Sparse Mixture-of-Experts Routing Impact on Multimodal RLHF Alignment Scores

Assignee Research

June 2, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does sparse mixture-of-experts routing in multimodal models influence alignment scores on RLHF benchmarks compared to dense baseline architectures. 19 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: RLHF-Blender: A Configurable Interactive Interface for Learning from Diverse Human Feedback. Research question: How does sparse mixture-of-experts routing in multimodal models influence alignment scores on RLHF benchmarks compared to dense baseline architectures?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.5/10.

3 Results

4 papers retrieved. 19 claims extracted; 0 independently verified. Quality review score: 2.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
RLHF-Blender consists of three major components: an interactive user interface, a feedback processor, and a consistent s	×	0.09
The RLHF-Blender interactive user interface enables browsing episodes or segments from a set of available state-action s	×	0.09
The RLHF-Blender user interface implements multiple feedback interactions that can be enabled or disabled dynamically.	×	0.06
The RLHF-Blender feedback processor consists of a sampling unit and a translator unit.	×	0.03
The RLHF-Blender feedback processor translates human feedback into a standardized format.	×	0.12
All three components of RLHF-Blender are highly modular to enable different combinations for human-subject studies.	×	0.07
RLHF-Blender supports an online training configuration where an RL agent is trained synchronously with a reward model.	×	0.06
In the online configuration, new trajectories are sampled by rolling out the online policy.	×	0.00
The preferred configuration for RLHF-Blender is an offline mode that uses pre-collected episode data to train reward mod	×	0.04
RLHF-Blender can dynamically serve episodes of different skill levels during an experiment session.	×	0.03
The system can replicate the reinforcement learning from human preferences setup described by Christiano et al. (2017).	×	0.11
For strict pairwise comparison in RLHF-Blender, the number of displayed episodes can be restricted to two.	×	0.02
RLHF-Blender allows for either random or progressive sampling of buffer elements when collecting human preferences.	×	0.04
RLHF-Blender supports a reward model architecture that receives single states as input and outputs a single scalar rewar	×	0.03
RLHF-Blender enables simultaneous multi-type feedback, such as ratings, ranking, and correction/action advice.	×	0.04
In multi-type feedback mode, participants can choose which feedback type to use for each episode.	×	0.04
RLHF-Blender automatically handles the translation of different feedback types into a standard encoding and subsequent l	×	0.08
RLHF-Blender supports training separate re	×	0.06

References

- http://arxiv.org/abs/2308.04332v1
- http://arxiv.org/abs/2407.17482v2
- http://arxiv.org/abs/2602.12375v1