Reviewing Advice

This is what I share with my student mentees any time they subreview a paper for me. It is tailored for *ACL, but I think the advice is generally applicable. As authors, we hate getting low-quality reviews on our papers, but I think it's also our responsibility to make sure we're not perpetuating the problem 😉

First, read through this, it's a pretty good guide:
https://aclrollingreview.org/reviewerguidelines

Some general guidelines:

- A good review should serve two purposes: 1) giving feedback to the authors on how they could improve their paper; and 2) giving the area chair (the person who will read the paper) a summary of the paper and what still needs to be addressed. Most people do (1) but forget about (2)
    - What this means is that make sure you are writing a review such that someone who has only read the title and skimmed the abstract can fully understand your review. Don't use acronyms without defining them, etc.

- The difference between "weaknesses" and "suggestions" is a bit subtle but important. People treat these as the same thing, but really weaknesses are much higher priority. Say you have issue X:
    - If the authors addressed X and it would make you more likely to want to accept the paper, put it under "weaknesses". These are usually issues in the experiments or conclusions, lack of discussions of prior work, and major issues in readability
    - If the authors fully addressed X and it would not change your assessment of the paper, put it under "suggestions". These are usually personal opinions, additional experiments, or extraneous things that are relatively unimportant. This guide from ARR can help a lot in disentangling these cases:
https://aclrollingreview.org/reviewerguidelines#-task-3-write-a-strong-review
    - Sometimes when writing a review, I like to explicitly say "My biggest problem with the paper is X", so the authors know exactly

- It's always good to do a check on if the current idea for the project has already been done. If something seems too simple to have not been tested or the authors are over-complicating what they're doing (i.e. you can't understand what they're writing), imo it's usually a sign that there has been some very relevant past work on this topic (very cynical, I know, but it's just my experience).

  - It's good to do a general search on Google Scholar with relevant keywords or smth like PaperFinder to quickly see if there's relevant literature. You're pretty junior but hopefully the AC or a more senior reviewer will know if there is existing work on the topic already
  - It's also a good time to recommend papers the authors should cite! This should go under "suggestions" unless it is an extremely critical citation (e.g. a missing baseline or paper that did the same thing)
  - For future Atrey, it's fine to recommend your own paper, but you should not only recommend your paper and should add other relevant literature that discuss the same thing. If the authors can guess who wrote the review, you wrote a really bad review lol

- Any time you state a weakness, it is extremely useful (and responsible) to add a sentence at the end of each bullet point explaining what you think the authors could do to alleviate this weakness, especially for the major ones. Like explicitly add "The authors could do XYZ to address this" (I often forget but try to do this)

Implementation-level Guidelines:

- I like to write the summary in paragraph form, around 5-6 sentences. So need to make this that long!
- The strengths, weaknesses, and comments/suggestions are better to keep as numbered lists. Later the authors will respond to your review, so it's useful if they can reference certain parts of your response later (e.g. The reviewer said [W1], so in response, we did XYZ…)
- Try really hard not to hedge and give a borderline score of 3. It's hard, but I think it's more useful for the AC if you give a stronger stance on whether you think the paper should be accepted or rejected, not just "maybe it can get in :)"

After writing a review:

- Please respond to your rebuttals! We hate when it happens to us, so don't become part of the problem!
- If you find yourself getting angry at the paper/rebuttal, it's probably a good idea to take a break. This happens to me, and it's great to be passionate, but sometimes that can get in the way of writing a good review.

At the bottom of this page I added two of my past reviews. They are anonymized and perturbed versions of two of my past reviews, as I did not want to call out any specific papers or authors. Please contact me (nbalepur@umd.edu) if you would like to see the full version!

Example Review #1 (lower score)

Paper Summary:

This paper studies the ability of LLMs to do X, which they deem "Y". They confirm 17 LLMs exhibit Y across 6 benchmark MCQA datasets and then disentangle three potential reasons behind this: 1) A; 2) B; and 3) C. Models can still score highly in A (reason 1) and B changes the model's answer (reason 2), so they conclude these are not indicative of X, but when they do C (reason 3), the model can do X much more accurately, so they attribute this as the primary factor (but I feel these conclusions are incorrect, see weaknesses). Based on this, they recommend Z.

Summary Of Strengths:

- Understanding Y is underexplored and has substantial consequences in LLM evaluation
- A large number of models, datasets, and hypotheses are considered to explain X (but I do not think all the hypotheses are fully sound, see weaknesses)
- The number of LLMs evaluated (17) is extensive, and well-above the bar for this type of analysis work
- Most of the figures were well-illustrated and easy to understand (barring some caption spacing, see suggestions)

Summary Of Weaknesses:

I am quite familiar with the first work on X [1], so most of my comments are based on differences between that work and this one. In particular, I find contradictions between the two works; it is okay if they do not say the same thing, but the authors should motivate why they made those decisions. I felt this discussion was necessary, since the two works have a very similar paper structure (i.e. discover X, then test hypotheses to explain why X happens):

1. The authors refer to this phenomenon as "Y", but this is not the correct term from prior work; an LLM being able to do X is not always a type of "Y", as argued by [1]. Standard terminology is often C, D, E, or F [2, 3, 4], so I would like the authors either need to explain why this is a Y or switch to one of these terms
2. The authors claim A [5] could explain X (L261), but I am failing to see why. In fact, since LLMs still often do X’ after A, it might actually suggest that LLMs are not exactly doing X (discounting the explanation of C proposed later). If the authors could further elaborate on this hypothesis in the rebuttal it would be useful.
3. I found the experiments on B extremely hard to understand. If the model does B, it may imply some self-recognition [6] or self-inconsistency [7] since the model favored its own generation, but I do not see how this is related to B. It seems to be more of a self-consistency check to me. Further, it was also unclear how B’ were generated; was it just an LLM prompt? In the rebuttal, I would appreciate it if the authors could provide a clearer definition of "B" and how it impacts X.
4. To test C, the authors do C’, and since C’’ happens, they conclude that C is the best explanation behind X. But importantly, correlation does not imply causation: when X and B both happen, it does not entail that X is a result of B. The authors do specify this is a correlation, but it is used to motivate their approach where they do Z, implying there is

causation. Indeed, this was a core point brought up in [1] that needs to be discussed more in-depth.

5. Finally, while the authors bring up filtering out MCQs where the model answered correctly just using the choices, this idea is not new. [8] uses the same protocol to do Z, while also weighing the model's confidence. This needs to be discussed in the paper

Overall, I think this work tackles an interesting and important problem, but I would like to see more rigorous hypotheses being tested and more connections to prior work that touch on similar concepts. I hope the authors find this feedback constructive as they revise the paper

References:

1. Redacted Paper
2. Redacted Paper
3. Redacted Paper
4. Redacted Paper
5. Redacted Paper
6. LLM Evaluators Recognize and Favor Their Own Generations
7. The Generative AI Paradox: "What It Can Create, It May Not Understand"
8. Redacted Paper

Comments Suggestions And Typos:

1. There are several grammar issues throughout the paper which might impact readability. For example, the section titles in 3.1, 3.2, and 3.3 are not valid questions. The question in the title is also not a grammatically correct question. I would recommend the authors put the paper through a grammar checker (e.g. Grammarly) before their next submission for readability!
2. You may want to cite more foundational and recent work discussing X

- Redacted Paper: Perhaps the first work on X, using BERT
- Redacted Paper: Explores whether LLMs do X on benchmarks
- Redacted Paper: A survey that includes progress in X that could be discussed

3. The authors write "Y*" in L306 but I'm assuming they mean "Y"
4. I personally think it's fine to use vspace to trim, but some margins are egregious (Table 4, Figure 7, Table 3). I would advise the authors to find other ways to save space.

Confidence: 5 = Positive that my evaluation is correct. I read the paper very carefully and am familiar with related work.
Soundness: 1.5
Excitement: 3 = Interesting: I might mention some points of this paper to others and/or attend its presentation in a conference if there's time.
Overall Assessment: 1.5 = Resubmit after next cycle: I think this paper needs substantial revisions that cannot be completed by the next ARR cycle.
Reproducibility: 2 = They would be hard pressed to reproduce the results: The contribution depends on data that are simply not available outside the author's institution or consortium and/or not enough details are provided.
Datasets: 1 = No usable datasets submitted.
Software: 2 = Documentary: The new software will be useful to study or replicate the reported research, although for other purposes it may have limited interest or limited usability. (Still a positive rating)

Example Review #2 (higher score)

Paper Summary:

This paper focuses on a specific but important design choice in multiple-choice question answering evaluation: [description of X]. The authors uncover that X can lead to statistically significant differences in accuracy; doing X is almost always more accurate and typically better calibrated on D across M LLMs. The authors also show that this issue persists with L across datasets

Summary Of Strengths:

- The paper is very well-written both in clarity and argumentation, and is especially well-scoped for a short paper. The paper was also generally entertaining to read (e.g. I liked the title of section 3)
- The thorough evaluation of models makes the result (i.e. we should do X) quite compelling. I feel usually many works that discuss prompt perturbations conclude with the result "changing the prompt can change the accuracy", but the authors focus on one specific perturbation and show that X leads to consistent gains, making the final recommendation more actionable and persuasive
- I feel calibration is often neglected in MCQA evaluation, so it was great to see that included, and hopefully will inspire more researchers to do the same

Summary Of Weaknesses:

I think the work is experimentally sound, so I don't have any problems there. However, I think the specific nature of this problem makes it feel less exciting (subjectively to me, of course) and thus may not have as much impact as it could have. So the weaknesses below are mainly to make the paper more appealing:

1. The main thing I want to know is whether X has different effects (larger/smaller, more/less consistent, etc.) compared to other perturbations. Essentially, is this just a reproduction of a finding that many of us know: LLMs are sensitive to prompts
2. The authors claim their findings are "robust and consistent" across datasets since L shows this issue across several datasets, but if the authors want to make this claim, they should test it across all of their models. If this finding does not hold for other models, they should discuss why
3. I understand the authors want to argue why this problem is important, but I think this came at the cost of overclaiming in certain areas. Apart from what I mentioned in (2), the authors claim this finding will impact "NLP and in an interdisciplinary context" (L77), but this supposes that MCQA is used in practical applications across disciplines (which certainly is not the case [1], and I'm assuming this is mentioned due to the special track). The authors also mention a few times that X is "fair" (L74, L301), but you could similarly argue that models are being treated unfairly if you do Y. Finally, even claiming in the title that this is a "large" effect in the title is a bit of an overstatement (if the maximum accuracy drop is N). The overall experiments are solid, so I would really just urge the authors to be careful with their word choice and make sure everything they claim is supported by citations or their experiments.

[1] The Shifted and The Overlooked: A Task-oriented Investigation of User-GPT Interactions

Comments Suggestions And Typos:

1. As a meta-comment, this was submitted to the Special Track (presumably interdisciplinary contexts for EMNLP), but this would fit much better and have better appeal in QA / Evaluation; I don't see how this work is interdisciplinary
2. Are there ways this finding could be applied to other tasks, settings, perturbations, etc., whether in evaluation or downstream applications? Adding this discussion would make the work more broadly appealing in my opinion
3. While not strictly necessary, there are other ways to design MCQA prompts (e.g. Y), and iterating over some of these different formats would be a simple way to address the limitation of a single prompt template in the limitations (L322) and make the work even more sound

Confidence: 5 = Positive that my evaluation is correct. I read the paper very carefully and am familiar with related work.
Soundness: 4 = Strong: This study provides sufficient support for all of its claims. Some extra experiments could be nice, but not essential.
Excitement: 2.5
Overall Assessment: 4.0 = Conference: I think this paper could be accepted to an *ACL conference.
Ethical Concerns:
There are no concerns with this submission

Needs Ethics Review: No
Reproducibility: 5 = They could easily reproduce the results.
Datasets: 1 = No usable datasets submitted.
Software: 2 = Documentary: The new software will be useful to study or replicate the reported research, although for other purposes it may have limited interest or limited usability. (Still a positive rating)