

Scaling Multimodal Datasets for PaLM-E Alignment with Human Problem-Solving Preferences

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: What is the impact of scaling the size of multimodal training datasets (e.g., combining MathQA, MATH, and GSM8K) on the alignment of PaLM-E with human problem-solving preferences, as measured by. 12 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Aligning Multimodal LLM with Human Preference: A Survey. Research question: What is the impact of scaling the size of multimodal training datasets (e.g., combining MathQA, MATH, and GSM8K) on the alignment of PaLM-E with human problem-solving preferences, as measured by benchmark accuracy and human evaluation scores?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

9 papers retrieved. 12 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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
MME-RealWorld, MMStar, MMBench, MMT-Bench, BLINK, MathVista, SQA3D, MMMU, MVBench, Mantis-Instruct are benchmarks for ev	×	0.03
Object HalBench, VideoHalluciner, VALOR-Eval, POPE, HaELM, OpenCHAIR, GAVIE, AMBER, Mementos, MMHal-Bench, VLind-Bench, M-	×	0.01
AdvDiffVLM, RTVLM, VLGuard, MultiTrust, VLLM-safety-bench, MOSSBench, MM-RLHF-SafetyBench are benchmarks for evaluating	×	0.03
Q-Bench, LLVisionQA, LLDescribe, LLaVA-Bench-Wilder, LiveBench, Vibe-Eval are benchmarks for evaluating conversation in	×	0.02
M-RewardBench, VL-RewardBench, RewardBench, MJ-Bench, MLLM-as-a-Judge, MM-RLHF-RewardBench are benchmarks for evaluating	×	0.02
ating		
Arena-Hard, AlpacaEval-V2, AlignBench, MM-AlignBench are benchmarks for evaluating alignment in multimodal models.	×	0.03
Fact-RLHF is the first multimodal RLHF algorithm, utilizing 10K human-labeled samples for the reward model and 50K hold-	×	0.03
DDPO assigns higher weights to corrected data in its loss function compared to standard DPO, using 1.4K manually refined	×	0.04
FDPO reuses InstructBLIP’s existing data for training.	×	0.04
The creation of alignment datasets involves three core factors: data sources, model responses, and preference annotation	✓	0.19
Most alignment algorithms are designed for specific tasks such as addressing hallucinations, ensuring safety, and improv	×	0.11
This survey is the first to specifically focus on the alignment of MLLMs.	×	0.09

References

- <http://arxiv.org/abs/2509.25160v1>
- <http://arxiv.org/abs/2503.14504v2>
- <http://arxiv.org/abs/2407.14477v4>