MIDI 2025 Conference

Comparative Evaluation of LLMs for Robotic Manipulation

Abstract

Large Language Models (LLMs) integrated with robotic systems enable natural language-driven manipulation. We evaluate five stateof-the-art LLMs (GPT-5.1, Claude Sonnet 4.5, Grok 4, Gemini 2.5 Flash, and Llama 4) for generating executable Python code through Arctos Studio. Using 10 diverse prompts with 3 semantic variants each (150 trials total), we assess code accuracy, execution success, semantic robustness, and error handling. Results show significant performance variations: GPT-5.1 achieved 60% success, Gemini 2.5 Flash 50%, Grok 4 and Llama 4 40%, and Claude Sonnet 4.5 16.7%. All models failed on Tasks 3 and 8, revealing universal limitations in state tracking and conditional reasoning. Syntax errors limited Llama 4 (23.3% rate), while commercial models achieved near-perfect syntax accuracy. Statistical analysis confirms significant differences (χ 2 = 11.94, $p = 0.018$). Llama 4 demonstrates competitive open-source performance with reproducibility advantages but increased vocabulary sensitivity.