Don’t Think About Pink Elephants

Future of Negation in LLMs

- the NOT problem
- the power of negation:
  - in human language: it can subvert the whole meaning of the sentence / paragraph / text
  - in LLM: often does not have a different role to any other word

Context	Match	Mismatch
A robin is a ___	*bird*	*tree*
A robin is not a ___	*bird*	*tree*

Figure: What is a robin? [1]

- LLM training datasets: mostly positive examples, impact on interpreting negation in text
- areas of interest
  - the impact of including more negative examples in the training data on the ability to interpret the negation
  - implementation of more slow thinking methodology in the reasoning of the LLM
  - addition of a special role to the negation operators

[1] Allyson Ettinger. “What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models”. In: *Transactions of the Association for Computational Linguistics* 8 (Jan. 2020), pp. 34–48. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00298. URL: https://doi.org/10.1162/tacl_a_00298.