Qeios

Peer Review

Review of: "Large Vision-Language Model Alignment and Misalignment: A Survey Through the Lens of Explainability"

Ravi Varma Kumar Bevara Bevara[1]

1. University of North Texas, United States

This paper presents a comprehensive survey on the alignment and misalignment of Large Vision-Language Models (LVLMs) with a particular focus on explainability. The authors introduce key concepts related to alignment in LVLMs, categorize misalignment into object, attribute, and relational misalignment, and analyze its causes across the data, model, and inference levels. The paper also reviews mitigation strategies, categorizing them into parameter-tuning and parameter-frozen approaches, and suggests future research directions related to explainability, architectural improvements, and benchmarking.

- The paper offers a fresh perspective by evaluating LVLM alignment and misalignment through an explainability lens, which is not commonly explored in prior surveys.
- The distinction between representation-level and behavior-level alignment adds clarity to the field, and the taxonomy of misalignment (object, attribute, relational) is well-structured.
- The survey is well-referenced with recent papers (2023-2024), ensuring that it covers the latest developments in LVLM research, and the discussion on alignment training stages (contrastive learning, fine-tuning, end-to-end training) is well supported by citations.
- The survey effectively highlights challenges in LVLM misalignment, making it useful for new researchers, while the discussion of explainability techniques provides a new perspective on alignment issues.

⚠ **Areas for Improvement**:

doi.org/10.32388/O5Q7TY

- The section on why alignment is possible is insightful but remains theoretical without empirical validation.
- Some major studies in explainability for multimodal models are missing—a stronger discussion on attribution methods (e.g., SHAP, LIME for LVLMs) would enhance the literature review.
- There is no systematic comparison of mitigation strategies (e.g., effectiveness across different LVLM tasks), and some sections repeat information from prior work without adding new insights.
- The evaluation of mitigation strategies is high-level and lacks critical analysis.
- The policy implications of LVLM misalignment (e.g., ethical concerns, bias) are underexplored.
- Some technical explanations could be simplified for accessibility to a broader audience beyond AI researchers.

To strengthen the paper's claims about LVLM alignment, the authors could consider incorporating or referencing empirical studies that validate their theoretical discussions.

- Conduct an ablation study where different stages of alignment (contrastive learning, adapter fine-tuning, end-to-end fine-tuning) are tested independently to quantify their impact on alignment quality. This is inspired by research on modality fusion in multimodal transformers (e.g., CLIP, LLaVA) using representation similarity scores (cosine similarity between embeddings) and task performance (e.g., zero-shot image captioning accuracy).
- Analyze cases where visual evidence contradicts the LLM's textual priors (e.g., a green tomato classified as red), inspired by cross-modality knowledge conflict detection in LVLMs (e.g., CLIP vs. GPT-4V inconsistencies). Conduct a contrastive test where contradictory visual inputs are provided and analyze how model outputs deviate from factual correctness.

The paper does a good job listing various parameter-tuning and parameter-frozen alignment methods, but some sections are repetitive. Here's how the discussion can be enhanced and differentiated:

- Instead of just listing mitigation methods, analyze their trade-offs in terms of:
  - Computational cost (e.g., RLHF is expensive, contrastive learning is scalable).
  - Robustness across tasks (e.g., some methods improve image captioning but fail in visual reasoning).

- The paper doesn't discuss challenges in deploying these mitigation methods in real-world applications. Add a section on failure cases where mitigation strategies do not generalize well (e.g., why post-decoding corrections might fail in long-form VQA tasks).

Declarations

Potential competing interests: No potential competing interests to declare.