SOVEREIGN: How do alignment techniques such as RLHF and DPO affect the performance of LLMs on LLM-as-a-Judge benchmarks l

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Abstract The rapid evolution of large language models (LLMs) has driven a transformative shift in artificial intelligence (AI), reshaping both research paradigms and practical applications. Distinguished from their predecessors by unprecedented scale and advanced capabilities, LLMs necessitate new frameworks for understanding their development, behavior, and societal impact. This survey systematically reviews recent advancements in LLM techniques across four key dimensions: (1) pre-training methodologies, which establish core model capabilities through large-scale self-supervised training, arc

1 Introduction

Analysis of: A Survey of Large Language Models. Research goal: How do alignment techniques such as RLHF and DPO affect the performance of LLMs on LLM-as-a-Judge benchmarks like MT-Bench, and what are the trade-offs in terms of helpfulness vs. harmfulness scores?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

12 papers retrieved. 9 claims extracted, 9 verified. Tribunal: 9.2/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
The rapid evolution of large language models (LLMs) has driven a transformative shift in artificial intelligence (AI), r	✓	0.34
LLMs are distinguished from their predecessors by unprecedented scale and advanced capabilities.	✓	0.21
LLMs necessitate new frameworks for understanding their development, behavior, and societal impact.	✓	0.25
This survey systematically reviews recent advancements in LLM techniques across four key dimensions: (1) pre-training me	✓	0.33
Pre-training methodologies establish core model capabilities through large-scale self-supervised training, architectural	✓	0.37
Post-training techniques include supervised fine-tuning and reinforcement learning, which adapt foundational models to d	✓	0.31
Utilization strategies, such as in-context learning, prompt engineering, and agentic reasoning, optimize real-world depl	✓	0.37
Evaluation methods encompass benchmarks for key ability dimensions such as core language capabilities, reasoning, and sa	✓	0.33
The survey identifies critical research issues, including those concerning theoretical foundations, efficient scaling, a	✓	0.29

References

- https://doi.org/10.1007/s11704-026-60308-3
- https://doi.org/10.4230/oasics.icpec.2025.4
- https://doi.org/10.1007/s10676-025-09837-2