摘要

本研究針對大模型在漢語否定理解能力評估方面的空白，構建了首個涵蓋顯性、隱性、多否定詞及特殊否定句對的漢語否定表達評測基準。該基準依託於北京大學現代漢語語料庫，以半自動化方式構建，並在ChatGPT上進行系統測試。結果顯示，提示詞優化可顯著提升模型在否定理解任務上的表現，但模型在處理句子級否定、多個否定詞疊加、否定詞移位及否定溢出等複雜情況時仍存在挑戰。本研究為全面評估大模型在漢語否定的語義推理方面的實際能力提供了依據，並提出了改進方向。

關鍵詞：大模型；評測基準；漢語否定理解；語義推理

Abstract

This study addresses the gap in assessing the ability of large models to understand Chinese negation by constructing the first benchmark for evaluating Chinese negation expressions, covering explicit, implicit, multiple negators, and special cases of negated pairs. The benchmark is built in a semi-automated manner based on the Corpus of Centre for Chinese Linguistic Peking University (CCL) and systematically tested on ChatGPT. The results show that prompt optimization can significantly improve the model's performance on negation understanding tasks, but challenges remain for the model in handling sentence-level negation, multiple stacked negators, negators displacement, and negation semantic overflow. This study provides a basis for comprehensively assessing the actual capability of large language models in semantic inference with negation expression in Chinese and suggests directions for improvement.

Keywords: Large Language Models, Chinese Negation Understanding, Evaluation Benchmark, Semantic Inference