Large Language Model Sensitivity to Data Perturbations in Radiology Report Classification

Vera Sorin, MD, Radiology Informatics Fellow, Radiology, Mayo Clinic, Rochester
Jeremy Collins, MD; Panagiotis Korfiatis, PhD

Introduction

Large language models (LLMs) are increasingly evaluated and applied to radiology reporting tasks. This study aimed to assess the impact of different types and levels of input text perturbations on LLM performance in classifying radiology reports.

Hypothesis

LLM performance may vary under different noise levels introduced into the radiology reports.

Methods

This was a retrospective IRB-approved study. We evaluated two Google LLMs, Gemini-1.5-Flash-001 and Gemini-1.5-Flash-002, on a balanced dataset of 2,200 CT pulmonary angiography reports (1,100 positive and 1,100 negative for pulmonary embolism). Three forms of noise were introduced: (1) Character noise: random removal of 20, 30, 60, or 120 characters, (2) Symbol noise: random insertion of 3, 9, 12, 24, or 64 symbols, and (3) Word shuffle: random rearrangement of 10, 30, or 50 words. Performance metrics for both models under each level of noise were calculated.

Results

Without noise, Gemini-1.5-Flash-001 achieved accuracy 0.967, recall 0.935, and F1-score 0.966. Gemini-1.5-Flash-002 performed at accuracy 0.984, recall 0.971, and F1-score 0.983. At the highest noise levels, Gemini-1.5-Flash-001's accuracy declined to 0.937 with character noise, 0.958 with symbol noise, and 0.932 with word shuffle. Gemini-1.5-Flash-002 had higher accuracy under the same conditions: 0.975 for character noise, 0.981 for symbol noise, and 0.975 for word shuffle. Overall, Gemini-1.5-Flash-002 demonstrated more resilience to noise, with smaller drops in accuracy across all perturbation types and levels.

Conclusion

Our results show that LLMs may be sensitive to data perturbations, including typos, formatting errors, and word shuffle. This vulnerability raises concerns about these models' performance when handling imperfect clinical data, as well as a potential sensitivity to cyber-attacks. Understanding the robustness and carefully validating LLMs is necessary prior to integrating into clinical practice.

Keywords

Artificial Intelligence/Machine Learning; Clinical Workflow & Productivity; Emerging Technologies