SOVEREIGN: Does Dynamic Clue Bottlenecks improve interpretability faithfulness (e.g., via causal mediation analysis or at

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

The burgeoning interest in Multimodal Large Language Models (MLLMs), such as OpenAI's GPT-4V(ision), has significantly impacted both academic and industrial realms. These models enhance Large Language Models (LLMs) with advanced visual understanding capabilities, facilitating their application in a variety of multimodal tasks. Recently, Google introduced Gemini, a cutting-edge MLLM designed specifically for multimodal integration. Despite its advancements, preliminary benchmarks indicate that Gemini lags behind GPT models in commonsense reasoning tasks. However, this assessment, based on a lim

1 Introduction

Analysis of: Gemini in Reasoning: Unveiling Commonsense in Multimodal Large Language Models. Research goal: Does Dynamic Clue Bottlenecks improve interpretability faithfulness (e.g., via causal mediation analysis or attention attribution metrics) compared to post-hoc rationales on VCR reasoning splits using LIME and SHAP baselines?.

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. 15 claims extracted, 1 verified. Tribunal: 4.2/10 → REJECT (revision_round=0). Policy: ESCALATE_TO_OWNER.

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
Gemini lags behind GPT-4 Turbo in accuracy on language-only commonsense reasoning tasks.	✓	0.20
Gemini encounters challenges in temporal and social reasoning, as well as in emotion recognition in images.	×	0.05
GPT-4V outperforms Gemini Pro Vision across all subtasks in the VCR dataset.	×	0.07
Gemini Pro Vision's performance either matches or is slightly lower than GPT-4V's, except in temporal-type questions, wh	×	0.07
Commonsense reasoning involves understanding that a person carrying an umbrella on a cloudy day likely anticipates rain.	×	0.05
Commonsense reasoning includes understanding that a closed door in a library signifies a need for quiet.	×	0.07
Commonsense reasoning involves understanding that birds typically fly and fish live in water.	×	0.05
Contextual commonsense includes understanding that a person wearing a coat and shivering is likely cold.	×	0.03
Abductive commonsense includes inferring that wet streets are likely due to recent rain.	×	0.02
Event commonsense includes understanding that eating spoiled food can lead to feeling sick.	×	0.02
Temporal commonsense includes understanding that breakfast is typically eaten in the morning.	×	0.03
Numerical commonsense includes understanding that a cube has six faces.	×	0.03
Physical commonsense includes understanding that a glass will break if dropped on a hard floor.	×	0.02
Science commonsense includes understanding that water boils at a higher temperature at sea level than in the mountains.	×	0.04
Social commonsense includes recognizing that a person is likely upset if he/she is crying.	×	0.04

References

- `http://arxiv.org/abs/2312.17661v1`

- http://arxiv.org/abs/2603.09988v1
- http://arxiv.org/abs/2105.02657v2