Intellectual Merit Criterion

Overall Assessment of Intellectual Merit
Very Good

Explanation to Applicant
- Intersectional sociocultural bias in LLMs addresses an interesting and important point of view to look into bias, in which multiple social identities are overlapped and interdependent. - The application shows the capability of the applicant as a researcher in learning, thinking, working, collaborating, and contributing to the community. - In terms of probing, line probing could be a good start. Some other non-linear approaches may worth trying further. - RLHF fine-tuning depends on high-quality human annotations. This might be even harder for cross-culture aspect. Try to think in more details when building the entity actions dataset.

Broader Impacts Criterion

Overall Assessment of Broader Impacts
Very Good

Explanation to Applicant
The proposed work would enhance fairness in the deployment of LLMs which further benefits people with different sociocultural background.

Summary Comments
This application proposes research on intersectional sociocultural bias in LLMs which is important to ensure the fairness of AI systems based on LLMs, also shows the capability of the applicant to conduct it.

Intellectual Merit Criterion

Overall Assessment of Intellectual Merit
Excellent

Explanation to Applicant
You have already published/submitted a few different papers, and your proposed research idea seem very well thought out. You have also identified multiple research questions and came up with proposed solutions for both, backed by previous work (some of which is your own)! Your previous work as a researcher in different labs is also very impressive, and I commend you wanting to collaborate with people across disciplines. I couldn't find any weaknesses.

Broader Impacts Criterion

Overall Assessment of Broader Impacts
Excellent

Explanation to Applicant
I think it's important that you are recognizing that it is individuals who use these tools. One small weakness that I found was that I believe you should think about *how* the end user might actually use a system that you make. Consider in your future work how different sociocultural groups might use these tools. For example, depending on the application, will a white man use your system the same way as a Black woman would? And if not, how will you ensure that these are comparable and equitable experiences? Furthermore, you might want to think about how your system would determine what norms to use for a particular

user.

Summary Comments

I appreciate your ability to recognize the current issues of LLMs and how they can be problematic for minority groups, and you have also come up with clever solutions to your research questions. It seems like you have a very clear vision and that you would be ready and willing to go through with it.

Intellectual Merit Criterion

Overall Assessment of Intellectual Merit

Excellent

Explanation to Applicant

Strengths: 1. The student acquired ample research experience as an undergraduate and contributed to original work. It is clear that they are well-positioned to take on their research plan. 2. The proposed research agenda is well-positioned to advance knowledge at the intersection of NLP and social science. 3. The proposed research is novel. The student has identified clear gaps (i.e., isolating representation of social norms, rigidness of RLHF frameworks) and has proposed a direction to close these gaps. 4. The plan seems reasonable, and the evaluation plan is clear for at least 1 of the 2 approaches. I would rate the maturity of the proposal in line with those prepared by successful early graduate students. Weaknesses 1. Unclear what the mechanism to access success will be for Approach 2

Broader Impacts Criterion

Overall Assessment of Broader Impacts

Excellent

Explanation to Applicant

Strengths: 1. The student has been very proactive and has ample experience in peer-mentorship, tutoring and educational outreach. 2. The student was part of a research initiative that led to establishing a research unit for the democratization of open-source human preference learning. They have already demonstrated broader impact in their past work. 3. As more technology incorporates LLMs, being able to steer them in an inclusive direction is potentially transformative. 4. Since we lack clear understanding of what and how they represent information, the interpretability research agenda is potentially transformative as well.

Summary Comments

Student has a strong background, having engaged in multiple successful research efforts and maintaining a near-perfect GPA. The student also has demonstrated to be a very involved citizen in their community, and has acquired skills that will be invaluable in their graduate career (peer-mentorship, teaching). The research plan is both well-thought out, clear and high-impact. Letter writers all support the picture of a very promising scholar.