AutoRef : Generating Refinements of Reviews Given Guidelines

Soham Chitnis[1], **Manasi Patwardhan**[2], **Ashwin Srinivasan**[1],
Tanmay Tulsidas Verlekar[1], **Lovekesh Vig**[2], **Gautam Shroff**[2]

[1]Department of Computer Science & Information Systems, BITS Pilani, India
[2]TCS Research, India

Correspondence: manasi.patwardhan@tcs.com

Abstract

When examining reviews of research papers, we can distinguish between two hypothetical referees: the maximally lenient referee who accepts any paper with a vacuous review and the maximally strict one who rejects any paper with an overly pedantic review. Clearly, both are of no practical value. Our interest is in a referee who makes a balanced judgement and provides a review abiding by the guidelines. In this paper, we present a case study of automatic correction of an existing machine-generated or human review. The AutoRef system implements an iterative approach that progressively "refines" a review by attempting to make it more compliant with pre-defined requirements of a "good" review. It implements the following steps: (1) Translate the review requirements into a specification in natural language, of "yes/no" questions; (2) Given a $(paper, review)$ pair, extract answers to the questions ; (3) Use the results in (2) to generate a new review ; and (4) Return to Step (2) with the paper and the new review. Here, (2) and (3) are implemented by large language model (LLM) based agents. We present a case study using papers and reviews made available for the International Conference on Learning Representations (ICLR). Our initial empirical results suggest that AutoRef progressively improves the compliance of the generated reviews to the specification. Currently designed specification makes AutoRef progressively generate reviews which are stricter, making the decisions more inclined towards "rejections". This demonstrates the applicability of AutoRef for: (1) The progressive correction of overly lenient reviews, being useful for referees and meta-reviewers; and (2) The generation of progressively stricter reviews for a paper, starting from a vacuous review ("Great paper. Accept."), facilitating authors when trying to assess weaknesses in their papers.

1 Introduction

The modern practice of peer-review of scholarly articles is attributed to William Whewell's proposal in 1831 that manuscripts submitted to the Transactions of the Royal Society be subject to a review by two other members of the Royal Society. The referees–as they were called–would submit a written report on the claims of the paper. The process immediately ran into trouble, as the very first paper reviewed in this manner resulted in a strong disagreement between the two referees (Whewell was one of them). While conflicting opinions about a paper are not new, the number of such occurrences is greatly increased in the modern day, where 100s of 1000s of manuscripts are received for peer-review.[1] This high demand for refereeing effort inevitably results in researchers being unable to write the kind of comprehensive review that they would have liked and authors expect. It is in this context that we investigate the possibility of machine– assistant capable of refining reviews–by way of correction or completion–given some pre- defined notion of the ingredients of a "good" review. A natural extension of the machine- assisted correction of a review is machine-assisted generation of a review (viewed as a refinement of some default trivial review). In this paper, we demonstrate the working of the AutoRef system that is capable of both aspects.

The rest of the paper is organized as follows. In Section 2 we describe the AutoRef system. Related work is in Section 3. A case study of using AutoRef using conference submissions and reviews for the International Conference on Learning Representations (ICLR) is presented in

[1] The AI Index Report (Maslej et al., 2024) estimates the number of AI publications in journals to be around 230,000 in 2022 and the number of conference publications to be around 40,000. Obviously, the number of submissions reviewed would be much higher.

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 175–190
August 16, 2024 ©2024 Association for Computational Linguistics

Section 4. Section 5 highlights some limitations of the current AutoRef implementation and concludes the paper.

2 The AutoRef System

We first present AutoRef conceptually, referring to the main computations that are performed.

2.1 Specification

We assume the following sets: P: the set of papers; D: the set of text-based reviews; G: the set of guidelines for review; T: the set of task-orderings that specify the guidelines; R: the set of "structured-reviews"; Q: the set of "structured-questionnaires" about any review in R; and F: the set of structured-feedbacks for elements of R (the elements of T, Q R, and F will be described shortly). Then AutoRef can be seen as an implementation of the following relations and functions[2]:

$\tau : G \hookrightarrow T$ τ non-deterministically returns a task-ordering t given a guideline g;

$\kappa : T \hookrightarrow Q$ κ returns a structured-questionnaire q, given a task-ordering t;

$\alpha : P \times D \times T \hookrightarrow R$ α returns a structured-review r, given a paper p, a (possibly empty) textual review d and a task-ordering t;

$\beta : R \hookrightarrow D$ β returns a textual review d from a structured-review r;

$\phi : P \times Q \times R \hookrightarrow F$ ϕ returns feedback f about a review r, given a paper p and a questionnaire q;

$\rho : P \times T \times R \times F \hookrightarrow R$, ρ returns a (new) review r' given a paper p, a partially-ordered set of tasks t, an review r for p, and feedback f about r;

Suppose we are given a paper p; and the ICLR review guidelines g. The following is a hypothetical description of the computations performed by AutoRef : (1) Given the review guideline g, obtain a task-ordering $t = \tau(g)$ (2) Given a task-ordering t, obtain a structured-questionnaire $q = \kappa(t)$ (3) Since no initial text-based review is provided, AutoRef calls $\alpha(p, t, \emptyset)$ to generate an initial structured-review r_0 (4) $\phi(p, q, r_0)$ provides feedback f_0 in the form of a pair consisting of a score and justification (s_0, e_0); (5) $\rho(p, t, r_0, f_0)$ results in review r_1 provided the score in s_0 satisfies some checks; and the process returns to Step (4) and iterates for some k steps. Finally, AutoRef returns the textual review $\beta(r_k)$.

Figure 1 shows a diagrammatic representation of the various components of AutoRef . D is simply the set of text-based reviews. We now specify T, Q, R, and F. For this, we assume we are given reviewer-guidelines that act as requirements, in a natural language, of a review (an example, used in experiments here, is available at [3]).

Task Orderings (the set T)

We assume we are given guidelines that consist of a set of requirements, in a natural language, of a review. We formalize these requirements in the form of a (partially-ordered) set of tasks. Example tasks are: 'Goal Determination', 'Methodology Extraction', 'Experiment Extraction', 'Clarity & Ambiguity Resolution', 'Questions for Authors', 'Decision', etc. Let Γ denote the set of all possible tasks. We will represent any partially-ordered set of tasks t by a labeled directed acyclic graph (DAG). Each vertex v_i in t is labelled with some task $t_i \in \Gamma$; and an edge from v_i to v_j in t denotes that task t_j depends directly on task t_i. For example, there may be an arc from 'Experiment Extraction' to 'Experiment Claim Support'. The set T is the set of all such labeled DAGs.

Structured Questionnaires, Reviews and Feedbacks (the sets Q, R and F)

Given a DAG $t \in T$, a structured questionnaire is a DAG with the same vertices and edges as t. However for every vertex v_k in t the vertex-label is (t_k, q_k), where q_k is a text-string denoting a "yes/no" question pertaining to task t_k.[4] The set Q is the set of all possible structured-questionnaires.

A structured review is a DAG with the same vertices and edges as t, with the vertex-label for v_k being (t_k, r_k), where r_k is a text-string denoting the part of a review pertaining to task t_k. The set

[2] We use $\hookrightarrow$ to denote nondeterministic function. That is, the function may return a different answer each time it is invoked, even with the same input.

[3] https://iclr.cc/Conferences/2023/ReviewerGuide

[4] Each question is intended to verify some aspect of the review, such as 'sufficiency of literature review', 'identification of claims', *etc*.

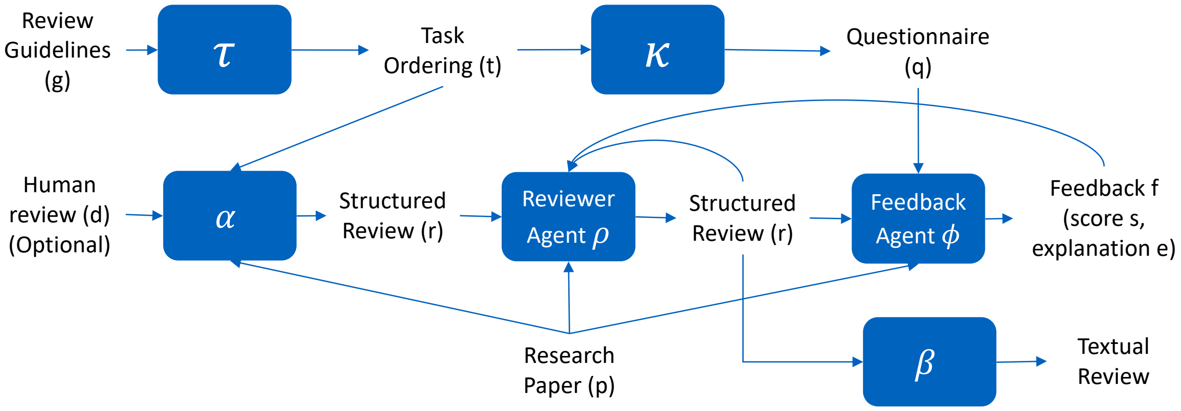

Figure 1: `AutoRef` System Components

R is the set of all possible structured-reviews.

Structured feedback is a DAG with the same vertices and edges as t, with the vertex-label for v_k being (s_k, e_k), where s_k is the score and e_k is the explanation for the score s_k pertaining to the question q_k. The set F is the set of all possible structured-feedbacks. Appendix A, B, C show examples of tasks and its partial ordering, structured-questionnaires and reviews from the experiments in this paper.

2.2 Implementation

In the implementation used in this paper, the relation τ between guideline(s) and task-ordering(s); and κ between task-ordering(s) and structured-quesionnaire(s) are manually done. The function α, β are translators from text-based reviews to structured reviews and *vice versa*. For reasons of space, we do not present procedures for these.[5] We focus here instead on implementing the ρ and ϕ functions. We call the former the 'Reviewer Agent' and the latter the 'Feedback Agent'.

The Reviewer Agent

The Reviewer Agent is an implementation of ρ, which is shown in Procedure 1.
ρ employs several auxiliary functions. We only provide brief descriptions of the main ones here:

- $Parents(\cdot)$: Given a vertex node v_k of a DAG r this function returns a sequence of parent vertices of v_k;

Algorithm 1 The function ρ

```
Input: $p$: a paper from $P$;
  $r$: a structured-review from $R$;
  $f$: feedback from $F$
Output: a structured-review from $R$
  Let $r'$ be a DAG with the same vertices and edges as $r$
  Let $\langle v_1, v_2, \ldots, v_n \rangle$ be an ancestral vertices of $r$
  for k= 1 to n do
      Let $(s_k, e_k)$ be the vertex-label for $v_k$ in $f$
      Let $(t_k, r_k)$ be the vertex label for $v_k$ in $r$
      Let $(t_k, r'_k)$ be the vertex label for $v_k$ in $r'$
      Let $h$ be a sequence of $(t_i, r_i)$ pairs s.t. $v_i \in Parents(v_k)$ and $(t_i, r_i) \in r$
      if $(s_k = 1)$ then
          if $ParentsUpdated(v_k)$ then
              $r'_k := UpdateReview(v_k, p, h, \emptyset)$
          else
              $r'_k := r_k$
          end if
      else
          $r'_k := UpdateReview(v_k, p, h, e_k)$
      end if
  end for
  return $r'$
```

- $ParentsUpdated(\cdot)$: Given a vertex node v_k of a DAG r, this function returns if the reviews of the patent vertices of v_k are updated after the last call to this function;
- $UpdateReview(\cdot)$: It refines the review r_k for task t_k, in the context of reviews of the parent tasks provided as the history h, the paper p and an optional explanation e_k in the feedback f. This function is implemented

[5] The α function is slightly more involved, requiring identification of portions of the text-based review that correspond to specific tasks identified from the guidelines. β simply concatenates the review-fragments associated with each task, using an ancestral ordering of the DAG representing the structured review.

using an LLM with a higher value of the temperature parameter. The use of the LLM makes $UpdateReview$, and hence ρ, non-deterministic.

The reviewer agent refines the review for each task t_k. The refinement happens under any of the following two scenarios: (1) Reviews of any of the parent tasks are updated, (2) The score s_k assigned to the review r_k as a part of the feedback f is zero, indicating the need for improvement in the quality of the review. For (2), the refinement occurs in the context of explanation e_k provided as the part of the feedback f.

The Feedback Agent

The Feedback Agent is an implementation of ϕ, which is shown in Procedure 2.

Algorithm 2 The function ϕ

```
Input: $p$: a paper from $P$;
  $q$: a structured-questionnaire from $Q$;
  $r$: a structured-review from $R$
Output: a structured-feedback from $F$
  Let $f$ be a DAG with the same vertices and edges as $r$
  Let $\langle v_1, v_2, \ldots, v_n \rangle$ be an ancestral ordering of vertices of $r$
  for k = 1 to n do
      Let $(s_k, e_k)$ be the vertex-label for $v_k$ in $f$
      Let $(t_k, r_k)$ be the vertex label for $v_k$ in $r$
      Let $(t_k, q_k)$ be the vertex label for $v_k$ in $q$
      $ans$ := $Answer$($q_k$,$r_k$,“yes”)
      $rel$ := $Relevant$($p$,$r_k$)
      if $ans$ and $rel$ then
          $s_k := 1$
      else
          $s_k := 0$
      end if
      $e_k := Explain(q_k, r_k, p, ans, rel)$
  end for
  return $f$
```

The main functions employed by ϕ are:

- $Answer(\cdot)$: The function returns a value “True” if the answer to the question q_k for review r_k is “Yes”, otherwise “False”.

- $Relevant(\cdot)$: The function returns a value “True” if the review r_k is relevant to paper p. If the cosine similarity of the vector representations[6] of the paper p and the review r_k is greater than a threshold[7] they are considered to be relevant.

- $Explain(\cdot)$: The function generates an explanation e_k for scoring a review r_k as either 0 ot 1 for the question q_k, in context of paper p. This function is implemented using an LLM with a higher value of temperature parameter. The use of the LLM makes $Explain$, and hence ϕ, non-deterministic.

The feedback agent assesses the review r_k for task t_k for (1) the question q_k and (2) the relevancy to the paper p. It provides a score s_k for the review as 1 if both the assessments are positive, otherwise 0. It provides an explanation e_k for the same.

Computation in `AutoRef` proceeds by repeatedly calling the Reviewer and Feedback agents in the simple iterative manner shown in Procedure 3.

Algorithm 3 `AutoRef`

```
Input: $p$: a paper from $P$;
  $g$: a set of guidelines from $G$;
  $d$: a textual review from $D$; $N$: Number of iterations
Output: a textual review from $D$
  Let $t$ be a task-ordering in $\tau(g, t)$
  Let $q$ be a structured-questionnaire in $\kappa(t, q)$
  $r_0 := \alpha(p, d, t)$
  $f_0 := \phi(p, q, r_0)$
  $S_0 := AggregateScore(f_0)$
  for $i = 1, \ldots, N$ do
      $r_i := \rho(p, r_{i-1}, f_{i-1})$
      $f_i := \phi(p, q, r_i)$
      $S_i := AggregateScore(f_i)$
      if $S_i < S_{i-1}$ then
          $r_i := r_{i-1}$
      end if
  end for
  return $\beta(r_N)$
```

Procedure 3 is a simple greedy procedure that stops after a fixed number of iterations N.The main function employed by `AutoRef` is $AggregateScore(\cdot)$. It adds the scores s_k provided for each task t_k, as the part of the feedback f_i for an iteration i. Thus, it provides an aggregated score for the complete review r_i for the

[6] OpenAI’s $text-embedding-3-small$ is used for generating embeddings (OpenAI, 2024)

[7] By default the threshold is 0.5 except for “Significance & Value Assessment" has a threshold of 0.3.

iteration i. If the aggregated score for an iteration i is inferior to the prior iteration $i-1$, AutoRef ignores the review r_i and keeps the review of the prior iteration r_{i-1}. At the end of the iteration N, AutoRef generates a textual review for the final structured review r_N.

3 Related Work

The recent work on the peer-reviewing system is through introducing new datasets, novel tasks, and analysis. DISAPERE is a labeled dataset with fine-grained annotation of reviews and rebuttal sentence-wise that showcases the understanding of argumentative strategies used by the reviewers and authors (Kennard et al., 2022). ArgSciChat is a dataset of argumentative dialogues between researchers addressing both exploratory and argumentative interactions on the scientific paper (Ruggeri et al., 2023). NLPeer is a multidomain corpus of 5k papers and 11k review reports from five venues with unified data representations (Dycke et al., 2023). QASA is a question-answering dataset on scientific articles in AI and ML. The questions require full-stack reasoning through a 3-step process: associative selection of paragraphs followed by evidential rational generation, which is utilized to compose the answer systematically (Lee et al., 2023). ARIES (Aligned, Review-Informed Edits of Scientific Papers) is a dataset of Computer Science papers (drafts & revised versions, reviewer feedback, and authors' responses to the review (D'Arcy et al., 2023). ARIES introduced the novel task of editing scientific papers using peer reviews. PeerSum introduces a dataset for generating meta-reviews of scientific papers with RAMMER (Relationship-aware Multi-task Meta-review Generator) to generate meta-reviews (Li et al., 2023). Similarly, (Wu et al., 2022) attempts to generate meta-reviews through the reviews and rebuttal. All these datasets have been designed for fine-grained analysis of review data, generation of meta-reviews, or updating the paper based on the review. Our task is a review generation task and none of these datasets can be directly used for this task. Moreover, all of these datasets contain information sourced from the open web, which is dated earlier than the data employed to train the LLM utilized in our experimentation. There is a possibility that this data has been memorized by the LLM, making it unsuitable to showcase the effectiveness of our approach. Hence, we synthesized our own dataset elaborated in Section 4.2.

ReviewerGPT provided an exploratory study on reviewing papers using LLM, showing that LLM can identify errors and verify checklists, but LLMs struggled to discern a good paper against a relatively bad paper, with a naive prompt-based approach (Liu and Shah, 2023). (Liang et al., 2023) empirically analyze if large language models can provide useful feedback on research papers. A monolithic prompt is fed to the GPT-4 (Achiam et al., 2023) to generate reviews for a scientific paper. This monolithic prompt only focuses on these four aspects: significance, novelty, potential reasons for acceptance or rejection, and suggestions for improvement. MARG (Multi-agent Review Generation) (D'Arcy et al., 2024) proposes a multi-agent system to generate reviews for a scientific paper. The system has three multi-agent groups, each one of them focusing on different aspects: Experiments & evaluation, Clarity & reproducibility, and Novelty & Impact. Additionally, a multi-agent group is defined for refinement. Every multi-agent group has several worker agents, a leader agent, and an expert. The leader agent communicates with the rest of these agents, and when the task is complete, the final answer is returned. These approaches do not render the final decision of acceptance or rejection. On the other hand, our work renders the final decision for the paper following a task-ontology-based workflow. It focuses on several aspects, much more than the prior approaches, encompassing all the dimensions of the review generation process, leading to a comprehensive review. Furthermore, unlike previous approaches, AutoRef prioritizes the refinement process, aiming specifically to enhance the robustness and rigor of the review.

4 A Case Study

4.1 Goal(s)

The goal of AutoRef is to generate good reviews. We note that it does not immediately follow that the outcome of a good review will be necessary match the true outcome of the paper, although a case can perhaps be made for the converse (a bad review is likely *not* to match the true outcome). To this end, we focus our experiments on investigating the question: "How does the iterative refinement process in AutoRef manifest itself in changes to reviews?". We will also provide some preliminary observations on the question: "Can AutoRef be used to improve the reviews generated by human reviewers?".

4.2 Materials

Guidelines and Data

We focus information available for a single conference, namely: The International Conference on Learning Representations (ICLR). We have the following information available: (a) Guidelines for reviewers (from 2023:)[8]; (b) Papers submitted along with their outcomes; (c) Reviews from (human)-referees. From this corpus, we performed a stratified selection of 300 papers (we call this the dataset D). The papers have the following distribution of outcomes: 140 Accept, and 160 Reject.

Algorithms and Machines

We use Gemini Pro-1.0 for all our experiments (Team et al., 2023). We use the temperature value of 0.7 for the reviewer agent ρ to have diverse reviews as part of our iterative process. To ensure the feedback agent ϕ is deterministic, we set the temperature value to 0 and top-p to 0. For parsing the PDF version of the paper, we utilize the ScienceBeam parser (Ecer and Maciocci, 2017), which is the same as the baseline (Liang et al., 2023).

4.3 Method

We want to examine the progressive performance of AutoRef starting without any prior review (human- or machine-generated). For this, we follow these steps:

1. Generate a task-ordering t given the ICLR guidelines.
2. Generate a structured-questionnaire q using the task- ordering t.
3. For each paper p in the dataset D
   (a) Generate a text-version of the paper p.
   (b) Generate an initial structured-review r_0 assuming no prior review (that is, an empty text-file is used as a prior review).
   (c) Repeatedly refine the review using the AutoRef procedure, and on each iteration, record the change (if any) in the review score S.
4. Summarise the number of reviews on each iteration with a positive change in Review Score (review is improved according to AutoRef)

The following details are relevant:

- t and q are obtained manually. That is, the functions τ and κ are performed by a human.
- The number of iterations performed by AutoRef is specified by the value N in Algorithm 3. We obtain results for $N = 0, 1, 2, 3, 4$ ($N = 0$ results in no refinement being done, and the algorithm returns r_0 for paper p). The value of N denotes the maximum number of attempts to generate a review for paper p with a higher review score than that of r_0.

Additionally, we seek to gain some preliminary insights into performance of AutoRef when given an existing review. It is evident that with the benefit of knowing the final outcome of the paper, we can categorise human-reviews as follows: (A) Reviews where the review-outcome is *Accept* and the true outcome is *Accept*; (B) Reviews where the review-outcome is *Accept* and the true outcome is $Reject$; (C) Reviews where the review-outcome is $Reject$ and the true outcome is *Accept*; (D) Reviews where the review-outcome is $Reject$ and the true outcome if $Reject$.

The method we adopt is largely the same as above, with some differences:

- We restrict ourselves to 5 reviews each in categories (B). The reason this is simply logistic issues arising from a step that is currently not automated (converting a human-review to a structured-review). In this event, we have assumed that the reviews in categories (B) and (C) are probably likely to require refinement. We call this dataset E. (for "erroneous")
- Initial structured reviews are obtained for each review in E. This is currently done manually. (that is, the function α is performed by a human)

[8] https://iclr.cc/Conferences/2023/ReviewerGuide

- The refined reviews are obtained as before.
- The summary statistics are now tabulations of the number of reviews whose review score improves across iterations as before.

Clearly, with small numbers, we do not expect results of statistical significance. Instead, our purpose is to understand if AutoRef can be used as an assistant to reducing false-positive errors made by a human reviewer.

4.4 Results and Discussions

Iteration		# Samples with	Average of %
Old	New	Increase in Score	Increase in score S
0	1	131	4.89%
1	2	125	4.53%
2	3	92	2.95%
3	4	79	1.89%
0	4	252	15.03%

Table 1: Progressive performance of AutoRef without any prior review. At 0^{th} iteration the structured review is generated by the LLM. The results are on dataset D.

Iteration		# Samples with	Average of %
Old	New	Increase in Score	Increase in score S
0	1	5	172.67%
1	2	4	10.68%
2	3	3	4.03%
3	4	2	1.26%
0	4	5	217.67%

Table 2: Progressive performance of AutoRef given an existing review. At 0^{th} iteration the review is generated by a human is converted to a structured review using α. The results are on dataset E.

AutoRef starts with an automatically generated review), Table 1 tabulates the numbers of reviews whose score increases for for $N = 0, \ldots, 4$. The sample distribution over scores is shown diagrammatically in Fig. 2. Table 2 and Fig. 3 show the same results for the second experiment (that is, AutoRef starts with a human review). We observe that at every i^{th} iteration, there exist samples with an increase in the score. The mean and the mode of the scores shift towards the right. This indicates that the reviews are becoming better as AutoRef progresses, complying more with the natural language specifications of the conference review guidelines (feedback questionnaire). We can see that for auto-generated reviews, the number of reviews being refined to better ones progressively decreases. The average % increase in scores also decreases over the iterations. This trend can also be tentatively observed when AutoRef starts with a human-review. This indicates that reviews for some papers are converging, with no or less increment in the scores as AutoRef progresses.

Iteration	TP	FP	FN	TN
0	107	124	33	36
1	97	115	43	45
2	98	118	42	42
3	87	118	53	42
4	92	117	48	43

Table 3: Comparing the decisions of AutoRef generated Reviews with the Meta- Reviewer's decisions for the papers in the dataset D.

We are also able to examine the outcome of reviews purely generated by AutoRef in the usual 4 cases of: TP (Actual = $Accept$, AutoRef = $Accept$); FP (Actual = $Reject$, AutoRef = $Accept$); FN (Actual = $Accept$; AutoRef = $Reject$), and TN (Actual = $Reject$, AutoRef = $Reject$). This is shown in Table 3. Here, the 'Actual' outcome is the decision provided by the meta-reviewer for the papers in D. We observe a decrease in the overall 'Accepts' ('FP's and 'TP's) and an increase in 'Rejects' ('FN's and 'TN's) over the AutoRef iterations. This indicates AutoRef is generating stricter reviews over the iterations, which are more inclined towards rejection. However, in the attempt to generate stricter reviews along with converting falsely detected 'Accepts' to 'Rejects' (reducing FPs), AutoRef is also converting correctly detected 'Accepts' (reducing TPs).

Reviews	Precision	Recall
Human-reviews	0.70	0.81
AutoRef generated (0^{th} iteration)	0.46	0.76
AutoRef generated (4^{th} iteration)	0.44	0.66

Table 4: Comparing the review-outcome with the true (meta-review) outcome for human-reviews and AutoRef for iterations 0 & 4. Precision and Recall are computed for the "Accept" class.

We compare against human-generated reviews for the same set of papers (Table 4). The Precision of AutoRef is substantially lower than a human-reviewer, but Recall is roughly comparable. The Task-ordering t synthesized for AutoRef

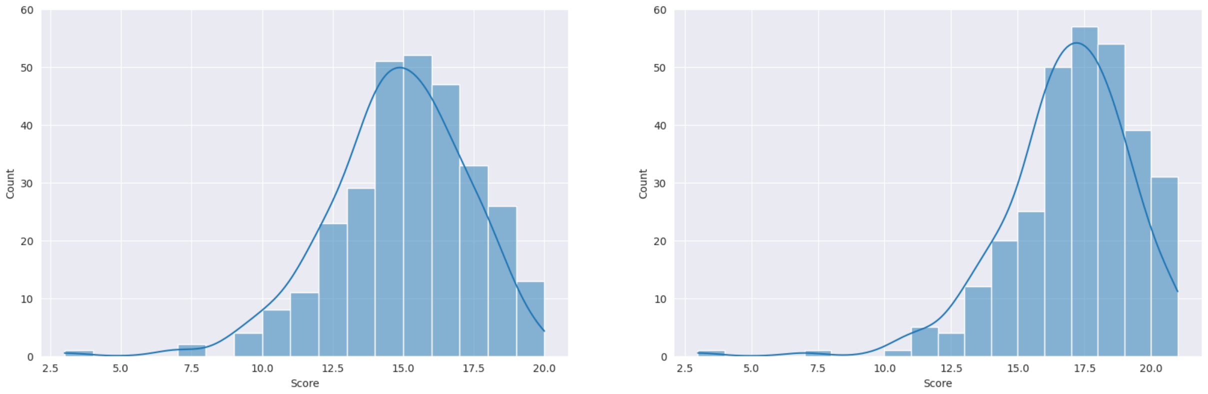

Figure 2: Progressive performance of AutoRef without any prior review. The histogram shows the distribution of scores of the papers in Dataset D at iteration 0 (left) and iteration 4 (right) of AutoRef

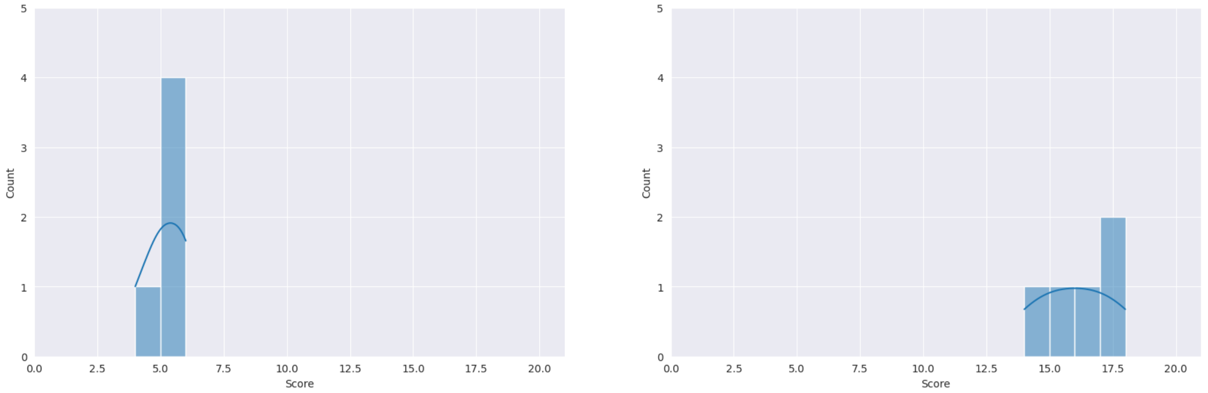

Figure 3: Progressive performance of AutoRef starting with a human-generated review. The histogram shows the distribution of scores of the papers in Dataset E at iteration 0 (left) and iteration 4 (right) of AutoRef

from guidelines g using τ can vary from the set of implicit task-orderings of distinct human reviewers. This leads to variations in reviews r by AutoRef and humans, leading to distinct decisions. Meta-reviewer refers to the human reviews while taking the final decision. Hence, having higher precision and recall for human-reviewers is obvious. Lower precision for the “Accept” class by AutoRef again points at AutoRef flipping the TPs into FNs in an attempt to reduce the FPs.

5 Concluding Remarks

We present AutoRef system, which implements an iterative approach that progressively “refines” a machine-generated or human provided review of a research paper. Our case study with ICLR papers demonstrates the capability of AutoRef to make the reviews more compliant with a review requirement specification, pre-defined in terms of “yes/no” questions. We observe a decrease in the ‘Accepts’ over the reviews generated over iterations of AutoRef , leading to the fact that AutoRef is generating stricter reviews. This demonstrates the applicability of AutoRef towards progressive correction of overly lenient review or vacuous reviews, being useful for referees, meta-reviewers, as well as authors to detect the weaknesses of the paper.

Further comparison of the outcomes of AutoRef generated reviews with meta-reviewers decisions yields lower precision as compared to outcomes of human reviews with comparable recall. In the attempt to generate stricter reviews along with converting falsely detected ‘Accepts’ to ‘Rejects’ (reducing FPs), AutoRef is also converting correctly detected ‘Accepts’ (reducing TPs), further reducing the precision as well as recall. This hints at the need for improvement in AutoRef to ensure better precision. Currently, the review decision is taken at the final leaf node of the task orderings based on the outcomes of the parent tasks. Reviewer agent being an LLM, this decision can be hallucinated. Instead, a more deterministic algorithm can be designed to make this decision based on LLM-generated outcomes of the parent tasks. Moreover, the current scoring system is very crude (+1 for each "yes" and 0 for each "no"). Clearly, several alterations could be made like increasing the weightage of more important questions and "negative marks" for incorrect answers. The questions themselves are fairly shallow and only examine superficial aspects of the review. Instead of just checking the relevancy of the review to the paper with respective vector representations, there is a need for having the paper in context for questionnaire-based evaluation of the review. More in-depth questions are possible, but it will need the language model to be able to detect those subtleties in the paper. We have also not utilized several features of language models, like using "few-shot" instances of good and bad reviews or the ability to generate multiple reviews when called repeatedly. Consequently, the current implementation should be seen as a first step in an automated review generation system. However, we do expect that the system's specification has been sufficiently abstracted to allow scope for significant improvements.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers. *Preprint*, arXiv:2401.04259.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2023. Aries: A corpus of scientific paper edits made in response to peer reviews. *Preprint*, arXiv:2306.12587.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. NLPeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.

Daniel Ecer and Giuliano Maciocci. 2017. Sciencebeam - using computer vision to extract pdf data.

Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1234–1249, Seattle, United States. Association for Computational Linguistics.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-in Lee, and Moontae Lee. 2023. Qasa: advanced question answering on scientific articles. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Miao Li, Eduard Hovy, and Jey Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7089–7112, Singapore. Association for Computational Linguistics.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. 2023. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. In *arXiv preprint arXiv:2310.01783*.

Ryan Liu and Nihar B. Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *Preprint*, arXiv:2306.00622.

Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, et al. 2024. The ai index 2024 annual report. *AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA*.

OpenAI. 2024. New embedding and api updates.

Federico Ruggeri, Mohsen Mesgar, and Iryna Gurevych. 2023. A dataset of argumentative dialogues on scientific papers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7684–7699, Toronto, Canada. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Po-Cheng Wu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Incorporating peer reviews and rebuttal counter-arguments for meta-review generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 2189–2198, New York, NY, USA. Association for Computing Machinery.

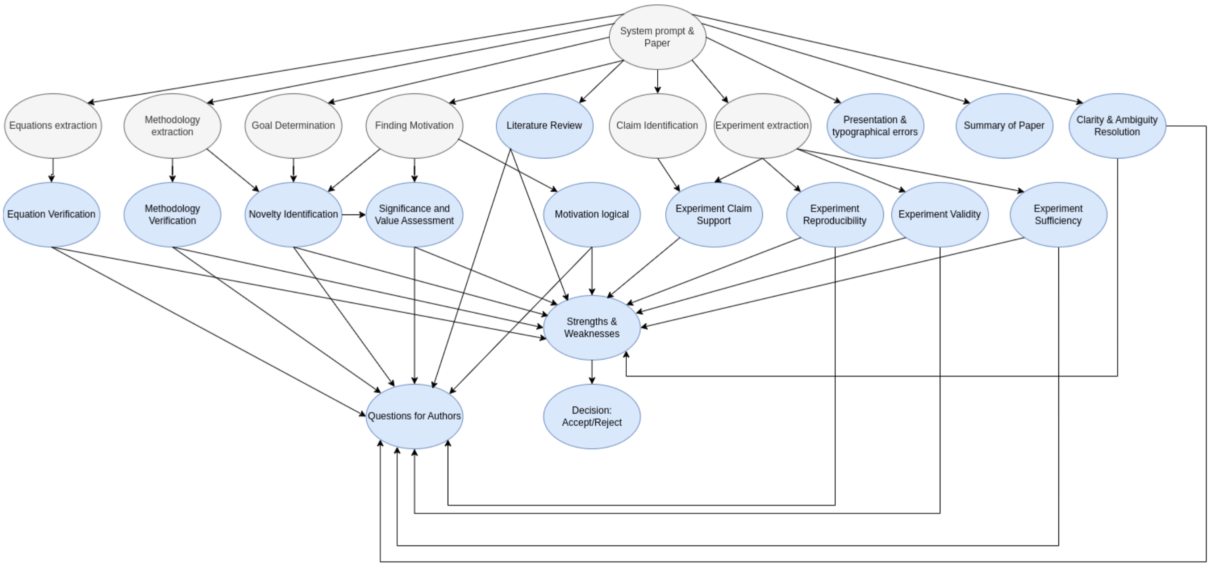

Figure 4: Directed acyclic graph representing the partially-order tasks

A Tasks

1. **Goal Determination:** What is the aim or objective of the paper? (Improvement, novel application, new findings)

2. **Finding Motivation:** What factors drove the authors to pursue the proposed approach, basically the motivation behind the work?

3. **Motivation Logical:** Is the motivation behind the paper logical?

4. **Literature Review:** Has the paper done sufficient literature review? In other words, has the paper cited all work related to it?

5. **Methodology Extraction:** What is the methodology or the approach of the paper?

6. **Equation Extraction:** What are the equations involved in the paper?

7. **Claim Identification:** Identify the specific claim(s) made by the paper. Ensure the claims are well-defined.

8. **Experiment Extraction:** What are the experiments conducted by the paper?

9. **Clarity & Ambiguity Resolution:** Assess the clarity and organization of the paper's writing. Identify any sections that require further clarification or explanation.

10. **Presentation & Typographical Error Detection:** Check for typographical errors and suggest improvements to the presentation. Evaluate the overall quality of the paper's presentation.

11. **Summary of the Paper:** Briefly summarize the paper.

12. **Methodology Verification:** Thoroughly examine the technical details of the proposed methodology and check for logical correctness. In case of mistakes, highlight them.

13. **Equation Verification:** Verify the accuracy and soundness of the mathematical equations.

14. **Novelty Identification:** Based on the previous conversation, identify any novel findings or contributions presented in the paper. Evaluate the significance and originality of these findings.

15. **Experiment Claim Support:** Based on the previous conversation, check if the experiments conducted support the claims.

16. **Experiment Reproducibility:** Determine if the experiments are reproducible.

17. **Experiment Validity:** Check the experiments for technical mistakes from a machine learning point of view.

18. **Experimental Sufficiency:** Determine if the experiments are sufficient. You should strictly assess the experiments. You may suggest any

additional experiments, including ablation experiments or experiments with more datasets, to strengthen the claims made by the paper.

19. **Significance and Value Assessment:** Determine the significance and value of the work presented in the paper. Consider its potential impact on the Machine Learning community.

20. **Strengths & Weaknesses Identification:** Based on the previous conversation, any positively identified points are the strengths, and any negatively identified points are the weaknesses. Consider the correctness of the methodology and equations, novelty, significance, claims, experimental analysis, and clarity of the work.

21. **Questions for Authors:** Based on the previous conversation, formulate questions for the authors to clarify any ambiguities or address concerns raised during the review process. While asking questions you must provide supporting evidence to what you are asking. **Do not ask trivial or shallow questions**. Ask only in-depth and specific questions that will help the community and the authors. Example: explanation of a certain part.

22. **Acceptance/Rejection Decision:** Compare the strengths and weaknesses of the paper. The following are labeled as major weaknesses: 1. Lack of motivation behind the proposed approach 2. Insufficient literature review 3. Proposed Methodology is unsound or logically incorrect. 4. Paper has limited novelty 5. Paper adds limited value to the Machine Learning community (Limited significance from a machine learning point of view) 6. Claim's made by the paper are factually incorrect. 7. Conducted experiments are insufficient to support the paper's claim. 8. Too many unclear and ambiguous parts in the paper. Minor weaknesses are: 1. Small mistakes like typographical errors in equations. 2. One or two sections need elaboration. If there are more than or equal to 1 major weakness, the paper must be rejected. When the weaknesses numerically outweigh the strengths, the paper must be rejected.

B Structured Questionnaire

1. **Goal**: Does the review identify the aim of the paper?

2. **Motivation 1**: Does the review find the motivation behind the paper?

3. **Motivation 2**:Does the review clearly state whether the motivation is logical or not?

4. **Literature Review**: Does the review assess the sufficiency of the literature review?

5. **Methodology Extraction**: Does the review extract the methodology of the paper?

6. **Equation Extraction**: Does the review extract the mathematical equations in the paper?

7. **Methodology Correctness**: Does the review assess soundness and logical correctness of the proposed methodology in the paper?

8. **Equation Correctness**: Does the review assess the correctness of the mathematical equations in the paper?

9. **Claims**: Does the review identify the claims made by the paper?

10. **Experiment Extraction**: Does the review extract the experiments from the paper?

11. **Experiment Claim Support**: Does the review check if the claims are supported by the experiments?

12. **Experiment Validity**: Does the review evaluate the validity and reliability of the experiments, highlighting any potential limitations or biases?

13. **Experimental Sufficiency**: Does the review clearly determine the sufficiency of the experiments and if the experiments are not sufficient then does it suggest any additional specific experiments to strengthen the claims?

14. **Experiment reproducibility**: Does the review assess whether the results presented in the paper can be reproduced by other researchers?

15. **Novelty Identification**: Does the review acknowledge and discuss any novel findings or significant contributions presented in the paper?

16. **Presentation Errors**: Does the review assess presentation of the paper and suggests any typographical errors or grammatical mistakes in the paper?

17. **Clarity**: Does the review assess the clarity and organization of the paper's writing?

18. Significance: Does the review evaluate the significance and contribution of the work to Machine Learning?

19. **Questions**: Are the questions posed to the authors non-trivial from machine Learning point of view?

20. **Strengths & Weaknesses**: Does the review provide a detailed analysis of both the strengths and weaknesses of the paper?

21. **Decision**: Does the review provide a clear and justified recommendation regarding the acceptance or rejection of the paper?

C Examples of Refinement of a Human Review

Table 5 and 8 showcase examples of human textual review from ICLR 2023. These reviews are converted into a structured-review using the function α, which in this case is manual. For $N = 4$, we refine the initial structured review. The initial structured reviews are shown in Table 6 and 9 and the structure review after 4 iterations of refinement are shown in Table 7 and 10. In the structured review, only some parts are shown. We follow the task-ordering as shown in Figure 4.

Summary of the paper: This paper aims to solve the exploration-exploitation tradeoff problem in the context of multi-agent reinforcement learning. While there have been a myriad of works on exploration-exploitation tradeoff on single age reinforcement learning, there are not many on the multi agent RL. This work proposes an adaptive entropy-regularization framework that learns adequate amount of exploration for each agent. To this end, this work proposes to decompose the joint soft value function into pure return and entropy sum. This disentanglements enable a more stable while updating the temperature parameters. This work focuses entropy-based MARL.

Strengths and Weaknesses: - The strength of the paper comes from the idea that, while previous works encourage same level of exploration across agents, this work proposes to differentiate the level of exploration across agents in multi-agent RL setting.
- Another strength comes from the core idea of this work: joint soft value function decomposition / separated factorization.
- The motivation part 3.1 sounds convincing to me; one agent's exploration can hinder other agent's exploitation, resulting that simultaneous exploration of multiple agents can make learning unstable. Need a framework that can adaptively learn proper levels of exploration for each agent.
- Experiments are well done, not extensive though.
Questions.
Q1. Question about ADER performance shown in Figure 2a. It seems that ADER outperforms other methods like SER-DCE, SER-MARL, but there is a point where ADER's performance suddenly jumps up in the middle. Is there any explanation on why this happens?
Q2. In Appendix B, could you give me more justification on setting the coefficient beta_i? Especially, line B.7, beta_i are defined as softmax of expectation of \\partial V^R_{JT}(s,\\tau) / \\partial H (pi (|))) ? Could you give us more detailed explanations on it? And can you explain why it is difficult to directly obtain the partial derivative in discrete-action case, and using chain rule is justified?

Clarity, Quality, Novelty and Reproducibility: This paper is well-written with high clarity. Somewhat novel, but not groundbreakingly novel. I think the authors showed good amount of experiments and evaluations on various benchmarks, and ablation studies, which seem to be reproducible.

Summary of the review: I would give marginally above the acceptance threshold. It would be good if the authors could answer my questions. There might be some issues that I didn't catch, and if other reviewers have raised issues, I'm happy to discuss.

Correctness: 4: All of the claims and statements are well-supported and correct.

Technical, Novelty, and Significance: 3: The contributions are significant and somewhat new. Aspects of the contributions exist in prior work.

Empirical, Novelty and Significance: 2: The contributions are only marginally significant or novel.

Flag for Ethics: No

Recommendation: 6: marginally above the acceptance threshold

Table 5: Example of Human Textual Review

Goal Determination: This paper aims to solve the exploration-exploitation tradeoff problem in the context of multi-agent reinforcement learning.

Motivation Logical: The motivation part 3.1 sounds convincing to me; one agent's exploration can hinder other agent's exploitation, resulting that simultaneous exploration of multiple agents can make learning unstable. Need a framework that can adaptively learn proper levels of exploration for each agent.

Literature Review: None

Clarity & Ambiguity Resolution: None

Summary of the paper: This paper aims to solve the exploration-exploitation trade-off problem in the context of multi-agent reinforcement learning. While there have been a myriad of works on exploration-exploitation tradeoff on single age reinforcement learning, there are not many on the multi agent RL. This work proposes an adaptive entropy-regularization framework that learns adequate amount of exploration for each agent. To this end, this work proposes to decompose the joint soft value function into pure return and entropy sum. This disentanglements enable a more stable while updating the temperature parameters. This work focuses entropy-based MARL

Novelty Identification: The contributions are significant and somewhat new.

Experiment Claim Support: All of the claims and statements are well-supported and correct.

Experimental Sufficiency: Experiments are well done, not extensive though.

Decision: Accept

Table 6: Initial Structured Review converted from the human textual review using α

Goal Determination: This paper aims to solve the exploration-exploitation tradeoff problem in the context of multi-agent reinforcement learning.

Motivation Logical: The motivation of the paper is to handle the multi-agent exploration-exploitation trade-off. The motivation is logical because in multi-agent reinforcement learning, the degree of necessity and benefit of exploration can be different across multiple agents and one agent's exploration can hinder other agents' exploitation. Thus, the balance of exploration and exploitation across multiple agents should also be considered for MARL in addition to that across the time dimension. The motivation is logical.

Literature Review: The paper cites a number of related works, including MAVEN (Mahajan et al., 2019) Coordinated exploration (Wang et al., 2019) Restricted space selection (Liu et al., 2021b), Intrinsic motivation-based exploration (Zheng et al., 2021) Joint exploration (Gupta et al., 2021). These works all address the problem of exploration in multi-agent reinforcement learning, and the paper discusses how its approach differs from and improves upon these methods. I have checked the paper, and it seems to have cited all related work. Yes, the paper has done sufficient literature review and has cited all related work.

Clarity & Ambiguity Resolution: The paper is well-written and organized. It is easy to follow the authors' arguments and the flow of the paper. However, there are a few sections that could be improved for clarity. The introduction could be more concise. The authors could remove some of the background information and focus on the main points of their paper. For example, they could remove the discussion of the history of reinforcement learning and focus on the specific problem that they are addressing in their paper. The methods section could be more detailed. The authors could provide more information about the algorithms they used and the experiments they conducted. For example, they could provide more information about the hyperparameters that they used and the specific tasks that they evaluated their algorithms on. The results section could be more visually appealing. The authors could use graphs and tables to illustrate their findings. For example, they could use a graph to show the performance of their algorithms on different tasks. By making these changes, the authors could improve the clarity and organization of their paper and make it easier for readers to understand their work.

Summary of the paper: This paper aims to solve the exploration-exploitation tradeoff problem in the context of multi-agent reinforcement learning. While there have been a myriad of works on exploration-exploitation tradeoff on single age reinforcement learning, there are not many on the multi agent RL. This work proposes an adaptive entropy-regularization framework that learns adequate amount of exploration for each agent. To this end, this work proposes to decompose the joint soft value function into pure return and entropy sum. This disentanglements enable a more stable while updating the temperature parameters. This work focuses entropy-based MARL

Novelty Identification: The novel findings or significant contributions presented in the paper are:\- The paper proposes a new framework, ADER, for adaptive exploration in MARL. ADER is able to handle the multi-agent exploration-exploitation trade-off. ADER is able to learn the adequate amount of exploration for each agent based on the degree of required exploration.

Experiment Claim Support: Claim 1:** The paper proposes an adaptive entropy-regularization framework (ADER) for multi-agent reinforcement learning (MARL) to handle the multi-agent exploration-exploitation trade-off. **Supporting Evidence:** The paper presents the ADER framework, which is designed to tackle the multi-agent exploration-exploitation trade-off in MARL. The framework incorporates entropy regularization with adaptive target entropy learning for each agent, enabling it to balance exploration and exploitation effectively. ......

Experimental Sufficiency: The experiments are sufficient because they cover a variety of tasks, including continuous action tasks, StarCraft II, and an ablation study. They show that ADER outperforms other state-of-the-art baselines in terms of both training speed and final performance. The experiments are sufficient to demonstrate the effectiveness of the proposed ADER framework for multi-agent reinforcement learning

Decision: **Major Weaknesses:** None **Minor Weaknesses:** The paper could be improved for clarity and organization. The paper does not provide enough implementation details to check for technical mistakes from a machine learning point of view. The paper does not provide enough information to determine if the experiments are reproducible. **Overall:** The strengths of the paper outweigh the weaknesses. The paper presents a novel and significant framework for adaptive exploration in thatMARL. The experiments show that ADER outperforms state-of-the-art baselines on various tasks. Therefore, the paper should be accepted.

Table 7: Structured Review after 4 iterations of refinement

Summary of the paper: The paper proposes an architecture and a training methodology (termed RoCourseNet) to generate a robust counterfactual (cf) along with the prediction for a given factual point. RoCourseNet builds on earlier work CounterNet [1] by modifying its objective to generate robust cfs i.e cfs which stay valid even when the underlying model shifts. As part of the RoCourseNet objective, the inner 'adversary' itself is proposed as a bilevel problem (called VDS in the paper). The paper proposes to learn a 'worst-case' classifier by looking at how the training dataset can change such that a classifier learnt on this new dataset maximally invalidates the old cfs. Experiments are performed on 3 real-world datasets and they compare against 4 baselines.

Strengths and Weaknesses of the paper: + The paper deals with a practical problem; generating robust recourses is necessary for models which are to be deployed in the real world
+ RoCourseNet outperforms the baselines convincingly in generating robust recourses for the 3 datasets considered
+ RoCourseNet works with the full model and not its locally linear approximation (via LIME etc.) which allows it to model larger number of model shifts via the VDS algorithm

- RoCourseNet involves a tri-level optimization problem. How much additional computational effort does ReCourseNet require? A comparison of the training time taken vs CounterNet seems necessary.
- The method lacks some flexibility of post-hoc counterfactual generation methods. Ex, different people have different notions of cost (proximity) or actionability. Can RoCourseNet solve this without retraining?

Other points:
In Algorithm1 VDS line (8) how is this gradient w.r.t \\delta computed? Is the only dependence of \\delta through \\theta(\\delta)?
Although not completely fair, a comparison of the training time w.r.t ROAR [2] may also be instructive.
Cite the published version of ROAR
[1] Hangzhi Guo, Thanh Nguyen, and Amulya Yadav. Counternet: End-to-end training of counterfactual aware predictions. In ICML 2021 Workshop on Algorithmic Recourse, 2021.
[2] Upadhyay, Sohini, Shalmali Joshi and Himabindu Lakkaraju. "Towards Robust and Reliable Algorithmic Recourse." NeurIPS (2021).

Clarity, Quality, Novelty, and Reproducibility: The paper is written clearly and is easy to follow. Code and implementation details are provided for reproducibility. The paper builds on existing work (CounterNet [1]), and novelty is in learning the 'adversarial model' for which they propose the VDS algorithm.

Summary of the review: The paper solves an important problem. The experimental protocol and results are convincing. The main issue I have is I feel the method is computationally expensive, and it lacks some flexibility that post-hoc cf-generation methods have. I propose acceptance, conditional on some time-complexity analysis.

Correctness: 4: All of the claims and statements are well-supported and correct.

Technical Novelty and Significance: 2: The contributions are only marginally significant or novel.

Empirical Novelty and Significance: 2: The contributions are only marginally significant or novel.

Flag for Ethics: No

Recommendation: 6: marginally above the acceptance threshold

Table 8: Example of Human Textual Review

Goal Determination: The paper proposes an architecture and a training methodology (termed RoCourseNet) to generate a robust counterfactual (cf) along with the prediction for a given factual point.

Motivation Logical: None

Literature Review: None

Clarity & Ambiguity Resolution: None

Summary of the paper: The paper proposes an architecture and a training methodology (termed RoCourseNet) to generate a robust counterfactual (cf) along with the prediction for a given factual point. RoCourseNet builds on earlier work CounterNet [1] by modifying its objective to generate robust cfs i.e cfs which stay valid even when the underlying model shifts. As part of the RoCourseNet objective, the inner 'adversary' itself is proposed as a bilevel problem (called VDS in the paper). The paper proposes to learn a 'worst-case' classifier by looking at how the training dataset can change such that a classifier learnt on this new dataset maximally invalidates the old cfs. Experiments are performed on 3 real-world datasets and they compare against 4 baselines.

Novelty Identification: The paper builds on existing work (CounterNet [1]), and novelty is in learning the 'adversarial model' for which they propose the VDS algorithm.

Experiment Claim Support: The experimental protocol and results are convincing.

Experiment Sufficiency: The authors should compare the training time of RoCourseNet to CounterNet and ROAR. the authors should explore the flexibility of RoCourseNet to solve different notions of cost or actionability without retraining.

Decision: Accept

Table 9: Initial Structured Review converted from the human textual review using α

Goal Determination: The paper proposes an architecture and a training methodology (termed RoCourseNet) to generate a robust counterfactual (cf) along with the prediction for a given factual point.

Motivation Logical: Yes, the motivation behind the paper is logical. The authors observe that existing approaches to recourse generation assume that the underlying model remains stationary over time. However, in practice, models are often updated regularly when new data is available to improve predictive accuracy on the new shifted data distribution. This shifted model might render previously recommended recourses ineffective (Rawal et al., 2020), and in turn, diminish end users' trust towards our system. The authors propose RoCourseNet, a training framework that jointly optimizes for predictions and recourses that are robust to future data shifts, to address this problem.

Literature Review: The paper has not cited all relevant work. It does not cite Pawelczyk et al. (2020a) who analyze the model multiplicity problem, which studies the validity of recourses under different ML models trained on the same data, and Black et al. (2022) who propose methods to ensure consistency under the model multiplicity setting.

Clarity & Ambiguity Resolution: The paper is well-written and organized. However, there are a few sections that could benefit from further clarification. These sections are the section on the proposed method and the section on the experimental results.

Summary of the paper: The paper proposes an architecture and a training methodology (termed RoCourseNet) to generate a robust counterfactual (cf) along with the prediction for a given factual point. RoCourseNet builds on earlier work CounterNet [1] by modifying its objective to generate robust cfs i.e cfs which stay valid even when the underlying model shifts. As part of the RoCourseNet objective, the inner 'adversary' itself is proposed as a \bilevel problem (called VDS in the paper). The paper proposes to learn a 'worst-case' classifier by looking at how the training dataset can change such that a classifier learnt on this new dataset maximally invalidates the old cfs. Experiments are performed on 3 real-world datasets and they compare against 4 baselines.

Novelty Identification: The paper builds on existing work (CounterNet [1]), and the novelty is in learning the 'adversarial model' for which they propose the VDS algorithm. RoCourseNet makes the following three novel contributions:

1. (Formulation-wise) We formulate the robust recourse generation problem as a tri-level (min-max-min) optimization problem, which consists of two sub-problems: (i) a bi-level (max-min) problem which simulates a worst-case attacker to find an adversarially shifted model by explicitly simulating the worst-case data shift in the training dataset; and (ii) an outer minimization problem which simulates an ML model designer who wants to generate robust recourses against this worst-case bi-level attacker. Unlike prior approaches, our bi-level attacker formulation explicitly connects shifts in the underlying data distribution to corresponding shifts in the ML model parameters.
2. (Methodology-wise) We propose RoCourseNet for solving our tri-level optimization problem for generating robust recourses. RoCourseNet relies on two key ideas: (i) we propose a novel Virtual Data Shift (VDS) algorithm to optimize for the inner bi-level (max-min) attacker problem, which results in an adversarially shifted model; and (ii) inspired by Guo et al. (2021), RoCourseNet leverages a block-wise coordinate descent training procedure to optimize the robustness of generated recourses against these adversarially shifted models. Unlike prior methods (Upadhyay et al., 2021), we require no intermediate steps in approximating the underlying model or data distribution.
3. (Experiment-wise) We conduct rigorous experiments on three real-world datasets to evaluate the robustness of several popular recourse generation methods under data shifts. Our results show that RoCourseNet generates highly robust CF explanations against data shifts, as it consistently achieves >96% robust validity, outperforming state-of-the-art baselines by $\sim$10%.

Experiment Claim Support: The experiments conducted by the authors support the claim that RoCourseNet outperforms state-of-the-art baselines in generating robust counterfactual explanations. This is evident from the fact that RoCourseNet consistently achieves >96% robust validity, which is $\sim$10% higher than the best performing baseline.

Experiment Sufficiency: The experiments conducted by the paper are sufficient to support the claims made in the paper. However, the following additional experiments could be conducted to further strengthen the claims:
Ablation experiments to study the impact of individual components of RoCourseNet on its performance.
Experiments with more datasets to demonstrate the generalizability of RoCourseNet's approach.
A human study to evaluate the quality of the explanations generated by RoCourseNet.

Decision: "**Major Weaknesses:** * Insufficient literature review **Strengths:** * The paper presents a novel approach to generating robust counterfactual explanations. * The proposed method is able to generate explanations that are robust to data shifts, which is a significant improvement over existing methods. * The paper provides a comprehensive experimental evaluation of the proposed method, which demonstrates its effectiveness on a variety of datasets. * The paper is well-written and organized, and the writing is clear and concise. **Decision:** The paper should be **rejected**. The paper has more weaknesses than strengths, and the major weaknesses are significant.

Table 10: Structured Review after 4 iterations of refinement