THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

Master's Thesis Presentation

Yuetong (Cathy) Li
Department of Statistics
The University of Chicago

"NaCLO: A Benchmark for Evaluating Reasoning in Large Language Models via Low-Resource Linguistic Puzzles"

November 10, 2025, at 4:00 PM
Jones 111, 5747 S. Ellis Avenue

Abstract

This study introduces **NaCLO**, a new benchmark for evaluating the reasoning capabilities of large language models (LLMs) through linguistic puzzles from the North American Computational Linguistics Olympiad (NACLO). These puzzles, drawn from low-resource and typologically diverse languages, require solvers to infer grammatical patterns rather than rely on memorization, providing a controlled setting for testing structured reasoning.
NaCLO contains 46 curated problems across 20 language families and evaluates several frontier models—GPT-4o, Claude, Gemini, LLaMA, and DeepSeek—under zero-shot, one-shot, and no-context conditions. Results show that models perform best on **fill-in-blank** tasks, followed by **match-up**, with **inference** questions proving more difficult and **translation** tasks the most challenging overall. These patterns indicate that while models handle surface-level pattern recognition well, they struggle with deeper, multi-step reasoning and generalization across unseen linguistic structures.

The findings highlight persistent limitations in current LLMs' reasoning ability and underscore the importance of linguistically grounded, low-resource benchmarks for evaluating genuine language understanding.
