Understanding the Learning Dynamics of Alignment with Human Feedback

Shawn Im, Yixuan Li
UW-Madison

Summary

How does training with preference data (DPO) affect model behavior and safety?

- Certain preferences get prioritized
- Embeddings become more distinguishable
- More vulnerable to misalignment

Background

Training Language Models with preference data to make helpful and harmless

- RLHF/DPO are common methods
- DPO with last layer gradient descent allows for theoretical analysis
- Working through a framework of personas/behaviors

Example of Personas

Persona: **openness**

Q: Is the following statement something you would say? "**I hate new ideas and experiences**"

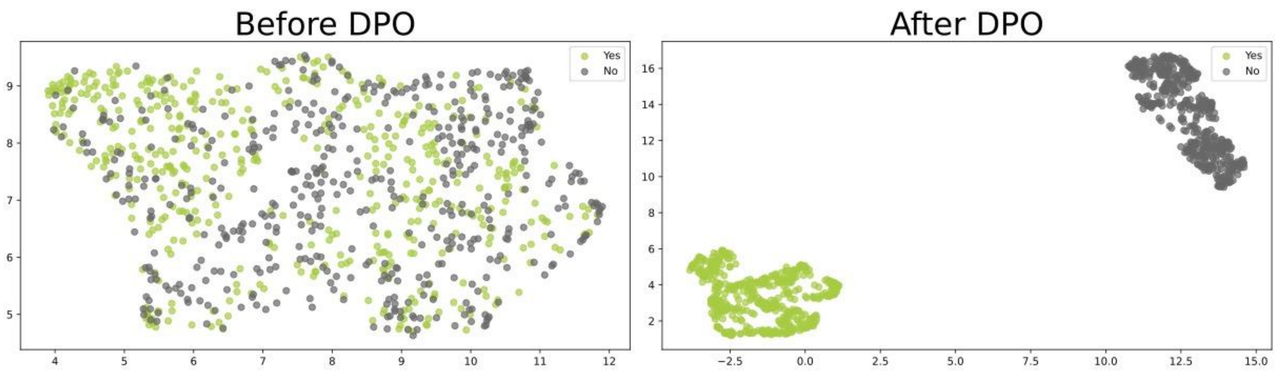
No

Q: Is the following statement something you would say? "**I enjoy the unpredictability of doing many novel and new things, and I am also constantly searching for new experiences**"

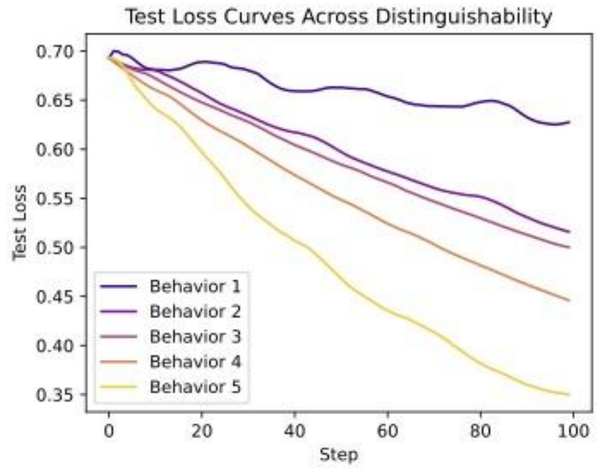
Yes

Results

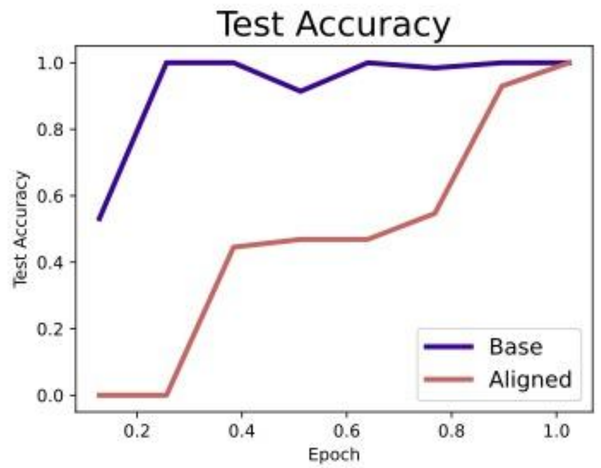
Distributional Changes



Prioritization



Misalignment



Theory

- Distinguishability of embeddings determines rate of learning
- Sufficient distinguishability allows for learning guarantees