澳門大學

協同創新研究院

協同

人文學院

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

語言大模型漢語否定推理能力評測研究

Benchmarking the Negative Inference Ability in Chinese for Large Language Models

周金晶, 學生編號: MC365000

理學碩士學位（數據科學—計算語言學）課程
項目報告

項目導師

袁毓林教授

2025 年 5 月 12 日

摘要

作爲自然語言推理的重要組成部分，否定推理長期以來被學術界納入各類推理評測集的構建範疇。然而，現有的研究多聚焦於英文語境下的否定推理的評測，關於否定推理的中文評測集還非常地少。爲了探索大模型在中文否定推理上的表現，我們以語言學家在否定表達領域的研究爲理論依據，構建了包含單否定詞格式、多否定詞格式和反問句的中文否定推理評測集 NITS。我們使用該測試集分別對以中文爲母語的人類和三種大模型 Qwen-Max-0125、DeepSeek-R1 和 GPT-4o 進行了測試。測試結果表明：大模型已經具備一定的否定推理能力，并在否定推理測試上的表現高於人類基綫水平；中文大模型在否定推理測試上的表現，優於英文大模型；通過改變提問方式，如少樣本學習，可以改善大模型的性能；模型在否定轄域變化、否定提升、否定冗餘、複雜隱性結構等特殊否定類型上，仍存在不同程度的推理困難。

關鍵詞： 自然語言推理，否定推理，大模型，NITS

Abstract

As an important part of the field of Natural Language Inference, Negative Inference has long been included in the construction of various types of inference test sets by academics. However, most of the existing research focuses on Negative Inference assessment in the English context and there are still very few Chinese test sets on Negative Inference. To explore the performance of Large Language Models on Negative Inference, we constructed a Chinese Negative Inference Test Set NITS containing single negation formats, multiple negation formats, and rhetorical questions based on linguists' research in negative expressions. We used this test set to evaluate the performance of Chinese native speakers and three Large Language Models, Qwen-Max-0125, DeepSeek-R1, and GPT-4o respectively. The test results show that: Large Language Models already have some negative inference ability and perform better than the human baseline; the Chinese Large Language Models outperform the English Large Language Models on the negative inference test; the performance of the Large Language Models can be improved by changing the questioning style, such as learning with fewer samples; the models still have different degrees of reasoning difficulties on special negation types such as negation scope change, negation lifting, negation redundancy, and complex implicit structures.

Keywords: Natural Language Inference, Negative Inference, Large Language Models, NITS