This work has been submitted twice before, we will summarize the overall reviewer's comments for each and how we addressed them.

2021 neurips (normal conference, i.e. non-dataset track)

openreview link : https://openreview.net/forum?id=NpL2pPJGKqP
overall comments:

1. as a dataset this is strong
2. as a method paper, the existing baselines are not novel, and performs poorly
3. unclear what sets this dataset (LARC) apart from other semantic parsing datasets
4. AC comment : "Moreover, I also know that NeurIPS explicitly created a datasets/benchmarks track this year. I think if that track didn't exist then I'd advocate for accepting this paper, but given that it does, I will recommend rejection. I do hope it gets into a future venue though."

our adjustments:

1. no changes as this is our strong point
2. made it clearer that it is not our intention to create SOTA on this exceedingly challenging dataset, but to highlight ways which traditional approaches perform poorly
3. added the distinction of "DSL-open" vs "DSL-closed" datasets, LARC is one of the very few DSL-open datasets, whereas all semantic parsing datasets are DSL-closed, making synthesis much more tractable. added in appendix explicit comparison between LARC and SCONE
4. submitting to the benchmark set to Neurips this year

2021 iclr (normal conference, as dataset track do not exist for ICLR)

openreview link :
https://openreview.net/forum?id=Z0XiFAb_WDr&referrer=%5Bthe%20profile%20of%20Yewen%20Pu%5D(%2Fprofile%3Fid%3D~Yewen_Pu1)
overall comments:

1. again a mis-match of conference fit, reviewers do not see this as a dataset paper, but insist on some SOTA results that we ourselves do not have (which is the point).
2. asked us to attempt some foundational model to this work, such as codex and clip

our adjustments:

1. submitting it to a dataset track for neurips this year so such mis-match no longer occurs
2. added preliminary studies using codex and clip in appendix, finding that they work poorly

Overall the issue of this work has been a mis-match of conference track, where it was assumed (implicitly and improperly) to be developing a SOTA method, whereas this paper is anything but that. We hope that this work will be well received in the dataset track, as we truly believe this is a highly valuable dataset that highlights the differences between how humans and machines differ when it comes to communicating and interpreting instructions.