Bayesian Non-Negative Reward Modeling Alignment Performance Across Model Scales on HELM

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the alignment performance of Bayesian Non-Negative Reward Modeling scale with model size when evaluated on the HELM Holistic Evaluation Benchmark. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: RewardBench 2: Advancing Reward Model Evaluation. Research question: How does the alignment performance of Bayesian Non-Negative Reward Modeling scale with model size when evaluated on the HELM Holistic Evaluation Benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

10 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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
The authors trained Bradley-Terry reward models using the Open Instruct library.	×	0.06
The study varied hyperparameters including learning rate and number of training epochs.	×	0.04
The study examined multiple strong open-weight base models.	×	0.03
The study utilized two training data mixtures: Tulu preference mix and Skywork preference mix.	×	0.04
Llama 3.1 Instruct-based models performed strongly at both 8B and 70B scales in the authors' setup.	×	0.03
Larger reward models performed better on the RewardBench 2 benchmark than smaller ones.	×	0.09
Skywork training data was particularly helpful for focus and safety domains.	×	0.05
Tulu training data was better for the factuality domain compared to Skywork data.	×	0.05
Combining Skywork and Tulu data sources improved average performance compared to training on either dataset alone across	×	0.07
Qwen 2.5 7B Instruct-based models outperformed 70B reward models trained on Llama 3.1 70B Instruct and Tulu 3 70B SFT in	×	0.03
Capabilities conferred in post-training of base models carry over to the trained reward model.	×	0.08
Eight of the eighteen best models on RewardBench 2 were trained for two epochs.	×	0.07
PPO scores saturate for on-policy and in-distribution reward models.	×	0.05
PPO scores are significantly lower for off-policy or out-of-distribution reward models compared to on-policy models.	×	0.07

References

- http://arxiv.org/abs/2406.12845v1
- http://arxiv.org/abs/2506.01937v2
- http://arxiv.org/abs/2410.14872v2