<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>

<https://arxiv.org/abs/1906.02361>

#### [Explain Yourself! Leveraging Language Models for Commonsense Reasoning](https://arxiv.org/abs/1906.02361) 6 Jun 2019

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, Richard Socher

#### [Leap-Of-Thought: Teaching Pre-Trained Models to Systematically Reason Over Implicit Knowledge](https://arxiv.org/abs/2006.06609) 11 Jun 2020

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, Jonathan Berant

#### [Measuring Mathematical Problem Solving With the MATH Dataset](https://arxiv.org/abs/2103.03874) 5 Mar 2021

### Prompting and In-Context Learning

#### Chain of Thought Prompting and Its Variants/Applications

#### [Chain of Thought Prompting Elicits Reasoning in Large Language Models](https://arxiv.org/abs/2201.11903) 28 Jan 2022

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou

#### [Iteratively Prompt Pre-trained Language Models for Chain of Thought](https://arxiv.org/abs/2203.08383) 16 Mar 2022

### Prompting and In-Context Learning

#### Chain of Thought Prompting and Its Variants/Applications

#### [Chain of Thought Prompting Elicits Reasoning in Large Language Models](https://arxiv.org/abs/2201.11903) 28 Jan 2022

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou

#### [Iteratively Prompt Pre-trained Language Models for Chain of Thought](https://arxiv.org/abs/2203.08383) 16 Mar 2022

Boshi Wang, Xiang Deng, Huan Sun

#### [Large Language Models are Zero-Shot Reasoners](https://arxiv.org/abs/2205.11916) 24 May 2022

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, Yusuke Iwasawa

#### [Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models](https://arxiv.org/abs/2209.08141) 16 Sep 2022

Ben Prystawski, Paul Thibodeau, Noah Goodman

#### [Language Models are Multilingual Chain-of-Thought Reasoners](https://arxiv.org/abs/2210.03057) 6 Oct 2022

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, Jason Wei

#### [Large Language Models are few(1)-shot Table Reasoners](https://arxiv.org/abs/2210.06710) 13 Oct 2022

Wenhu Chen

#### [Language Models of Code are Few-Shot Commonsense Learners](https://arxiv.org/abs/2210.07128) 13 Oct 2022

Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, Graham Neubig

#### [PaL: Program-Aided Language Model](https://arxiv.org/abs/2211.10435) 18 Nov 2022

Luyu Gao\*, Aman Madaan\*, Shuyan Zhou\*, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, Graham Neubig

#### [Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks](https://arxiv.org/abs/2211.12588) 22 Nov 2022

#### [On the Advance of Making Language Models Better Reasoners](https://arxiv.org/abs/2206.02336) 6 Jun 2022

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, Weizhu Chen

#### [Complexity-Based Prompting for Multi-Step Reasoning](https://arxiv.org/abs/2210.00720) 3 Oct 2022

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, Tushar Khot

#### [Automatic Chain of Thought Prompting in Large Language Models](https://arxiv.org/abs/2210.03493) 7 Oct 2022

Zhuosheng Zhang, Aston Zhang, Mu Li, Alex Smola

#### [Teaching Algorithmic Reasoning via In-context Learning](https://arxiv.org/abs/2211.09066) 15 Nov 2022

<https://www.technologyreview.com/2023/03/20/1070067/language-models-may-be-able-to-self-correct-biases-if-you-ask-them-to/>

<https://arxiv.org/pdf/2302.07459.pdf>

Prompt Engineering Our Q+IF, Q+IF+CoT, and Q+IF+Match Stats experiments all rely on prompts engineered to be appropriate for each experiment. Small variations in the prompts can sometimes yield large changes in model outputs. We have not systematically tested for this in any of our experiments. Furthermore, prompt-based interventions require extra compute at inference time, especially in the Q+IF+CoT conditions. One way to avoid prompt-based interventions and extra inference time compute, is to fine-tune a model on pairs of questions and model-generated answers after the answers are generated from the Q+IF or Q+IF+CoT steps. Along these lines, a recent technique called Constitutional AI, trains language models to adhere to a humanwritten set of ethical principles (a constitution) by first having models determine whether their outputs violate these principles, then training models to avoid such violations [4]. Constitutional AI and our work observe the same phenomenon: sufficiently large language models, with a modest amount of RLHF training to be helpful, can learn how to abide by high-level ethical principles expressed in natural language.

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

A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity

<https://arxiv.org/abs/2304.03612>

What does ChatGPT return about human values? Exploring value bias in ChatGPT using a descriptive value theory

<https://arxiv.org/abs/2305.10235>

Assessing Hidden Risks of LLMs: An Empirical Study on Robustness, Consistency, and Credibility

<https://huggingface.co/gouravsinha/finance-NER>

<https://huggingface.co/blog/rlhf>

https://arxiv.org/abs/2212.08073

What is Constitutional AI? And why have companies like Anthropic adopted it to train helpful, harmless and honest models?

Constitutional AI (CAI) is similar to RLHF except instead of human feedback, it learns through AI feedback.

At a high-level, there are two stages of Constitutional AI (CAI): the Reflection stage and the Reinforcement stage.

Ask the LLM to generate toxic responses.

2️⃣ Give the LLM a set of rules to follow (or a Constitution). Present the toxic responses back to the LLM and ask if they accord with the Constitution.

3️⃣ Ask the LLM to generate a revised response. (Repeat revision until the responses follow the Constitution.)

4️⃣ This creates a synthetic dataset, which you can use for training. Fine-tune (or train) the baseline model on this synthetic dataset to create responses that more closely follow the Constitution.

5️⃣ Through this process, you get the SL-CAI model, the intermediate model.