LLM Evaluation Report

Evaluation of a rule-based system supported by LLM judgment

Introduction

This report presents the evaluation of a rule-based system designed to detect logical inconsistencies in textual context. The system is tested using predefined test cases and its decisions are further reviewed by a Large Language Model acting as an independent judge. The goal of this evaluation is to identify weaknesses of deterministic rules and demonstrate how LLM-based judgment can support quality assurance in AI systems.

Evaluation Summary

Number of test cases: 7
Error rate: 0.14
Risky types: context_conflict
False positives: 0
False negatives: 1

Test ID	Result	Expected	Test type	LLM agrees	LLM comment
TC-01	FAIL	FAIL	context_conflict	N/A	N/A
TC-02	FAIL	FAIL	context_conflict	N/A	N/A
TC-03	PASS	PASS	context_conflict	N/A	N/A
TC-04	FAIL	FAIL	context_conflict	N/A	N/A
TC-05	FAIL	FAIL	context_conflict	N/A	N/A
TC-06	FAIL	FAIL	context_conflict	N/A	N/A
TC-07	PASS	FAIL	context_conflict	False	DISAGREE Reason: The context states that the individual has a dog and two cats, which totals three animals, contradicting the claim of having only two animals.