Strategic Misreporting

As noted in the paper, the Borda count is vulnerable to annotators intentionally misreporting their preferences to strategically influence the model's behavior. Although the same holds for DPL, there's no way to detect the strangeness of the situation in standard RLHF, whereas DPL can.

Suppose that we have two groups of annotators, Group A (90% of all annotators) and Group B (10% of all annotators) and two responses X and Y so that X is slightly preferred to Y by Group A whereas Y is slightly preferred to X by Group B. That is, suppose that

$u(r, z)$	Group $z = A$	Group $z = B$
Response $r = X$	1	0.9
Response $r = Y$	0.9	1

Also, suppose that $\hat{u}(r, A) = u(r, A)$, i.e., members of Group A answered truthfully while $\hat{u}(X, B) = 0$ and $\hat{u}(Y, B) = 100$, i.e., members of Group B massively exaggerated their preferences. In this scenario, both standard RLHF and risk-averse DPL would favor response X because of the majority bias. However, DPL would allow us to detect the high variance in the reported utility distribution whereas with standard RLHF, we cannot.

The fact that DPL can detect the high variance in the reported utility distribution allows us to identify and penalize strategic misreporting. Although in this setup we actually would want to show the response that the majority prefers, note that a different method of aggregating preferences could lead to the opposite outcome. For example, suppose that the aggregating method satisfied the desirable property of preference learning with hidden context converging to the expected utility for all alternatives, i.e., $\bar{u}(a) = E_{z\sim\mathcal{D}_z}[u(a, z)]$. Then, the aggregating method would favor response Y and we would not be able to detect anything suspicious going on, let alone regulate it with a risk-averse approach.

There's no such thing as a free lunch

Do note that risk-averse DPL has its downsides. For example, a risk-averse DPL would not show responses that mention the Pell grant in Example 1.1 whereas, on the other hand, non-risk-averse DPL would be more vulnerable to strategic misreporting in the setup described above.