SOVEREIGN: How does varying the number of active experts (k) in sparse MoE vision-language models affect VQA accuracy and

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

The rising popularity of explainable artificial intelligence (XAI) to understand high-performing black boxes raised the question of how to evaluate explanations of machine learning (ML) models. While interpretability and explainability are often presented as a subjectively validated binary property, we consider it a multi-faceted concept. We identify 12 conceptual properties, such as Compactness and Correctness, that should be evaluated for comprehensively assessing the quality of an explanation. Our so-called Co-12 properties serve as categorization scheme for systematically reviewing the eva

1 Introduction

Analysis of: From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. Research goal: How does varying the number of active experts (k) in sparse MoE vision-language models affect VQA accuracy and inference latency on benchmarks like VQAv2 or GQA, and does the optimal k correlate with visual complexity metrics such as object count or scene clutter?.

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

10 papers retrieved. 5 claims extracted, 5 verified. Tribunal: 8.3/10 → 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 rising popularity of explainable artificial intelligence (XAI) to understand high-performing black boxes raised the	✓	0.37
We identify 12 conceptual properties, such as Compactness and Correctness, that should be evaluated for comprehensively	✓	0.28
Our so-called Co-12 properties serve as categorization scheme for systematically reviewing the evaluation practices of m	✓	0.44
one in three papers evaluate exclusively with anecdotal evidence	✓	0.25
one in five papers evaluate with users	✓	0.17

References

- https://doi.org/10.1145/3583558
- https://doi.org/10.48550/arxiv.2206.04615
- https://doi.org/10.1109/access.2021.3140175