns 2,3
16: $a \leftarrow$ chooseArm(B)
17: $W \leftarrow$ UpdateRewardWeight($W, B_w, a, 1$) ▷ Eqns 2,3
18: **end if**
19: $i \leftarrow i + 1$
20: **end while**

PAIR Dataset. PAIR (Min et al., 2022) contains interactions between clients and counselors consisting of single-turn exchanges. The data collection process involved a combination of expert and crowdsource annotations. Expert annotations were employed for the reflection category while crowd sourced annotations were utilized to obtain examples of non-reflective language. Following the Motivational Interviewing Treatment Integrity (MITI) (Moyers et al., 2016) scheme, the current gold standard in Motivational Interviewing literature, each counselor's response was categorized as Complex Reflection (CR), Simple Reflection (SR), or Non-Reflection (NR). Examples illustrating these categories can be found in the bottom three rows of Table 3. Complex Reflections (CRs) are deemed as preferred responses in comparison to Simple Reflections (SRs), which, in turn, are rated higher than Non-Reflections (NRs).

CounselChat and Reddit Dataset. The dataset compiled by Welivita and Pu (2023) comprises conversations extracted from online peer support forums, such as CounselChat and Reddit. The dialogues feature a mixture of counseling conforming and non-conforming responses, reflecting the diverse nature of peer support interactions. This dataset also used MITI to categorize counselor response types at the utterance-level. We process

the dialogs in the dataset to extract only client prompt and counselor responses where the counselor responses are either Simple (SR) or Complex reflections (CR), based on the MITI annotations.

Table 1 shows statistics for each dataset.

Statistics	PAIR	Welivita and Pu (2023)
# of Exchange Pairs	2,544	1,184
Avg # of Words	32.39	36.92
# of Complex Reflections	636	768
# of Simple Reflections	318	416
# of Non-Reflections	1,590	0

Table 1: Counselor reflection dataset statistics.

4.2. Counselor Reflection Generation

Counselor reflection generation is the task of automatically generating empathetic responses that mirror and affirm a client's thoughts and feelings, fostering a therapeutic and collaborative dialogue. While multiple counseling strategies are available, we follow the Motivational Interviewing strategy. Specifically, we aim to generate counselor reflections (Complex and Simple) given client prompts about issues such as drug cessation, weight loss, or health problems.

We evaluate our proposed approaches, DynaOpt and C-DynaOpt in generating counselor reflections, alongside with the following baselines:

- Cross Entropy: This model is trained using standard supervised learning methods. It serves as a warm-start model for all the RL-trained models, providing an initial reference point for comparison.
- Round: Within the *Alternate* category, the Round baseline employs a fixed round-robin strategy, cyclically switching between reward functions. Each reward function is allocated a round size of 20 steps. This approach represents a simplistic but systematic way of alternating between different rewards during training.
- Uniform Weighted: Falling under the *Combine* class, the Uniform Weighted baseline employs a straightforward uniform weighting scheme to average the reward metrics.
- DORB: We implement the Single Multi-armed Bandit (SM) DORB method proposed in Pasunuru et al. (2020)'s. We focus on the single bandit version of DORB since previous studies have shown on par or superior performance to more complex hierarchical models (Pasunuru et al., 2020). Moreover, we avoid the additional overhead required while evaluating and tuning the controller bandit.

4.2.1. Evaluation

Reward Metrics. We use three main metrics to measure the quality of generated counselor reflections in terms of counseling style, fluency and coherence.

- Reflection Score (Min et al., 2022): This metric quantifies the quality and relevance of generated counselor reflections, ensuring that the model produces responses that are contextually appropriate and meaningful within a counseling context. It is computed by a RoBERTa scoring model that was pretrained using the PAIR dataset. We use the original weights trained by Min et al. (2022).
- Fluency: Fluency is assessed to evaluate the smoothness and coherence of generated reflections, ensuring that they read naturally and coherently. We implement our fluency reward as the inverse of the perplexity of the generated responses, following (Sharma et al., 2021).
- Coherence: Coherence evaluates the logical flow and consistency of the generated counselor reflections. We implement the coherence reward by training a RoBERTa classifier trained to detect coherent and incoherent client prompt and counselor response pairs, where incoherent pairs are created by matching prompts to randomly sampled responses (Sharma et al., 2021).

Evaluation Metrics. In our evaluation, we extend our assessment beyond the reward metrics and delve into two additional linguistic-based metrics.

- Diversity (dist-2) (Li et al., 2016a): This metric gauges the linguistic diversity of the generated counselor reflections. It measures the variety and richness of language used in the model's responses.
- The Levenshtein edit rate (Edit Rate): This quantifies the extent to which the model successfully avoids verbatim repetition of client words. This aspect of the evaluation ensures that the generated counselor reflections strike a balance between maintaining consistency and avoiding excessive repetition, aligning with the principles of effective counseling within MI (Lord et al., 2014).

Human Evaluation. In addition to automated evaluation, we conducted human annotation of 100 randomly sampled generated reflections from four models (DORB, Uniform Weighted, DynaOpt, C-DynaOpt) to assess generation quality. Two motivational interviewing experts collaborated as consultants, rating generations on a 3-level scale.

Then, the ratings were normalized to [0, 1]. The guidelines for the human annotators are included below:

> ANNOTATION GUIDELINES: Use the following guidelines to evaluate responses as either 0 (Non-Reflection), 1(Simple Reflection), or 2 (Complex Reflection):
>
> Non-Reflection (0): A response is considered a non-reflection when it does not engage with the client's input or the task at hand. It may be off-topic, irrelevant, or simply fail to address the client's query.
>
> Simple Reflection (1): A response is categorized as a simple reflection when it acknowledges the client's input or question without adding substantial depth or insight. It might repeat or rephrase the client's words, showing understanding but not extending the conversation significantly. Simple reflections demonstrate basic engagement with the client's query.
>
> Complex Reflection (2): A response is identified as a complex reflection when it goes beyond mere acknowledgment and engages deeply with the client's input or question. It demonstrates an understanding of the client's thoughts, feelings, or concerns and provides a thoughtful, insightful, or elaborate response. Complex reflections contribute to the conversation by expanding upon the client's ideas or by offering new perspectives and information.
>
> When evaluating responses, choose the most appropriate category (0, 1, or 2) based on these criteria. Keep in mind that responses may vary in complexity, and your judgment should be guided by the degree to which they reflect upon the client's prompt.
>
> Coherence. Rate the coherence of the counselor on a scale of 0 to 2 (0=not coherent at all, 1=somewhat coherent, 2=very coherent). Coherent counselor responses should effectively address the client's concerns and maintain a logical flow of conversation.
>
> Fluency. Assess the linguistic naturalness and smoothness of the counselor's responses. Responses are rated on a scale from 0 to 2, where 0 indicates responses that lack fluency, 1 signifies somewhat fluent responses, and 2 represents responses that are highly fluent and natural in their expression. Fluent counselor responses should convey information in a clear and easily understandable manner, ensuring effective communication with the client.

4.3. Experimental Setup

During our experiments, we employ the k-Self-Critical Sequence Training (k-SCST) (Laban et al., 2021), an RL algorithm known for its simplicity and effectiveness. In the k-SCST technique, k (≥ 2) samples are generated, and then their rewards $R^{S_1}, \cdots, R^{S_k}$, alongside the average reward achieved by the samples, $\overline{R}^S$, which serves as the baseline.

We use a KL-divergence loss between the initial policy p_0 and trained policy p_θ to prevent the model from deviating from the original model and generating unnatural text (Ramamurthy* et al., 2023). Thus, our RL training objective is as follows:

$$L_{RL} = \frac{1}{k}\sum_{j=1}^{k}[(\overline{R}^S - R^{S_j}) \log p_\theta(\cdot|c) - \beta\, \mathrm{KL}(p_\theta(\cdot|c)\|p_0(\cdot|c))] \quad (5)$$

where c is the prompt, $\cdot$ is the generated sequence conditioned on c, β is the KL divergence coefficient.

Additionally, we train and test the pretrained `t5-base` model (Raffel et al., 2020) on Nvidia's GeForce GTX 2080 GPUs, and use a batch size of 10, which is also the k parameter for the k-SCST algorithm. We tune our hyperparameters on the validation set. During our evaluations we combine and shuffle the datasets and use a split of 50%/10%/40% for the train/dev/test split. We report the averaged results of 5 different runs for our automated evaluations.

5. Results and Analyses

5.1. Overall Results

Not All Multi-reward Optimization Methods Are Effective for Counselor Reflection Generation. In our experiments, methods within the *COMBINE* class exhibit superior performance compared to the *ALTERNATE* methods (Table 2). We note that *COMBINE* methods such as DORB or Round failed to improve over the Cross Entropy baselines in automated reflection evaluation. This result is in contrast to Pasunuru et al. (2020)'s experiments on data-to-text generation and question generation tasks, which showed that their *ALTERNATE*-based bandit method DORB was able to achieve improvements in overall metrics. This stresses the absence of a universally optimal method for multi-reward optimization, highlighting the nuanced nature of reward combinations and the influence of task specifics on the efficacy of different approaches.

Models	Reflection (↑)	Fluency (↑)	Coherence (↑)	Edit Rate (↑)	Diversity-2 (↑)
Round	-5.02%	11.36%	5.51%	-8.75%	-0.20%
Uniform Weighted	4.48%	8.13%	**5.36**%	-6.28%	**-0.23**%
DORB (Pasunuru et al., 2020)	-3.03%	9.54%	5.42%	-7.00%	-0.08%
DynaOpt	**7.80**%	7.03%	5.02%	**-4.90**%	-0.63%
C-DynaOpt	6.14%	**8.73**%	5.02%	-5.75%	-0.46%

Table 2: Automated evaluation results on the counselor reflection generation task. We compute the average measurements of 5 different runs and report relative change over the Cross Entropy baseline. Green indicates the model achieved improvement over all reward metrics.

Client Prompt: I'm a teenager. My entire family needs family therapy, and more than likely individual therapy. My parents refuse to take action, and I'm tired of it. Is there any way I can get out of this myself?	
Automatic Generation	
Uniform Weighted	Your parents are telling you that family therapy is needed for your entire family. You're worried that as a teenager, it's getting too overwhelming and overwhelming.
DORB (2020)	Your parents are fed up with what they see as a lack of action from your part of the family.
DynaOpt	You're questioning whether or not you can get help for your entire family.
C-DynaOpt	You're questioning if there's any way you can get out of this situation.
Expert Ground Truth	
Complex Reflection	You're frustrated that your parents are hesitant to take action, and you're wondering if there's a way for you to address these concerns on your own.
Simple Reflection	You're tired of your parents not taking action.
Non-Reflection	Have you talked to your parents about how you're feeling?

Table 3: Sample reflection generations of different models on the counselor reflection generation task.

	Uniform Weighted	DORB (2020)	DynaOpt	C-DynaOpt
Reflection (↑)	69.44	64.16	73.50	74.10
Fluency (↑)	45.65	47.40	46.27	46.58
Coherence (↑)	86.49	86.86	86.37	86.40
Edit Rate (↑)	83.79	83.11	85.52	85.08
Diversity-2 (↑)	92.07	92.35	91.41	91.80

Table 4: Automated evaluation results on the counselor reflection generation task (run seed $= x$).

	Uniform Weighted	DORB (2020)	DynaOpt	C-DynaOpt
Reflection	28.29	25.30	**32.10**	29.93
Fluency	**60.31**	55.85	59.38	58.91
Coherence	62.48	62.79	63.62	**63.68**

Table 5: Human evaluation results on the counselor reflection dataset.

Comparative Advantage of Our Methods. Our results show that DynaOpt and C-DynaOpt outperform not only the *Alternate* methods but also the Uniform Weighted baseline in terms of both automatic and human reflection levels while achieving similar levels in other metrics (Tables 2, 5). Specifically, while the bandit-based DORB training leads to degraded performance over automated reflection, our methods show consistent improvement over all reward metrics.

5.2. Automated Evaluation

In the Counselor Reflection Generation experiment (see Table 2), we observe that the *Combine* methods over performed the *Alternate* models. While the *Combine* and *Alternate* models achieve similar levels of Fluency and Coherence, the *Combine* models exhibit notable improvements over the Cross Entropy baseline in terms of Reflection measurements. In contrast, the *Alternate* models show degraded performance in the Reflection metric. Interestingly, DORB achieves higher reflection levels compared to the Round approach.

Furthermore, in contrast to the bandit-based models, the Round approach exhibits higher overall variance over random runs (reflection variance of 3.59 vs 1.43 & 1.29 of our methods), indicating less stability in the training process. This further suggests that in certain settings, the bandit ap-

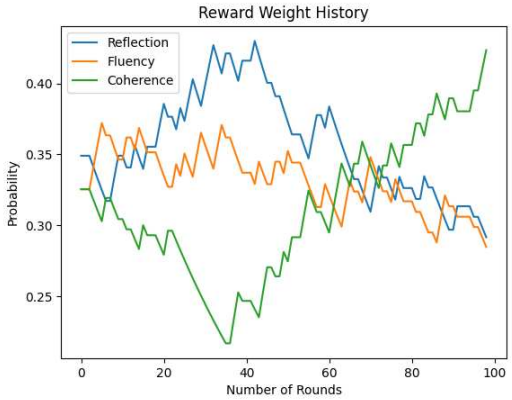

Figure 3: Reward weight trajectory of DynaOpt on the counselor reflection generation task.

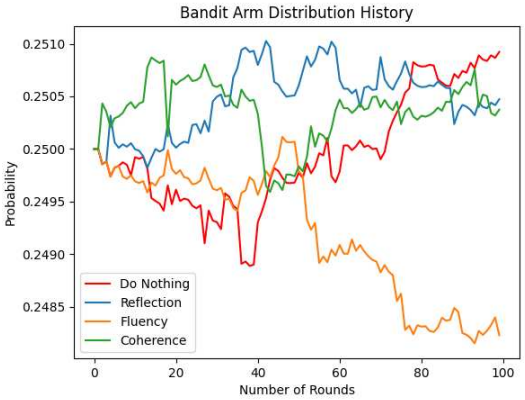

Figure 4: Bandit arm weight history of DynaOpt on the counselor reflection generation task.

proach may contribute to a more stable and adaptable training process by dynamically adjusting reward weights and optimizing multiple rewards as training progresses.

5.3. Human Evaluation

The results of our human evaluation are presented in Table 5, with sample-generated reflections shown in Table 3. In this evaluation, we compared DORB, DynaOpt, and C-DynaOpt, with the Uniform Weighted model. For human evaluation, we utilized the models trained in a single run (automated evaluations are shown in Table 4).

The evaluation results confirm the trends observed in our automated evaluation. Specifically, the *Combine* models outperform the Uniform Weighted model, which falls within the *Alternate* class. Among the *Combine* models, our proposed approaches (DynaOpt and C-DynaOpt) outperform the Uniform Weighted baseline in achieved reflection. We also see that in terms of human-evaluated fluency and coherence, our models slightly outperform DORB despite having lower automated results. This could be attributed to the higher reflection levels contributing to the overall naturalness of the generated responses of *Alternate* models. Our human evaluation reaffirms the effectiveness of our bandit-based methods, particularly in terms of enhancing reflection quality in counselor responses.

5.4. Bandit Visualization

To understand the dynamics of bandit-based reward adjustment for DynaOpt we visualized the trajectory of reward weights over the RL development set throughout training.[2] Notably, we observe that the relative importance of each reward dynamically changes over time, underscoring the adaptive nature of our bandit-based control of reward weights. This dynamic adjustment allows the model to optimize multiple rewards effectively as training progresses.

We also plot the history of the probability distribution of each arm during training in Figure 4, where "Do Nothing" corresponds to the action of not updating the reward weight distribution. We note that the trajectory of the reward weight history can be understood by tracking the evolution of bandit arm probabilities. For example, the increase of coherence reward weight around round #40 coincides with the corresponding increase of the coherence arm weight, the decline of reflection and fluency weights, as well as the rapid boost in the "Do Nothing" arm.

6. Conclusion

Our study addressed the problem of optimizing multiple linguistic rewards in reinforcement learning in the context of counselor reflection generation in motivational interviewing (MI). We explored two primary optimization strategies, the *Alternate* and *Combine* approaches, and also presented bandit-augmented versions of the latter class. Our two novel bandit methods, DynaOpt and C-DynaOpt, operate by dynamically adjusting reward weights during training using multi-armed bandits. Based on our empirical assessments, we observed that previous naive and bandit-based approaches to multi-reward optimization fail to improve response generations over the reward metrics. In addition, our proposed techniques, DynaOpt and C-DynaOpt, outperform existing baselines in the counselor response generation task, demonstrating their potential for enhancing the RL step of training language models.

[2]For all our visualizations we use the same random seed run used for results reported in Tables 4 and 5

7. Limitations

There are limitations in our study that suggest directions for future investigation. First, we have yet to examine whether the trajectory of rewards during training influences holistic model behavior, especially when *Alternate* and *Combine* models achieve similar performance metrics after optimization. In addition, our study's focus on moderate-scale language models overlooks the implications of applying our approach to larger models with billions of parameters, such as Llama 2 (Touvron et al., 2023).

Also, although our method is independent of specific RL optimization algorithms, we have not conducted experiments with popular RL algorithms such as proximal policy optimization (Schulman et al., 2017). Addressing these limitations in future research will help provide a more comprehensive understanding of the applicability and efficacy of our approach in a wider range of contexts and settings.

Finally, we emphasize that in this study our priority was to explore and compare different strategies for optimizing reflection generators with reinforcement learning, rather than creating state-of-the-art models.

8. Ethical Considerations

The datasets used in our study include motivational interviewing conversations between counselors and patients. We ensured that the source datasets processed the dialogues so that personally identifiable information was redacted. In addition, we stress that we do not advocate for the deployment of our models in clinical or mental health settings, both because human understanding and communication are vital in these domains and the behavior of language models is not fully understood. We recommend that current MI and counseling systems are best considered as tools that are best used for training and coaching learning practitioners.

Acknowledgements

The authors would like to thank researchers and students from the University of Michigan School of Public Health, for their valuable feedback and participation in this project. This material is based in part upon work supported by a National Science Foundation award (#2306372). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

9. Bibliographical References

Alekh Agarwal, Daniel J. Hsu, Satyen Kale, John Langford, Lihong Li, and Robert E. Schapire. 2014. Taming the monster: A fast and simple algorithm for contextual bandits. *CoRR*, abs/1402.0555.

Victor Ardulov, Torrey A. Creed, David C. Atkins, and Shrikanth Narayanan. 2022. Local dynamic mode of cognitive behavioral therapy.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. 2009. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902. Algorithmic Learning Theory.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. 2002a. The non-stochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. 2002b. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Alberto Bietti, Alekh Agarwal, and John Langford. 2021. A contextual bandit bake-off. *J. Mach. Learn. Res.*, 22(1).

Sébastien Bubeck and Nicolò Cesa-Bianchi. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

Giuseppe Burtini, Jason L. Loeppky, and Ramon Lawrence. 2015a. A survey of online experiment design with the stochastic multi-armed bandit. *CoRR*, abs/1510.00757.

Giuseppe Burtini, Jason L. Loeppky, and Ramon Lawrence. 2015b. A survey of online experiment design with the stochastic multi-armed bandit. *CoRR*, abs/1510.00757.

David Cortes. 2018. Adapting multi-armed bandits policies to contextual bandits scenarios. *CoRR*, abs/1811.04383.

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.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nikolaos Flemotomos, Victor R. Martinez, Zhuohao Chen, Torrey A. Creed, David C. Atkins, and Shrikanth Narayanan. 2021. Automated quality assessment of cognitive behavioral therapy sessions through highly contextualized language representations. *CoRR*, abs/2102.11573.

Cristina Garbacea and Qiaozhu Mei. 2022. Why is constrained neural language generation particularly challenging?

Alex Graves, Marc G. Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. 2017. Automated curriculum learning for neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1311–1320. PMLR.

Gholamreza Haffari, Tuan Tran, and Mark James Carman. 2017. Efficient benchmarking of nlp apis using multi-armed bandits. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Julia Kreutzer, David Vilar, and Artem Sokolov. 2021. Bandits don't follow rules: Balancing multi-facet machine translation with multi-armed bandits. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3190–3204, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. 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 6365–6378, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. 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.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. *CoRR*, abs/1003.0146.

Sarah Lord, Elisa Sheng, Zac Imel, John Baer, and David Atkins. 2014. More than reflections: Empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior Therapy*, 46.

Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. 2022. QUARK: controllable text generation with reinforced unlearning. In *NeurIPS*.

Do June Min, Verónica Pérez-Rosas, Kenneth Resnicow, and Rada Mihalcea. 2022. Pair: Prompt-aware margIn ranking for counselor reflection scoring in motivational interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Theresa Moyers, Lauren Rowell, Jennifer Manuel, Denise Ernst, and Jon Houck. 2016. The motivational interviewing treatment integrity code (miti 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65.

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,*

Volume 2 (Short Papers), pages 646–653, New Orleans, Louisiana. Association for Computational Linguistics.

Ramakanth Pasunuru, Han Guo, and Mohit Bansal. 2020. DORB: Dynamically optimizing multiple rewards with bandits. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7766–7780, Online. Association for Computational Linguistics.

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. *J. Mach. Learn. Res.*, 21(1).

Rajkumar Ramamurthy*, Prithviraj Ammanabrolu*, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In *International Conference on Learning Representations (ICLR)*.

S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195, Los Alamitos, CA, USA. IEEE Computer Society.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Ashish Sharma, Inna Wanyin Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. *Proceedings of the Web Conference 2021*.

Siqi Shen, Veronica Perez-Rosas, Charles Welch, Soujanya Poria, and Rada Mihalcea. 2022. Knowledge enhanced reflection generation for counseling dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3096–3107, Dublin, Ireland. Association for Computational Linguistics.

Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–20, 1st virtual meeting. Association for Computational Linguistics.

Satinder Singh, Diane Litman, Michael Kearns, and Marilyn Walker. 2002. Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. *J. Artif. Int. Res.*, 16(1):105–133.

Charlie Snell, Ilya Kostrikov, Yi Su, Sherry Yang, and Sergey Levine. 2023. Offline RL for natural language generation with implicit language Q learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Artem Sokolov, Julia Kreutzer, Christopher Lo, and Stefan Riezler. 2016. Learning structured predictors from bandit feedback for interactive NLP. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1610–1620, Berlin, Germany. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Yujie Wang, Chaorui Huang, Liner Yang, Zhixuan Fang, Yaping Huang, Yang Liu, and Erhong Yang. 2023. Cost-efficient crowdsourcing for span-based sequence labeling: Worker selection and data augmentation. *ArXiv*, abs/2305.06683.

Anuradha Welivita and Pearl Pu. 2023. Boosting distress support dialogue responses with motivational interviewing strategy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5411–5432, Toronto, Canada. Association for Computational Linguistics.

Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2021. Reinforcement learning for abstractive question summarization with question-aware semantic rewards. 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 249–255, Online. Association for Computational Linguistics.

Mingyang Zhou, Josh Arnold, and Zhou Yu. 2019. Building task-oriented visual dialog systems through alternative optimization between dialog policy and language generation. 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 143–153, Hong Kong, China. Association for Computational Linguistics.

A. Appendix

A.1. DORB Bandit Reward Computation

Following Graves et al. (2017), Pasunuru et al. (2020) defines the bandit reward at time t as a mean of scaled rewards:

$$\hat{r}^t = \begin{cases} 0 & \text{if } R_t < q_t^{lo} \\ 1 & \text{if } R_t > q_t^{hi} \\ \frac{R_t - q_t^{lo}}{q_t^{hi} - q_t^{lo}} & \text{otherwise} \end{cases} \quad (6)$$

where R_t = is the history of unscaled rewards at time t and q_t^{lo}, q_t^{hi} are the lower and upper quantiles of R_t.

A.2. Experiment Hyperparameters

We include the hyperparameter values we used in Table 6. We optimize the learning rate, the bandit coefficient γ, and the parameters of the contextual bandit exploration's online cover by conducting a grid search based on the criterion of average reward maximization.

B. Evaluation Results in Absolute Number

We include the automated evaluation results in absolute values and also include the standard deviation over the five different runs (Table 7).

Supervised Learning (Cross Entropy model)	
Language Model	`t5-base`
Training epochs	5
Learning Rate	1e-4
Reinforcement Learning (k-SCST)	
Language Model	`t5-base`
Learning Rate	1e-4
Sampling Temperature	1.0
Testing Temperature	0.5
k	10
KL weight β	0.05
m_{train}	1000
m_{bandit}	200
$\text{round}_{\text{bandit}}$	10
bandit coefficient γ	0.07
Bandit History Size H	200
Contextual Bandit Exploration	Online Cover $= 3$

Table 6: Experiment models & parameters.

Models	Reflection (↑)	Fluency (↑)	Coherence (↑)	Edit Rate (↑)	Diversity-2 (↑)
Cross Entropy	68.46	43.08	82.14	89.66	92.26
Round	65.03 ± 3.59	47.97± 0.94	86.67± 0.20	81.81± 1.56	92.08± 0.42
Uniform Weighted	71.53± 1.86	46.58± 0.69	86.55± 0.14	84.03± 1.04	92.05± 0.08
DORB (Pasunuru et al., 2020)	66.39± 1.84	47.18± 0.41	86.59± 0.25	83.38± 0.21	92.19± 0.21
DynaOpt	73.80± 1.43	46.11± 0.42	86.27± 0.31	85.27± 0.70	91.69± 0.30
C-DynaOpt	72.66± 1.29	46.84± 0.67	86.27± 0.21	84.50± 0.81	91.84± 0.38

Table 7: Automated evaluation results on the counselor reflection generation task. We compute the average and standard deviation measurements of 5 different runs. Alternate model resuts are highlighted in Cyan. Combine model resuts are highlighted in Red.