Dense vs. Sparse Multimodal Model Alignment on VQAv2 and OK-VQA Benchmarks

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the alignment score (e.g., via RLHF or DPO) of a dense multimodal model compare to a sparse model with varying numbers of experts on the VQAv2 benchmark, and does this correlation hold for. Reinforcement learning from human feedback (RLHF) has emerged as the primary method for aligning large language models (LLMs) with human preferences. The RLHF process typically starts by training a reward model (RM) using human preference data. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts. Research question: How does the alignment score (e.g., via RLHF or DPO) of a dense multimodal model compare to a sparse model with varying numbers of experts on the VQAv2 benchmark, and does this correlation hold for OK-VQA reasoning accuracy?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

14 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
Most existing reward models for LLM alignment are trained with Bradley-Terry loss on pairwise data with annotated prefer	×	0.11
The pairwise preference annotations are essentially binary labels, e.g., {0, 1}, indicating which response is preferred	×	0.05
UltraFeedback is curated with 5-objective absolute ratings: Overall Score, Instruction Following, Truthfulness, Honesty,	×	0.05
Each objective in UltraFeedback has 5 distinct ratings based on pre-defined rubrics.	×	0.03
The dataset is further binarized into pairwise comparisons, using the Overall Score, or the average score of the remaini	×	0.06
The original ratings in UltraFeedback are fine-grained, as each objective has continuous integer rating scores (e.g., 1,	×	0.04
The binarization process discards some fine-grained information.	×	0.03
A pair of examples with scores 1:5 is labeled in the same way as another pair with scores 2:3.	×	0.03

References

- http://arxiv.org/abs/2406.12845v1
- http://arxiv.org/abs/2601.15021v1

- http://arxiv.org/abs/2603.12895v1