Note May 3rd 2022:

This document contains the original guidelines used by annotators in producing the DISAPERE dataset, including records of changes that were made during the annotation process. Much of the original terminology used during the annotation period was modified for clarity in the version of the manuscript that was submitted. If you would like to annotate additional data using DISAPERE guidelines, the authors would be happy to assist with any questions or clarifications. Please feel free to reach out.

Table of contents

An overview of the review - rebuttal process

We are doing two tasks. A review is a simple document composed of the reviewer's opinions about a paper, which is written both for the sake of the Paper's author, and for the sake of the other reviewers, including the "area chair" who reads all the reviews and decides upon a final accept or reject.

Reviewers end up therefore playing two roles, that of **Gatekeeper --** maintaining the quality of the venue by determining whether the paper "meets the bar" -- and of **Supporter,** supporting the author's science by providing constructive feedback to improve the paper. They do this by making evaluative statements (about what is good or bad about the paper), as well as requesting information, additional experiments, and other changes to the paper, in order to improve their own understanding or improve the paper itself.

In this review setting, authors can then provide a **rebuttal**, which allows the author to respond to the problems and questions from the review. They might promise to make changes to the paper (or, in open review settings, might actually update the paper during the review process) and might otherwise hope to answer questions.

What we are annotating

Imagine we have a tiny example document:

Review:

Reviewer1: This paper is impressive. Its results are the best. The writing is clear. However, they compare only to a random chance baseline.

Rebuttal:

Authors: Thank you for the review. We agree that our results are impressive. Moreover, we did not only compare to random chance; Table 4 shows the results of other models.

We want to get four pieces of information:

- What are the evaluative statements the reviewer is making, and what requests are they making?
- When these statements imply good or bad things about the paper, what aspects of the paper are they describing as good or bad?
- When we read the rebuttal, what parts of the review are being responded to?
- When the rebuttal responds to a part of the review, is it agreeing or disagreeing with the review?

A small example of annotations

For practical reasons, we break down every review and each rebuttal into *individual sentences*, and we have a set of labels for reviews and rebuttals. In our toy example, the review would thus get the following labels, which will be described below:

Reviewer1:
This paper is impressive. [Evaluative ; "substance" aspect = positive]
It's results are the best. [Evaluative ; "substance" aspect = positive]
The writing is clear. [Evaluative ; "clarity" aspect = positive]
However, they compare only to a random chance baseline. [Evaluative ; "meaningful comparison" aspect = negative]

When annotating a rebuttal, we will link each rebuttal sentence with zero or more sentences in the review which they are responding to, and give a rebuttal label to each:

Review	Rebuttal	Rebuttal Types
This paper is impressive. It's results are the best. The writing is clear. However, they compare only to a random chance baseline.	Thank you for the review. We agree that our results are impressive. Moreover, we did not only compare to random chance; Table 4 shows the results of other models.	Social Accept Praise Reject Criticism

Review Annotation

Overview of the Review Annotation Phase

The reviewer has an agenda of presenting their opinion of the paper to the area chairs (who make the final accept/reject decision). Approximately, reviewers tend to view themselves as filling one of two roles:

1. **Gatekeeper** - maintaining the quality of the venue by determining whether the paper "meets the bar"
2. **Supporter** - supporting the author's science by providing constructive feedback to improve the paper.

The reviewer generally carries out their agenda using arguments, which fall into one of three types:

1. **Request** - a request for information or change in regards to the paper
2. **Evaluative Statement** - a subjective judgement of an Aspect of the paper
3. **Fact** - an objective truth, typically used to support a claim

In this task, we will also label statements that are not arguments. These include two types:

1. **Structuring** - text used to organize an argument

Argumentative Test

Does this sentence directly express the reviewer's judgment about the paper?

Yes : It is likely to be Argumentative.

No : It is likely to be Non-argumentative.

2. **Social** - non-substantive text typically governed by social conventions

We call these five labels **Coarse statement types**. Each of these types may have sub-types and other associated labels. The other labels include:

1. Requests and Structuring labels will have **subtypes**
2. **Aspect** - which qualities (e.g. originality, correctness…) of the paper a statement is commenting on
3. **Polarity** - whether comments on aspect are positive (recommending acceptance) or negative (recommending rejection)

These are summarized in the table below, and are described in the sections that follow:

Category	Coarse statement type	Has sub types?	Aspect	Polarity
Argumentative	Request	Yes (see below)	🆗	🆗
	Evaluative	N	✔️	✔️
	Fact	N	-	-
Non-argumentative	Social	N	-	-
	Structuring	Yes (see below)	-	-

✔️: Required 🆗: Allowed

Sub-types of Request:

Request category	Request type
Response in manuscript	Typo
	Edit
	Experiment
Response in rebuttal	Clarification
	Explanation

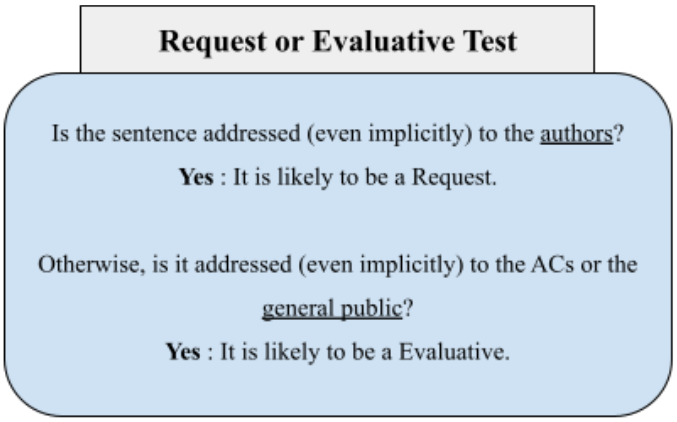
Sub-types of Structuring:

Structuring type
Summary
Heading
Quote

Statement types and Subtypes

Argumentative statements

Request



Definition: A Request is a statement that expresses a suggested course of action, including things from simple typo fixes to requesting further experimentation.
Additional Labels: You must add **Aspect** and **Polarity**, as well as a request subtype, as follows:

Subtypes:

1. Manuscript subtypes: requests which involve making changes to the manuscript
    a. **Typo**: Requires authors to fix a typo in the manuscript . “Typo” should be viewed as something uncontroversial, which might be fixed very quickly -- e.g. spelling changes.
    b. **Edit**: Requires authors to edit the text in the manuscript (including adding citations)
    c. **Experiment**: Instances in which the reviewer asks for further experiments or other results. If they simply ask the author to reformat existing results, use “edit”

2. Rebuttal subtypes (requests for information): requests which can be fulfilled by responding in the rebuttal. These are usually questions, but could be framed as direct requests for clarification. We distinguish between:
    a. **Clarification**: A question that asks for clarification of the meaning of some item in the manuscript. This encompasses clarification on factual details -- results, model details, etc.
    b. **Explanation**: A question that asks for explanation of some aspect of the manuscript -- usually for the reasons why a particular choice was made.

Statement Type Decision boundaries Examples

Evaluative

Definition: A sentence that expresses the reviewer’s judgment or opinion, and does not ask the authors to carry out any task, answer a question, or suggest a course of action.
Additional Labels: You must add **Aspect** and **Polarity**.

Statement Type Decision boundaries Examples

Fact

Definition: A generally incontrovertible statement; usually stating something that (from the perspective of the reviewer) is a mathematical fact, commonly held knowledge, or a factual statement about the paper.
Additional Labels: None

Tests and Notes:

- If a seemingly “factual” statement implies problems with the paper, use “evaluative” or “request”.

Statement Type Decision boundaries Examples

Non-argumentative statements

Social

Definition: These are sentences such as greetings or shows of appreciation, which do not contribute to the reviewer's argument.
Additional Labels: None

Tests and Notes:

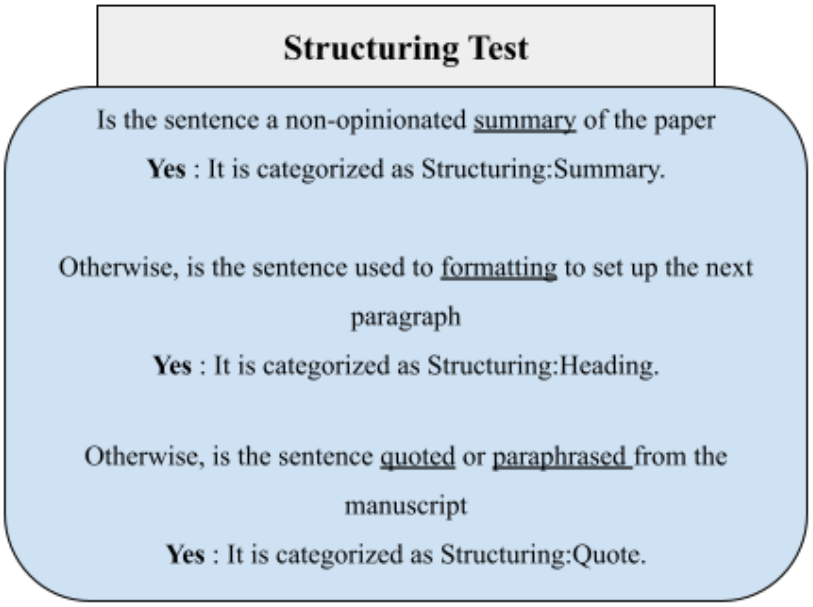
- Use "Social" for any "interactional" work referring to the authors or even to other reviewers.

Statement Type Decision boundaries Examples

Structuring

Definition: These are sentences that are included to organize the review, such as headings or quotes from the paper
Additional Labels: No aspect; select subtype below, based on the test

1) Summary
2) Heading
3) Quote



Statement Type Decision boundaries Examples

Statement Type Decision boundaries

- **Evaluative vs Factual**
  - **Facts that are leading up to evaluative statements:** The common "edge case" involves statements of fact which are laying the groundwork for a criticism. If a fact only implies something negative about the paper in context -- e.g. it asserts a general statement that is later used to show that the paper is wrong -- then treat it simply as a fact. If, in contrast, facts presented as the final culminating assertion for why an approach is good or bad, use evaluative:
    - Fact: The main reason that CMA-ES can explore better come from the randomness of parameter generation (line 2 in Algorithm 2).
    - Evaluative: For some dataset, this is beyond a spot, it could actually be a huge portion of the input space!
  - **Facts that discuss novelty:** If something discusses a method which may or may not be novel, you'll mark "evaluative" if there's any implication that it's unoriginal or incremental. If it simply says "X is an extension of Y" or "X is based on Y" without comment, you don't need to assume that it's a statement about the originality of the work:
    - Fact /Structuring: 2. The paper proposes a metric for the saliency map naming FSM which is an extension of existing metric SSR.
      -
    - Evaluative: The work is rather incremental from current state-of-the-art methods.
- **Request vs Evaluative:**
  - **Implicit requests for missing content:** Negative statements about missing results or missing analysis can be viewed as a request for those results, but don't go overboard: only do this when it feels pretty clearly like a "request"; e.g. phrases like "it would be great if you could do X" are pretty commonly treated as request in English:
    - Request.Experiment: Lack of theoretical analysis. It could have been nice if the authors could show the observed phenomenon analytically on some simple distribution. (substance_negative; analysis)
  - **Rephrasing:** We're going to ask you to make a judgment call whenever a reviewer says things like "It seems more appropriate to say that _____". Often these are really evaluative statements: if a reviewer says "It's more appropriate to say that they just took Smith et al. and tweaked the loss function", they're mostly emphasizing some weakness in the paper's originality, rather than literally suggesting that the paper's framing should be rewritten. Other situations will actually be asking the author to rewrite a text. Pick "Request.Edit" or "evaluative" based on whether it seems like the statement would be "addressed" by adding that text to the paper.
  - **Deontic statements** (i.e. "one should not X", "you should generally Y"): choose between Request.Edit and Evaluative (often +soundess/correctness), based whether a complaint is being presented as a fixable issue or not. You can use your own world knowledge on this if you have to, but don't dig too deep: If you can't tell, assume that it's just Evaluative.
-
- **Types of Requests:**
  - **Request.Explanation vs Request.clarification**: requests for explanation should ask *why* something was done a certain way (i.e. ask for motivation), or should be asking why a particular result occurs (asking for speculation).
  - **Request.Experiment vs Request.Explanation:** Use "Explanation" for requests for explanation regarding "why did you do X" or "why didn't you do Y" unless it seems to genuinely express that those experiments should be done\.
    - Request.Explanation : 2. What's the purpose of larger budget? You choose a bigger iteration budget than origin PPO implementation.

- **Request.Experiment with Speculative Musings:** Use the “request.experiment” label for questions about experiments that might be viewed as soft requests -- e.g. “I’d be curious to see what would happen if you applied this to dataset X” .

Aspects and Polarity

Aspect of statements are labeled according to ACL review guidelines. (The descriptions below are from Yuan et al. 2021). All sentences with an aspect should also have a Polarity.

- **Motivation/Impact**: Does the paper address an important problem? Does the review sentence discuss the importance of the task?
- **Originality**: Are there new research, topic, technique, methodology, or insight?
- **Soundness/Correctness**: Is the proposed approach sound? Are the claims in the paper convincingly supported?
    - Use this for statements (NEG) in a review that suggest that methods are anomalous
    - Assertions that a method breaks under various conditions can be considered a “soundness” issue
- **Substance**: Does the paper contain substantial experiments to demonstrate the effectiveness of proposed methods? Are there detailed result analysis? Does it contain meaningful ablation studies?
    - “Impressive performance” or “insufficient performance” counts as substance.
    - Statements about whether the experiments have good enough results, or whether the experiments/analysis result in clear answers, are also substance.
    - Requests for more experiments can often be viewed as a substance issue
- **Replicability**: Is it easy to reproduce the results and verify the correctness of the results? Is the supporting dataset and/or software provided?
- **Meaningful Comparison**: Are the comparisons to prior work sufficient given the space constraints? Are the comparisons fair?
    - Don’t use this for measuring impressiveness of the performance, even in comparison to other works!
    - This is for statements about whether you selected the right thing to compare against, be it the correct SOTA or good baselines.
- **Clarity**: For a reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?
    - This is for both writing comments as well as rhetoric and visualization elements -- complaints about the clarity of a table or figure are also “clarity”.

Aspect Decision Boundaries

- **Substance vs Soundness:** Negative “substance” could be fixed with additional experiments (you aren’t asking enough/ the right questions), while negative “soundess” implies something about the theory or implementation is wrong, or that the conclusions are incorrectly drawn from the current experiments.
    - Substance NEG The synthetic data experiments could also have been repeated on larger document sets for better understanding of model behavior
    - Substance NEG Other such experiments and analysis would be very helpful.
    - Substance NEG I think several important experiments are missing
    - Soundness NEG This is not entirely supported by the results in Tables 2, 3 and 4.
    - Soundness NEG The sigmoid activation function assumption is unprincipled
    - Soundness NEG My main concern is about the feasibility of using a neural network to learn cumulative quantities.
- **Soundness vs Meaningful Comparison:** Treat discussion of the choice of baselines as part of meaningful comparison, not soundness.

- **Substance vs Meaningful Comparison:** When discussing results; use meaningful comparison when we care about whether we are comparing to the right works, and substance when simply judging current results.
    - *Meaningful Comparison POS: The authors make a good effort at comparison their model to alternative architectures, such as a generic RNN and CNN+LSTM.*
    - *Substance POS: The paper provided results from multiple molecular optimization tasks.*
    - *Meaningful Comparison NEG: Also there is no comparison with CoGAN, which I believe is the most relevant work for coupled image generation.*
    - *Substance NEG: The system performance can be further improved.*
- **Replicability vs Clarity:** The main edge case here is with technical ambiguity -- lack of details in a method. If a sentence merely asserts that it lacks a particular detail, use “replicability”. Switch to clarity if sentences more generally describe a description as “unclear”: i.e. there may be sufficient details to understand the approach, but they are not presented clearly.
    - *Replicability NEG "Overall, I think the paper makes sense to me, but several details need to be verified."*
    - *Replicability NEG “I am not even clear how the system is trained”*
    - *Clarity NEG: "The authors could try to give some more intuition of what's happening"*
    - *Clarity NEG: "IN section 2.1, a large number of notations are introduced"*
- **Substance vs Impact:** Use “Impact” when discussing the *value* of succeeding at the intended task, and substance when discussing *whether* they succeeded at that task. “Good results at task X” are substance, “the downstream impact of task X” is impact. If something is very vague and just saying “this is interesting”, you can default to “impact”.
    - Impact POS The new WikiText language modeling dataset is very interesting
    - Impact POS This paper approaches aa problem that is not well studied in the literature
    - substance POS The experimental results seem impressive
    - Substance POS The many experiments conducted in the paper support these claims

- **Originality vs Substance**:
    - If a sentence discusses impressive results or comprehensive experiments, it will be clearly [substance]. The edge case between originality and substance occurs when discussing intellectual contributions. If a paper is “providing insights” but is not explicitly coded as providing a new approach or a new theory, label it as [substance]**:**
        - Substance NEG -- The analysis presented does not give new insights
        - Substance POS-- The connection to SMLC is interesting and it may contain a lot of insights.
        - Substance POS -- - Provides insights on why adversarial training is less effective on some datasets.
    - However, if it more directly discusses novelty, or specifically frames a contribution in terms of being an interesting idea or a theoretical contribution -- even if there are discussions of results as well -- then use ``originality'':
        - Originality POS I believe the experiments are novel and the results are interesting
        - Originality NEG If you just reduce all the negative advantage value to zero and calculate its gradient, the method is similar to just use half of step-size in policy gradient.
-

- **Soundness vs Clarity:** Sometimes people say “it is not clear that…” and it’s actually a question about soundness -- e.g. the implication that the equations or methods are not well-thought-out. If the thing that’s not clear could be resolved with rhetoric alone, mark it as clarity; if it seems to be implying “you didn’t explain X well because your idea is confused”, so with soundness:
    - *Soundness NEG: "I do not think you adequately explained why you chose to use aa GAN-like loss to learn these models."*
    - *Soundness NEG: The authors do not provide any explanation as to *why* language modeling is a better pretraining objective than translation.*

Polarity

Polarity of a statement should be labeled using one of the options below. We only label polarity for sentences which have an Aspect, i.e evaluative and request statements.

- **Positive**: Positively describes an aspect of the paper, which contributes to reasons the paper should be *accepted*
    - Requests should be “Positive” if rejecting the request likely does not have a negative connotation, but good answers could have positive results for that aspect. This will mostly occur with positive suggestions or requests for future work, e.g. “this is great, is there any chance one could use this for semantic parsing?” -- a negative answer is not really going to negatively impact the work.
- **Negative**: Negatively describes an aspect of the paper, which contributes to reasons the paper should be *rejected*
    - Requests should be “Negative” if they imply something negative about the paper if not corrected. One heuristic: if there is an innocent reason why this request/question cannot be satisfied or answered, does this likely count as a meaningful problem with the paper?
- **Neutral**: The statement does not seem to perceptibly commit to be ing positive or negative (this should be rare for evaluative statements)
    - For evaluative statements: Try to use this only when things are ambivalent; if at all possible, attempt to assign a positive or negative polarity.
    - For requests: use this for requests and questions whenever it’s not clearly negative, i.e. for many clarification questions.

Examples

Argumentative:Request
Argumentative:Evaluative
Argumentative:Fact
Non-argumentative:Social
Non-argumentative:Structuring
Request subtypes

Aspect

- **Motivation/Impact**
  - *POS: "The issue researched in this work is of significance because understanding the predictive uncertainty of a deep learning model has its both theoretical and practical value."*
  - *NEG: "The method seems limited in both practical usefulness and enlightenment to the reader."*
- **Originality**
  - *POS: "Novel addressing scheme as an extension to NTM."*
  - *NEG: "The reviewer believes that the idea of the paper is similar to the one in [1]."*
- **Soundness/Correctness**
  - *POS "The proposed method is sensible and technically sound."*
  - *NEG: "The required condition is rather implicit, and it is unclear how this condition can be checked in practice."*
  - *NEG: "There is not much theory to support the method." < ????*
- **Substance**
  - *POS "This is a thorough exploration of a mostly under-studied problem."*
  - *POS "The experiment section shows extensive experiment."*
  - *NEG "There are several modules introduced in the paper, but there isn't much analysis of them during the experiments."*
  - *NEG "The theoretical contribution is very limited."*
- **Replicability**
  - *POS "Release of the dataset and code should help with reproducibility."*
  - *NEG "There are some technical ambiguities."*
- **Meaningful Comparison**
  - *POS "The authors do a good job of positioning their study with respect to related work on black-box adversarial techniques."*
  - *NEG "Since the attention based aggregation is similar to GAT, a discussion on the difference is important."*
- **Clarity**
  - *POS "The paper is well-written and easy to follow."*

- *NEG "The presentation of the results is not very clear."*

Special Cases

When do I use the "add an argument" button?

Try to avoid using this! The instance where it's appropriate is when a sentence is summarizing a bunch of evaluative statements, and is saying multiple things at once.

- Experiments on real-world data do not validate the method, and the method itself is not novel. (Evaluative+Substance+NEG) + (Evaluative+Originality+NEG)

Sentences seem to be split in the wrong places!

Peer review data can be noisy --- there are numerous issues in this domain that make sentence splitting harder than you might expect. If you see a few instances of split sentences, try to label both sentences with the same label, and click the "tokenization error" button . If the whole document is full of them, feel willing to skip to the next document and let us know that you are skipping it.

Rebuttal annotation

What happens in a rebuttal?

The author's agenda is to assuage any doubts that the reviewer might have, and convince the area chair that the paper should be accepted despite the criticisms in the review. Authors tend to react to the reviewer's argumentative statements in one of two ways:

1. **Accepting** - agreeing that what the reviewer said is valid and true, and carrying out the course of action the reviewer suggested.
2. **Rejecting** - rejecting a claim, either as invalid or false, or declining to carry out the reviewer's suggested course of action.

Similar to the review annotations, rebuttal sentences either directly carry out the author's agenda (Argumentative), or support the statements that do so (Non-argumentative).

Rebuttal annotation is done in two steps: **selecting the context** of a rebuttal sentence in the review, followed by **annotating the label** describing the relationship between the rebuttal sentence.

Context selection

Sentence selection

How do we know what to select when there are many related sentences? We will provide heuristics for how to know how much to select, but this is also a slippery issue and you want to treat is as a negotiation -- we will all have intuitions about what we want to select, and you want to be alert for moments where you are unsure of

what to annotate, and bring those issues up -- often the guidelines won't cover a particular case! As some general starting points:

1. If a review has a general statement (e.g. “this is hard to understand”) which is followed by additional sentence fleshing those out. (review structure is <general statement>. <subpoint1> <subpoint2> <subpoint3>)
    a. Select **each subpoint** which is addressed by the rebuttal sentence
    b. If all are addressed, select the **general statement** as well
    c. If none are addressed specifically, but general statement is addressed, select **general statement only**
    d. If in doubt, default to selecting the whole thing (generalization + subpoints)
2. If a rebuttal responds to a review with multiple sentences, but you can view that entire paragraph/sequence as being a single response with a role like “reject-criticism::
    a. There’s no actual place in the annotation where you provide the “rebuttal paragraph label”, this is just useful to think about sometimes.
    b. When a rebuttal provides a fact supporting a larger point, label it with whatever larger paragraph label (e.g. reject-criticism) it’s contributing to.
    c. When a rebuttal provides a promise to do work (done-manuscript, future work, etc.), use that label rather than the larger paragraph
    d. **If you have two equally good options**, try to be consistent (i.e. matching boundaries used in prior rebuttal annotations), but don’t do incorrect annotations to achieve this consistency.
3. Even if only part of a sentence is relevant to a rebuttal, and if our tool is allowing you to select only a part of the sentence (i.e. there are “sentence splitting” errors), try to select a whole sentence rather than just a fragment.

Non-sentence context

In some cases, the context of a rebuttal sentence cannot be directly described as a subset of the review sentences. These cases fall into one of the following categories:

- **Global context** is used for assertions about the entire review -- thanking the authors, making a catch-all promise about fixing their suggestions, etc.
- **Context in Rebuttal** is used when a part of the rebuttal is adding auxiliary information to another part of the rebuttal, but cannot be viewed as part of a reply. This can range from internal signposting ("here are our answers to the questions:") to citations that support their points.
- **No context** is for things that neither engage with the review nor are referring to other parts of the text. This is rare, but is mostly seen when authors write a combined response and paste it to all the reviewers, resulting in some parts being irrelevant to each review.

Relation labels

These responses are categorized according to the argument type that they respond to. These categories are meant to help organize the possible responses, however, there is no restriction to cases in which they can be applied.

Accept Responses

- **Answer** - Answer to requests for explanation or clarification. Use this for all sentences within a paragraph answering a particular question, as long as particular individual sentences help the task.
- **Accept praise:** Thanking reviewer for positive statements (including elaborating upon them)

- **Concede criticism:** Acknowledging valid arguments or thanking reviewers for pointing out a problem.
- **Left to future work** - Authors promise to pursue requested experiments/directions but not in current work. Use this for general statements about future directions.
- **Already done** - Rebuttal states that a needed change in the manuscript has already been made in the time since submission (since submission)
- **Promised by camera ready ("by CR")** - Rebuttal promises changes, but implies that they would be made to this paper before publication, if accepted.
-

Reject Responses

- **Refute validity of question** - Explain why the underlying premise of the request is incorrect. Also use this if the request or question was already addressed by the submitted paper.
- **Reject Request** - Request is either infeasible (e.g. not enough time, not enough compute, ..) or otherwise cannot be done
- **Contradict Assertion -** Disagree with a general statement of fact
- **Reject criticism** - Disagreeing with basis of criticism; utterances pushing back against negative statements. Remember to use this label for all sentences in a sequence, if that series of sentences collectively shows a disagreement with a reviewer's criticism.
- **Mitigate praise** - This will be rare: statements adding caveats to positive assertions.
- **Mitigate criticism** - Accepting the technical validity of criticism but implying that the particular issue is unimportant in context.

Non-argument

- **Social** - Use this for all interactional labels, most commonly thanking reviewers
-
- - Use this for any part of the rebuttal whose role is to identify the part of the review that it's replying to. This encompasses quotes, but also any other sentences like "Regarding Table 4:"
- **Summary** - In the rebuttal, use this for either (a) summarizing over the rebuttal itself, or (b) when the rebuttal re-summarizes or re-states relevant parts of the original paper. (use "No context")

Other rebuttal statements

- **Multiple** - Use this in the case that a rebuttal sentence embodies more than one of the labels above. This should be rare.
- **Follow up** - Use when a rebuttal sentence neither accepts or refutes the reviewer, but asks for more information from the reviewer in return. Use this rarely -- when a follow-up question is rhetorical, or otherwise seems to be clearly another rebuttal type, then use that other rebuttal type.
- **Other** - Use this if none of the other labels apply to the rebuttal statement.

General Rebuttal label decision boundaries:

Different Edit labels: "Take done" vs "Task Will be done": In general, you can use "already done" if an author discusses changes in the past tense ("we have clarified the terminology in section C"), "by CR" if they use future tense ("we will clarify the terminology in section C"), and "future work" only when people explicitly note that something is left for future work. ("CR" means "camera ready", which is the term for the final version of a paper which gets officially published after all revisions are done.)

Concede criticism vs "Task Will be done/ Task Done": If someone accepts a suggestion or complaint and promises edits that will address it, you might feel that both "task done" and "concede criticism" both apply, but default to using the task done label (or "will be done", whichever is more relevant).

Reject-criticism vs summary/answer/etc.: If a rebuttal sentence provides factual information that's part of a disagreement with the criticism of a reviewer, use "reject-criticism" instead of non-reject labels.

Reject criticism vs reject question: Focus on how the rebuttal is framed, rather than what that rebuttal is responding to. If the rebuttal is explaining why a particular question cannot be answered, use "reject question" even if the original review is evaluative. Similarly, if a review asks a question but the reviewer pushes back against the evaluative stance implied by the question, use "reject criticism"

Rebuttal Examples

(Key: the left column shows relevant parts of reviews: green means the span you should select, red refers to incorrect spans that people might be tempted to select. The right column shows relevant parts of each rebuttal, and yellow illustrates the rebuttal sentence being annotated

REVIEW	REBUTTAL SENTENCE
- MAAC does not consistently outperform baselines, and it is not clear how the stated explanations about the difference in performance apply to other problems. - Authors do not visualize the attention (as is common in previous work involving attention in e.g., NLP). It is unclear how the model actually operates and uses attention during execution. Reproducibility - It seems straightforward to implement this method, but I encourage open-sourcing the authors' implementation.	*We have added a new section 6.3 to the supplement that includes visualizations of the attention mechanism both over the course of training and within episodes.* (label: done) Guide: select the smaller span here. (see rule 1a about span selection -- just getting the subpoint)
- The writing looks very rushed, and should be improved. For example, I have trouble understanding the sentence "So the existed algorithms should be heuristic or it can get a bad result even we train the neural networks with lots of datasets." in the introduction. - The aspect ratio in Fig. 5 should be fixed. 3) The experiments are completely preliminary and not reasonable: - The WGAN-GP baseline is very weak, i.e. does not show any reasonable generated images (Fig. 9). There are countless open pytorch implementations on GitHub which out-of-the-box produce much better results. - The shown inception scores are far from state-of-the-art. It is unclear, why one should use the proposed duality gap GAN.	*3)For the experiment: we will spend some time to train GANs with more iteration and modify it.* *Guidelines: Grab only the limited span (in green);* (see rule 1a above again) *Label disagreement: [by-cr_manu vs concede-criticism]* *Use **"by-cr_manu"** : although it's true that agreeing to make changes does **imply** that you are maybe conceding criticism, the promise to make changes is much more clear-cut.*

-- Minor: In Eq. 1, the utility is evaluated as the probability Yi=Yi'. What randomness is considered in this probability? In Eq 2, privacy is defined as maxmin of |Ii - Ii'|. Do you mean privacy guaranteed by the proposed method is different for each data? This should be defined as expectation over T or max over T. In page 4. "The reason we choose this specific architecture is that an exactly reversed mode is intuitively the mode powerful adversarial against the Encoder." I could not find any justification for this setting. Why "exactly reversed mode" can be the most powerful adversary? What is an exactly reversed mode? Minimization of Eq. 3 and Eq. 4 contradict each other and the objective function does not converge obviously. The resulting model would thus be highly affected by the setting of n and k. How can you choose k and n?	*We call Eq. 3 and Eq.4 adversarial, as explained in out intuition, they need not be opposite all the time.* For this, select the green bit (that's what's being responded to here), and code it as **reject-criticism** -- it's disagreeing with that specific claim that they contradict each other, so you can focus on that.

It is interesting that the analysis using this framework on simple examples is in line with known results in the GAN literature (Dirac GAN). Although I personally enjoyed reading the results that from control theory perspective are inline with GAN literature, the paper does not provide novel surprising results. For e.g. the results on the oscillating behavior of Dirac-GAN are described in related works (e.g. Mescheder et al. 2018), and in practice, WGAN with no regularization is not used (as well as GAN with momentum, as normally beta1=0 in practice). In my understanding, the authors present these results to justify the validity of the approach. However, this limits the novelty of the results relative to existing literature. The authors do not focus (in the main paper) on GAN variants used currently, and it is not clear if the proposed approach provides improvement relative to the current state of the art implementations (see next paragraph) . Moreover, if I understand correctly the WGAN analysis does not take into account that G and D are non-linear, and it is unclear if these can be done. I am also wondering if the comparison with the baselines is fair.	*About the main concern on "novelty, improvement relative to the current state of the art implementations, and non-linearity of G and D":... ….as some useful examples, in this paper,...we added new results in Table 1 in the revision, which shows that the same technique of negative feedback can further improve the state-of-the-art method of SN-GAN [*5]. Specifically, we apply NF-GAN to SN-GAN [*5] and NF-GAN provides a significant improvement on the state-of-the-art inception score on CIFAR-10 (from 8.22 to 8.45). Such results indicate that our technique of NF-GAN can still benefit the state-of-the-art variants of GANs (e.g., SN-GAN).* The whole rebuttal paragraph refers to this whole cluster of complaints -- about both non-linearity of G & D and various novelty issues -- you want to grab the whole set of sentences continuing those complains (the answers in green) Remember to use summary for sentences like this which aren't even part of the rebuttal of the parts of the review, but are providing new paper details relevant to that larger argument.

In other words, although in the present results, the proposed NF-SGAN/WGAN outperforms the baseline, the reported performance of the baselines is worse than in related works on CIFAR10. In particular, FID of ~30 on CIFAR10 for the baselines is notably higher then current reported results on this dataset (e.g. Miyato et al. 2018; Chavdarova et al. 2019). In my opinion, the authors could start from the existing state of the art implementations on this dataset, and report if negative feedback (NF) improves upon.	

- MAAC is a simple combination of attention and a centralized value function approach. Con - MAAC still requires all observations and actions of all other agents as an input to the value function, which makes this approach not scalable to settings with many agents. - The centralized nature is also semantically improbable, as the observations might be high-dimensional in nature, so exchanging these between agents becomes impractical with complex problems. - MAAC does not consistently outperform baselines, and it is not clear how the stated explanations about the difference in performance apply to other problems.	*Your thinking of 'semantically probable' exchange of information is interesting.* … *We note that it is possible to compress each agent's actions/observations before they are sent to a central critic.* **Contradict-assertion vs reject-criticism vs answer vs concede-criticism?** For both of these, they should be reject-criticism , since (when you look at it in context): - Sentence 1: it is using "interesting" to start disagreeing with the reviewer. - Sentence 2: When a rebuttal states general facts which are part of a disagreement, use the label that best describes that agreement -- i.e. reject-criticism here

5. One of the anonymous comments on OpenReview is very interesting: samples from a CIFAR model look nothing like SVHN. This seems to call the validity of the anomalous into question. Curious what the authors have to say about this. Minor nitpick: There seems to be some space crunching going on via Latex margin and spacing hacks that the authors should ideally avoid :)	*4. "Samples from a CIFAR model look nothing like SVHN. This seems to call the validity of the anomalous into question. Curious what the authors have to say about this."* *This is a very good point.* *See our response to Shengyang Sun's comment below.* *We see think this phenomenon has to do with concentration of measure and typical sets, but we do not yet have a rigorous explanation.* **answer vs concede-criticism: t**his should be

Summary

Sometimes these types will not line up, and that is ok -- language is flexible, and people interpret evaluative statements as requests all the time. But nevertheless, you can think of the rebuttal type as usually responding to particular review types. The following table summarizes all labels, and what they expect to respond to:

			Author action		
Category	**Coarse type**		**Accept**	**Reject**	**Other**
Argumentative	Request	Manuscript	Done since submission Promised by camera-ready Accepted for future work	Reject Request	-
		Rebuttal	Answer	Refute validity of question	-
	Evaluative		Accept praise Concede criticism	Mitigate praise Reject criticism	-
	Fact		-	Contradict	-
Non-argumentative	Social		-	-	Politeness
	Structuring		-	-	Summary Subheading Quote
	Other				Multiple Other

Appendix

Guidelines History

4/16/2021: semi-daily digest of questions and answers:

- **Can we finalize the guidelines? ??**
  - The guidelines are mature now -- that doesn't mean they won't change, but it does mean that we'll post regular sets of changes (like this), so that you don't need to re-read the guidelines to find changes.
- **I've been "assigned" 30 annotations -- does that mean I have to finish those this week?**
  - You don't need to do a fixed amount. Since it's an hourly appointment, the main thing is to report the number of hours you spend on the task, and to do as much during those hours as you can.
  - There are many more reviews to be assigned: try to let us know if you are close to going through the 30 currently assigned, so that we have some time to assign you more! You don't want to run out of work to do.
- **What to do about "=====" lines?:**
  - Use Structuring:heading!
- **for introductory sentences like 'This paper proposes ...'**
  - They would come under the structuring -> summary
- **What to do about egregious tokenization issues ?**
  - Check egregious tokenization and skip
- **So if an author asks for advice to the reviewer in the rebuttal, which label should we select?:**
  - "followup" label
- **What to do about Latex symbols?:**
  - Just work around it (assuming it's latex code expressing equations: don't worry about decoding it)
- **I noticed there are some places where the reviewers make a revision to comment on the rebuttal made by the author. It isn't technically an evaluative statement on the paper and I feel an additional tag to specify context to the rebuttal would be more helpful in that case.**
  - Default to "social" when they are just talking about the review process (e.g. something like "i've kept my score the same"), but if they actually make evaluative statements (like "I still have concerns about the validity of their baselines") feel free to label those as evaluative
- **I also had another query: If there are comments on meaningful comparison but it isn't about the main model/algorithm, would it be a meaningful comparison or a substance/fact tag ? e.g. A comment on a comparison between some parameter in the model and a parallel of that in humans.:**
  - *"I think the comparison between prior lifetimes and humans mastering a language doesn't hold up and is distracting"*
  - I'd agree that it doesn't pass our tests for meaningful comparison -- an argument could be made for something like "evaluative - clarity" but you can leave it as fact.
- **There's also some degree of vagueness at times between what classifies as "evaluation" & "request". … For instance — "I believe one should not compare the distance shown between the left and right columns of Figure 3 as they are obtained from two different models."**
  - **We added an expanded section in Statement Type Decision boundaries** on request vs evaluative. The main relevant one:
    - **Deontic statements** (i.e. "one should not X", "you should generally Y"): choose between Request.Edit and Evaluative (often +soundess/correctness), based whether a complaint is being presented as a fixable issue or not. You can use your own world knowledge on this if you have to, but don't dig too deep: If you can't tell, assume that it's just Evaluative.

- **Reference: Add reference label?**
    - If a review mentions citations that are missing, treat them as request.edit
    - If you're sure that a review is only providing citations for its own claims (not missing references provided to the authors), then treat that as "other"

4/12/2021: Big update

- We have stopped annotation of *grounding* labels in this current project, to simplify annotation
- Removed "Simple answer" request type (use "request+clarification" or "request+explanation")
- Reorganized taxonomy of rebuttal label types
- Merged "infeasible" and "not valid" rebuttal types

Terminology used in this document

Argument: A statement that expresses evidence or reasoning used to either oppose or support a given point.