Synaptic Routing Impact on Codestral Adversarial Robustness Across Context Lengths

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 3 peer-reviewed papers addressing the following research question: How does MFOUR's Synaptic Routing affect Codestral's robustness to adversarial inputs in the AdvBench benchmark when scaling from 8K to 32K context lengths. 11 claims were extracted from source literature; 5 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records. Research question: How does MFOUR's Synaptic Routing affect Codestral's robustness to adversarial inputs in the AdvBench benchmark when scaling from 8K to 32K context lengths?.

2 Methodology

Systematic literature search across multiple databases yielded 3 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.7/10.

3 Results

3 papers retrieved. 11 claims extracted; 5 independently verified. Quality review score: 6.7/10.

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
MedAlign is a benchmark dataset containing 983 natural language instructions for Electronic Health Record (EHR) data.	✓	0.29
MedAlign was curated by 15 clinicians representing 7 medical specialties.	×	0.12
MedAlign includes clinician-written reference responses for 303 instructions.	✓	0.22
MedAlign provides 276 longitudinal EHRs for grounding instruction-response pairs.	✓	0.22
The study evaluated 6 general domain Large Language Models using the MedAlign dataset.	×	0.11
Clinicians ranked the accuracy and quality of each LLM response in the evaluation.	✓	0.16
GPT-4 exhibited an error rate of 35% in the MedAlign evaluation.	×	0.06
MPT-7B-Instruct exhibited an error rate of 68% in the MedAlign evaluation.	×	0.10
GPT-4 experienced an 8.3% drop in accuracy when context length was reduced from 32k to 2k tokens.	×	0.11
The study reports correlations between clinician rankings and automated natural language generation metrics.	✓	0.22
MedAlign is distributed under a research data use agreement.	×	0.13

References

- https://doi.org/10.48550/arxiv.2310.05869
- https://doi.org/10.48550/arxiv.2309.12307
- https://doi.org/10.1609/aaai.v38i20.30205