Robustness of Mistral 7B and Llama 3.1 8B to Distribution Shifts in Tabular Anomaly Detection

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 3 peer-reviewed papers addressing the following research question: How robust are Mistral 7B and Llama 3.1 8B to distribution shifts in tabular anomaly detection tasks when measured by the area under the precision-recall curve across different noise levels. 8 claims were extracted from source literature; 8 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Detecting hallucinations in large language models using semantic entropy. Research question: How robust are Mistral 7B and Llama 3.1 8B to distribution shifts in tabular anomaly detection tasks when measured by the area under the precision-recall curve across different noise levels?.

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 8.7/10.

3 Results

3 papers retrieved. 8 claims extracted; 8 independently verified. Quality review score: 8.7/10.

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
Large language model (LLM) systems, such as ChatGPT or Gemini, can show impressive reasoning and question-answering capa	✓	0.37
Answering unreliably or without the necessary information prevents adoption in diverse fields, with problems including f	✓	0.41
Encouraging truthfulness through supervision or reinforcement has been only partially successful.	✓	0.20
Researchers need a general method for detecting hallucinations in LLMs that works even with new and unseen questions to	✓	0.36
Entropy-based uncertainty estimators for LLMs can detect a subset of hallucinations—confabulations—which are arbitrary a	✓	0.31
The method addresses the fact that one idea can be expressed in many ways by computing uncertainty at the level of meani	✓	0.31
The method works across datasets and tasks without a priori knowledge of the task, requires no task-specific data, and r	✓	0.36
By detecting when a prompt is likely to produce a confabulation, the method helps users understand when they must take e	✓	0.36

References

- https://doi.org/10.1038/s41586-024-07421-0
- https://doi.org/10.48550/arxiv.2404.02151
- https://doi.org/10.1093/nar/gkae1082