Annotation Bias in GPT-4 Visual Instructions and Hallucination Rates in Vision-Language Models

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the impact of annotation bias in GPT-4 generated visual instructions on the hallucination rates of vision-language models evaluated on standard VQA datasets. Despite vision-language models' (VLMs) remarkable capabilities as versatile visual assistants, two substantial challenges persist within the existing VLM frameworks: (1) lacking task diversity in pre-training and visual instruction tuning, and (2) annotation error and bias in. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Vision-Flan: Scaling Human-Labeled Tasks in Visual Instruction Tuning. Research question: What is the impact of annotation bias in GPT-4 generated visual instructions on the hallucination rates of vision-language models evaluated on standard VQA datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

12 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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
VISION-FLAN BASE achieves state-of-the-art performance on comprehensive evaluation benchmarks including MME, MM-Bench, a	×	0.09
VISION-FLAN BASE reduces hallucination and catastrophic forgetting.	×	0.12
VISION-FLAN BASE scores significantly lower on the LLaVA-Bench dataset compared to VLMs trained using GPT-4 synthesized	×	0.13
VISION-FLAN CHAT achieves significant performance improvement on LLaVA-Bench through the second-stage tuning on 1,000 GP	✓	0.19
VISION-FLAN BASE and VISION-FLAN CHAT retain more than 90% of performance on most tasks when using pretrained MLPs from	×	0.06
The Pearson Correlation Coefficient between the parameters of pretrained MLPs and instruction-tuned MLPs is computed.	×	0.05
VISION-FLAN dataset contains 1.6M instances and 196 tasks.	×	0.06
VISION-FLAN dataset is based on publicly available datasets.	×	0.09
MultiInstruct dataset contains 510K instances and 62 tasks.	×	0.02
MultiInstruct dataset mainly focuses on visual grounding tasks and only contains 29 tasks that do not involve region-spe	×	0.06

References

- http://arxiv.org/abs/2306.09265v1
- http://arxiv.org/abs/2604.12659v1

- http://arxiv.org/abs/2402.11690v1