SOVEREIGN: How does GPT-4o's performance on ARC-Challenge compare to other state-of-the-art multimodal models like Flamin

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Getting a good feature representation of data is paramount for Human Activity Recognition (HAR) using wearable sensors. An increasing number of feature learning approaches-in particular deep-learning based-have been proposed to extract an effective feature representation by analyzing large amounts of data. However, getting an objective interpretation of their performances faces two problems: the lack of a baseline evaluation setup, which makes a strict comparison between them impossible, and the insufficiency of implementation details, which can hinder their use. In this paper, we attempt to a

1 Introduction

Analysis of: Comparison of Feature Learning Methods for Human Activity Recognition Using Wearable Sensors. Research goal: How does GPT-4o's performance on ARC-Challenge compare to other state-of-the-art multimodal models like Flamingo or Gemini in few-shot learning scenarios?.

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

14 papers retrieved. 7 claims extracted, 5 verified. Tribunal: 7.2/10 $\rightarrow$ REVISE (revision_round=1). Policy: SOFT_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 paper proposes an evaluation framework to allow rigorous comparison of features extracted by different methods.	✓	0.20
The paper provides all codes and implementation details to facilitate reproduction of results and re-use of the framewor	✓	0.18
Experiments were conducted on the OPPORTUNITY dataset.	×	0.04
Experiments were conducted on the UniMiB-SHAR dataset.	×	0.07
Hybrid deep-learning architectures involving convolutional and Long-Short-Term-Memory (LSTM) layers are effective for ob	✓	0.31
A lack of a baseline evaluation setup currently makes strict comparison between feature learning approaches impossible.	✓	0.27
Insufficiency of implementation details hinders the use of existing feature learning approaches.	✓	0.21

References

- https://doi.org/10.1109/access.2020.2983149
- https://doi.org/10.3390/s18020679
- https://doi.org/10.48550/arxiv.1505.00468