# Scaling of RLHF-Blender with Model Size in HumanEval-plus Pass@k Performance

Assignee Research

June 12, 2026

**Abstract**

We apply preference modeling and reinforcement learning from human feedback (RLHF) to netune language models to act as helpful and harmless assistants. We nd this alignment training improves performance on almost all NLP evaluations, and is fully compatible with training for specialized skills such as python coding and summarization. We explore an iterated online mode of training, where preference models and RL policies are updated on a weekly cadence with fresh human feedback data, efciently improving our datasets and models. Finally, we investigate the robustness of RLHF training, and ide

## 1 Introduction

This paper examines: Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. Research question: How does the RLHF-Blender approach scale with increasing model size in terms of pass@k performance on HumanEval-plus compared to independent sampling in CodeT5+?.

## 2 Methodology

Systematic literature search across multiple databases yielded 18 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

## 3 Results

18 papers retrieved. 9 claims extracted; 7 independently verified. Quality review score: 7.5/10.

## 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 |
|---|---|---|
| Preference models trained to primarily evaluate helpfulness perform very poorly on harmlessness, and vice versa. | × | 0.13 |
| Preference models trained on a mixture of both helpful and harmless datasets can learn to be helpful when appropriate an | ✓ | 0.16 |
| Purely helpful RLHF-trained models are far easier to red-team compared to helpful+harmless models. | ✓ | 0.23 |
| Helpful+harmless models are both very helpful and much less harmful. | ✓ | 0.16 |
| RLHF-trained models tend to perform better than their raw, generative counterparts on virtually all evaluations. | ✓ | 0.25 |
| One can mix specialized skills with alignment-related training without compromising either alignment or performance. | ✓ | 0.22 |
| Professional Writers, Context Distilled, Static HH RLHF, Online HH RLHF (52B), and Online Helpful RLHF (52B) are compare | ✓ | 0.30 |
| Train PM (52B) and Test PM (52B) scores are compared for RLHF policies with varying numbers of training samples in Table | ✓ | 0.24 |
| RLHF Policy Performance On Test Prompts is evaluated for Helpful Comparisons, Harmless Comparisons, Helpful Prompts, and | × | 0.07 |

## References


- `https://arxiv.org/abs/2507.04340`
- `https://arxiv.org/abs/2204.05862`

- https://arxiv.org/abs/2410.05116