<https://news.mit.edu/2023/large-language-models-are-biased-can-logic-help-save-them-0303>

While we may still be far away from a neutral language model utopia, this research is ongoing in that pursuit. Currently, the model is just for language understanding, so it’s based on reasoning among existing sentences. Unfortunately, it can’t generate sentences for now, so the next step for the researchers would be targeting the uber-popular generative models built with logical learning to ensure more fairness with computational efficiency.

<https://arxiv.org/abs/2305.11098>

<https://arxiv.org/abs/2305.14695>

**A Causal View of Entity Bias in (Large) Language Models**

IEEE, Springer, Elsevier, Wiley, Taylor and Francis

<https://arxiv.org/pdf/2302.04023.pdf>

# <https://arxiv.org/abs/2303.16634> --G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment

<https://www.amazon.science/blog/using-large-language-models-llms-to-synthesize-training-data>

<https://dl.acm.org/doi/10.1145/3544548.3580688>

<https://arxiv.org/abs/2302.02463>

Nationality Bias in Text Generation

<https://arxiv.org/abs/2305.10510>

ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages

<https://arxiv.org/pdf/2304.01358.pdf>

Title: Challenging the appearance of machine intelligence: Cognitive bias in LLMs

<https://ai.stanford.edu/blog/linkbert/> - LinkBERT

from transformers import AutoTokenizer, AutoModel

tokenizer = AutoTokenizer.from\_pretrained('michiyasunaga/LinkBERT-large')

model = AutoModel.from\_pretrained('michiyasunaga/LinkBERT-large')

inputs = tokenizer("Hello, my dog is cute", return\_tensors="pt")

outputs = model(\*\*inputs)

from transformers import AutoTokenizer, AutoModel

tokenizer = AutoTokenizer.from\_pretrained('michiyasunaga/BioLinkBERT-large')

model = AutoModel.from\_pretrained('michiyasunaga/BioLinkBERT-large')

inputs = tokenizer("Sunitinib is a tyrosine kinase inhibitor", return\_tensors="pt")

outputs = model(\*\*inputs)

<https://www.scribbr.com/research-bias/baader-meinhof-phenomenon/> - The Baader–Meinhof Phenomenon Explained

<https://aclanthology.org/2022.bigscience-1.3.pdf>

<https://www.amazon.science/publications/linguist-language-model-instruction-tuning-to-generate-annotated-utterances-for-intent-classification-and-slot-tagging>

<https://paperswithcode.com/paper/language-models-don-t-always-say-what-they>

<https://paperswithcode.com/paper/language-models-don-t-always-say-what-they>

Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting

<https://github.com/milesaturpin/cot-unfaithfulness>

<https://arxiv.org/abs/2305.03742>

<https://arxiv.org/abs/2212.09597>

Reasoning with Language Model Prompting: A Survey

<https://paperswithcode.com/paper/towards-reasoning-in-large-language-models-a>