

Peer Review

Review of: "Approximated Variational Bayesian Inverse Reinforcement Learning for Large Language Model Alignment"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The authors recast LLM alignment as a Bayesian inverse reinforcement learning problem. Instead of learning only the reward *difference* between a chosen and a rejected response (Bradley-Terry, DPO), they place a posterior over the reward itself and approximate it with the AVRIL objective of Chan and van der Schaar, adapted token by token to text generation. Two objectives come out of this. AVA-d (Eq. 5) learns from single demonstrations; AVA-p (Eq. 17, with the auxiliary CER term of Eq. 18) learns from preference pairs by treating the rejected response as a negative demonstration. Both are implemented on a "Transformer with Q-value and Reward Heads" (TQR, Figure 1), with attention-based reward weighting borrowed from the Preference Transformer (Eq. 11) and an option to initialise the Q-value head from the pretrained LM head via a log-softmax mapping (Eq. 12).

Empirically, the paper reports that AVA-p gives higher reward accuracy than Bradley-Terry and Pref-Trans with GPT-2 and BART-base backbones on Anthropic-Harmless, Anthropic-Helpful, and OpenAI-Summary (Table 1), that its reward model wins more often in Best-of-N sampling (Table 2) and as a PPO reward for Llama-2-7b-chat (Table 3), and that used as a direct-optimisation loss it beats DPO (Table 4) and, in the AVA-d form, AfD on Alpaca-GPT-4 and Math-GPT-4o (Table 5). The conclusion states that these results show "the alleviation of the reward hacking issue".

2. Whether the conclusion holds on the data

I want to start with what is genuinely good. The derivation is clean and well motivated: the step from the AVRIL ELBO (Eq. 2-4) to an NLG form (Eq. 5) is explained in prose, the contrastive extension CBIRL (Eq. 15-16) is a natural idea, and the paper is honest about the design compromise in Eq. 18, where it says plainly that CER uses only the last-timestep reward. The ablation design is also the right one. Five variants (w/o rwt, w/o neg, w/o irl, w/o cer, w/o ptq) isolate the components one at a time, and "w/o irl" is effectively a contrastive SFT control, which is exactly the comparison a reader wants to see. Testing the same objective in three pipelines (reward modelling, PPO, direct optimisation) is more than most alignment papers attempt.

The direction of the results is consistent. The size of the effects, and the evidence that they are not noise, is where I would ask for more.

Reward accuracy margins are often very small. In Table 1, column Harmless/gpt2, AVA-p scores 70.27 against Pref-Trans 70.26: a gap of 0.01 points. On Harmless/bart, the gap to Pref-Trans is $72.30 - 71.32 = 0.98$, and on Helpful/gpt2 it is $72.37 - 71.37 = 1.00$. The larger gains are on Helpful/bart (74.84 vs 72.37, +2.47) and Summary/bart (64.31 vs 59.27, +5.04). No test-set size, no standard deviation, and no number of seeds are given, and the checklist itself answers "no" to both measures of variation and statistical tests. Without those, the claim that AVA-p "surpasses Bradley-Terry and Pref-Trans in reward accuracy on all reported reward modeling tasks" is literally true, but I cannot tell whether a 0.01-point margin would survive a second seed.

Two pairs of rows in Table 1 look duplicated. I compared them cell by cell. "AVA-p w/o rwt" and "AVA-p w/o cer" are identical in the first five columns (70.06, 70.73, 69.81, 69.32, 60.55) and differ only in Summary/bart (58.89 vs 58.58). "AVA-p w/o neg" and "AVA-d" are likewise identical in the first five columns (70.54, 70.36, 69.75, 69.15, 62.06) and differ only in Summary/bart (58.65 vs 59.00). Two different training objectives agreeing to two decimals on five of six tasks is possible but unlikely. I would ask the authors to check whether these rows were copied by mistake. It matters for the text, too. The sentence "AVA-d achieves the best performance on 2 out of 6 tasks" rests on Harmless/gpt2 (70.54) and Summary/gpt2 (62.06), and in both columns "AVA-p w/o neg" has exactly the same value, so these are ties, not outright wins. The same two columns are also the ones where full AVA-p is *not* the best variant, so by my count AVA-p is best on 4 of 6 columns, which matches "the greatest number of tasks" but should be

stated with the count. Bradley-Terry also shows 59.27 in both Summary columns; that may be real, but it is worth a check.

Win rates are dominated by ties, and the judge is not described. I summed Win + Tie + Lose for every row of Tables 2-5; all are 100.0 within rounding (Table 3, Helpful/SFT gives 100.1; Table 3, Summary/SFT and Table 5, Math/AVA-p w/ AfD give 99.9). So the tables are internally coherent. What concerns me is the shape. In Table 3, row Helpful / PPO w/ Bradley-Terry, the result is 1.8 win, 97.2 tie, 1.0 lose; for Pref-Trans it is 2.6 / 95.8 / 1.6. In Table 4, Summary vs DPO, it is 44.9 vs 44.1. In Table 5, row Math / DPO w/ AfD, Win and Lose are both 44.3, an exact tie, so the sentence "AVA-d outperforms the AfD approaches" does not hold for that row. I could not find, anywhere in the Qeios text, who or what produced these judgements (human raters, GPT-4, a reward model), how many prompts were evaluated, or how ties were defined. With 97% ties, a 0.8-point difference could be a handful of prompts. I would not draw conclusions from these rows until the evaluation is specified and given confidence intervals.

The reward-hacking claim is not measured. The abstract and conclusion say that AVA reduces reward hacking. None of the reported metrics measures it directly: reward accuracy and win rate are both downstream quality measures. A KL-vs-reward curve during PPO, or gold-reward vs proxy-reward divergence in the style of Gao et al. (ref. 19), would turn this from an interpretation into a result.

The intermediate-reward argument and Eq. 18 pull in different directions. The introduction motivates the work by intermediate rewards, yet CER (Eq. 18) uses only the last-timestep reward, justified by refs 29-30. That is a reasonable engineering choice. But the ablation "w/o cer" then removes the only term that compares rewards between chosen and rejected answers, and on Table 1 its numbers are close to, sometimes identical with, "w/o rwt". I would like to see some evidence that the learned intermediate rewards carry information: for example, a plot of per-token rewards on a few examples, or reward accuracy computed on truncated prefixes.

Overall: the method is sound and novel enough, the trend in the tables is favourable, and the work deserves to be read. The quantitative claims need variance, evaluation details, and a check of Table 1 before "outperforms" can be stated without qualification.

3. Reproducibility

Model identities are clear and, helpfully, they are fixed open checkpoints: GPT-2 (117M), BART-base (140M), and Llama-2-7b-chat-hf (Experiment Setup, "Pre-trained Models"). Datasets are named, and the ablation variants are defined precisely. That is a good base.

Almost everything else is deferred to "our code and the Experiment Details section of the Technical Appendix", and neither is present in the Qeios version. I could not find a repository URL anywhere in the text. The ELBO derivation of Eq. 16 is also said to be in the Technical Appendix, so it cannot be checked here. Specifically missing from this version:

- learning rates, batch sizes, epochs, the prior on rewards, the weight of the TD-error term in Eq. 4/5, the CER weight in Eq. 18, and the temperature parameter in Eq. 12;
- PPO settings (KL coefficient, rollout length, number of steps) for Table 3, and the N used in Best-of-N for Table 2;
- decoding parameters for every generated sample that was judged;
- seeds and the number of runs per number (the checklist says "yes" to both, but I could not find them in the body);
- train/test splits and sizes for all five datasets; "Math-GPT-4o" in particular has no citation or source;
- hardware and compute (the checklist again says "yes").

As rendered on Qeios, the mathematics is duplicated character by character (each symbol appears once as text and once as a glyph), and I could not locate displayed labels for Eq. 8, 9, 10, and 13, although the text refers to them. Algorithm 2 is referenced, but its body is hard to separate from the surrounding text. This is a rendering issue rather than an author error, but it makes the method section hard to follow on this platform. A PDF with the appendix attached would solve most of it.

4. What specifically to fix

1. Table 1, rows "AVA-p w/o rwt" / "AVA-p w/o cer" and "AVA-p w/o neg" / "AVA-d". Please verify these four rows against the logs. If they are correct, say so explicitly in the text, since identical values to two decimals on five of six columns will make every reader suspect a copy error.
2. Table 1 text: change "AVA-d achieves the best performance on 2 out of 6 tasks" to say that in both columns (Harmless/gpt2, Summary/gpt2) it ties with AVA-p w/o neg.
3. Tables 1-5. Report the evaluation set size for each dataset and mean $\pm$ standard deviation over at least three seeds. For Table 1, a paired bootstrap interval on reward accuracy would be enough. This directly addresses the two "no" answers in the reproducibility checklist.
4. Experiment Setup, "Metrics": describe the win-rate judge (human or model, which model and version, the prompt used, position-bias control) and how a tie is decided.

5. Table 5, row Math / DPO w/ AfD (44.3 / 11.4 / 44.3): this is a tie. The sentence "AVA-d outperforms the AfD approaches" should be qualified for Math.
6. For the near-total-tie rows of Table 3 (Helpful, 97.2 and 95.8 tie), I would add the win rate among non-tied pairs together with the number of non-tied pairs; otherwise, the comparison carries almost no information.
7. Abstract and Conclusion: either soften "alleviation of the reward hacking issue" to a hypothesis, or add a direct measurement (proxy vs gold reward over PPO steps, or reward vs KL).
8. Eq. 18 and the intermediate-reward claim in the Introduction. Show at least one qualitative example of learned per-token rewards, or prefix-level reward accuracy, so the reader can see that intermediate rewards are meaningful and not only the last one.
9. Experiment Setup lists "AfD w/ Bradley-Terry, AfD w/ Pref-Trans, and AfD w/ AVA-p" for reward modelling, and AVA-d ablations ("AVA-p/d w/o rwt", "w/o ptq"). Neither set appears in Table 1. Add them, or remove them from the setup.
10. Attach the Technical Appendix (Experiment Details and the Eq. 16 derivation) and give a repository URL with a commit hash.
11. Small things. In the AVA-d section, "make the TD error of the Q-values 7 close to the reward" contains a stray "7". "Afd" should be "AfD". Give a source for Math-GPT-4o.

5. Recommendation

Minor revisions. The Bayesian IRL formulation of alignment is a real idea, carefully derived, and tested across three pipelines with a sensible ablation set. What holds it back is reporting: possibly duplicated rows in Table 1, an unspecified win-rate judge, no variance, and a reward-hacking claim that no table measures. None of this requires new methodology. I would expect the conclusions to stand, perhaps in a more measured form, once seeds, evaluation details, and the appendix are added.

Declarations

Potential competing interests: No potential competing interests to declare.