Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era

Lecture-Style Tutorial @ KDD 2024

Sunhao Dai[1], Chen Xu[1], Shicheng Xu[2], Liang Pang[2], Zhenhua Dong[3], Jun Xu[1]

1 Gaoling School of Artificial Intelligence, Renmin University of China

2 Institute of Computing Technology, Chinese Academy of Sciences

3 Huawei Noah's Ark Lab

https://llm-ir-bias-fairness.github.io/

Origanizers

Sunhao Dai

Gaoling School of Artificial Intelligence,
Renmin University of China
sunhaodai@ruc.edu.cn

Chen Xu

Gaoling School of Artificial Intelligence,
Renmin University of China
xc_chen@ruc.edu.cn

Shicheng Xu

Institute of Computing Technology,
Chinese Academy of Sciences
xushicheng21s@ict.ac.cn

Liang Pang

Institute of Computing Technology,
Chinese Academy of Sciences
pangliang@ict.ac.cn

Jun Xu

Gaoling School of Artificial Intelligence,
Renmin University of China
junxu@ruc.edu.cn

Zhenhua Dong

Noah’s Ark Lab,
Huawei Technologies Co.,Ltd
dongzhenhua@huawei.com

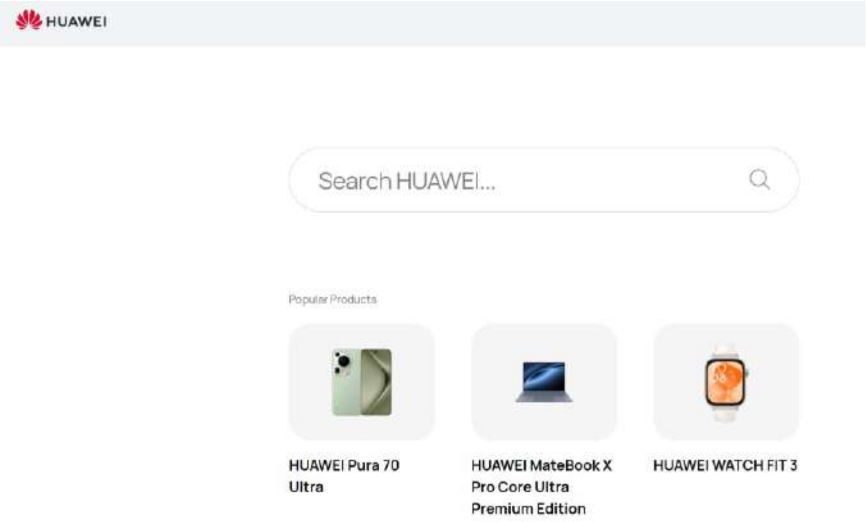
Schedule

- Part 1 (30 mins, 10:00 - 10:30)
  - Introduction (Jun Xu, 15 mins)
  - A Unified View of Bias and Unfairness (Jun Xu, 15 mins)
- Coffee Break (15 mins, 10:30 - 10:45)
- Part 2 (135 mins, 10:45 - 13:00)
  - Bias and Mitigation Strategies (Sunhao Dai, 75 mins)
  - Unfairness and Mitigation Strategies (Liang Pang, 45 mins)
  - Conclusion and Future Directions (Liang Pang, 10 mins)
  - Q&A (5 mins)

Outline

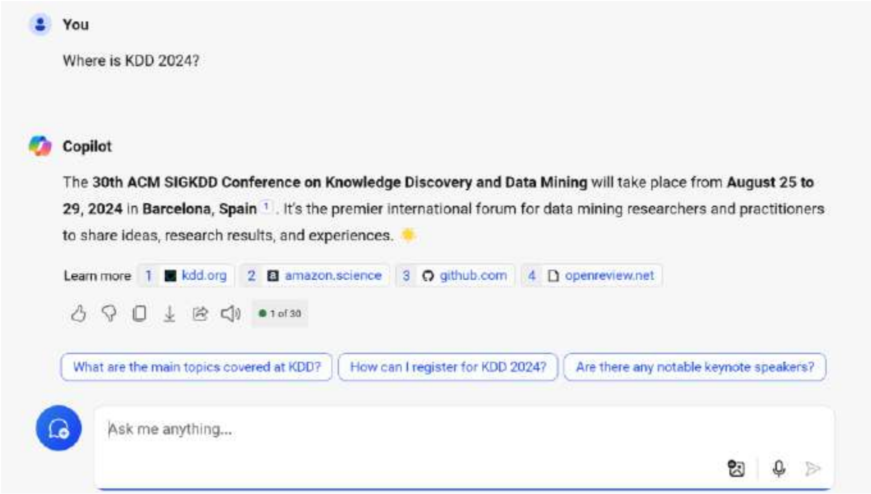
- Introduction
- A Unified View of Bias and Unfairness
- Bias and Mitigation Strategies
- Unfairness and Mitigation Strategies
- Conclusion and Future Directions

Information Retrieval Systems



- Product Search

- Music
- Video
- Apps



Information Retrieval is Everywhere

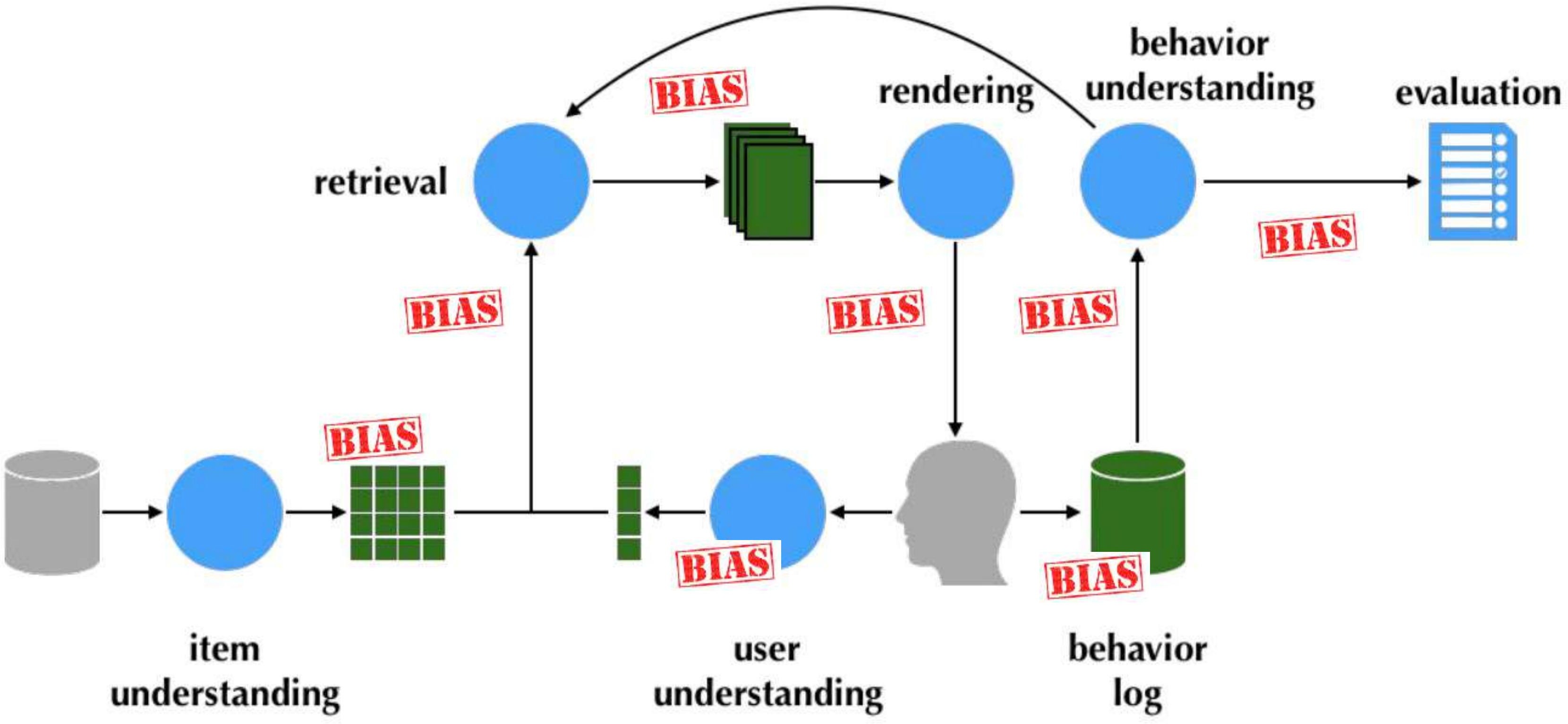
- New Bing

Biases in Information Retrieval

A disproportionate weight *in favor of* or *against* an idea or thing ……

In science and engineering, a bias is a systematic error

——*Wikipedia*



[1] Michael D. Ekstrand et al., *Fairness and Discrimination in Information Access Systems, Foundations and Trends in Information Retrieval 2022.*

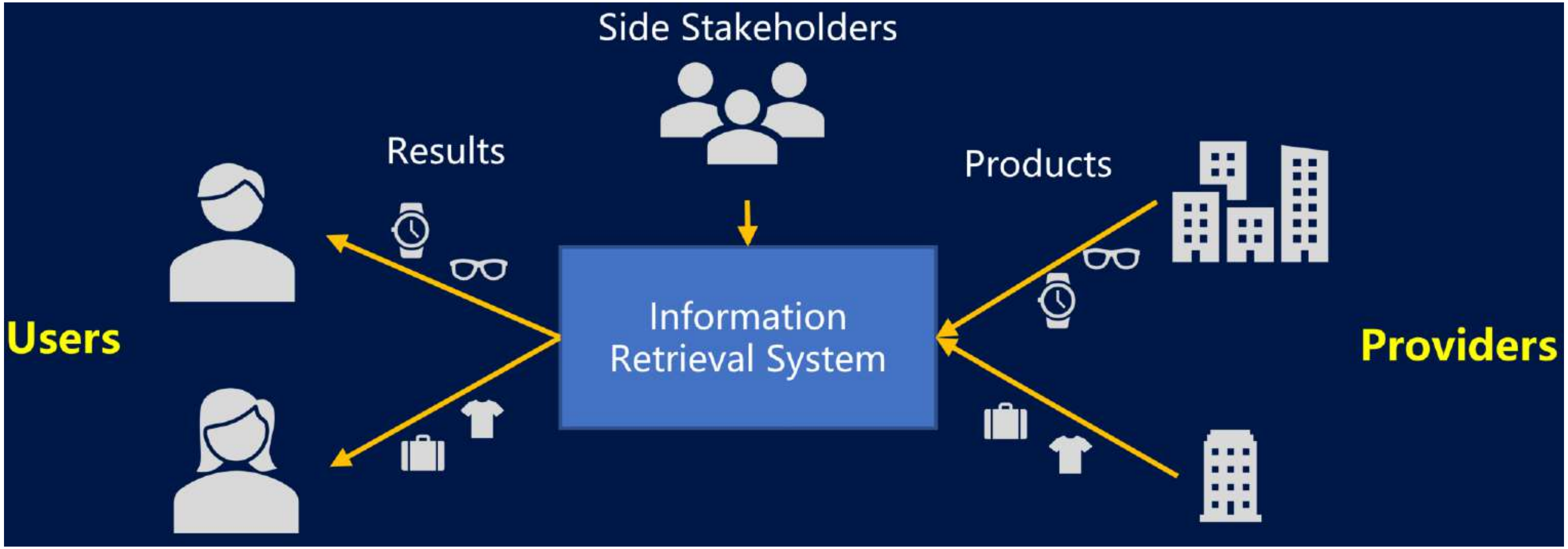
Unfairness in Information Retrieval

- User-fair: Equality

Everyone is treated the same and provided same resources to succeed

- Item-fair: Equity

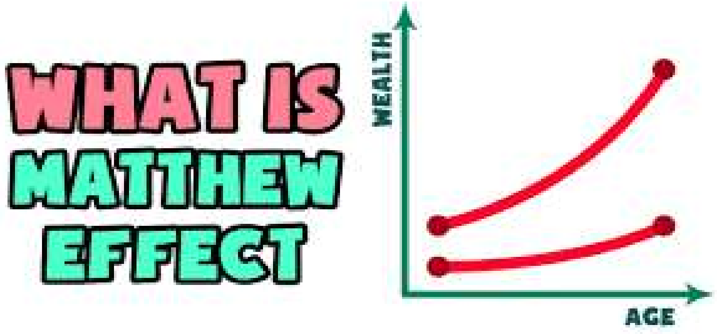
Ensuring that resources (e.g., exposures) are equally distributed based on needs



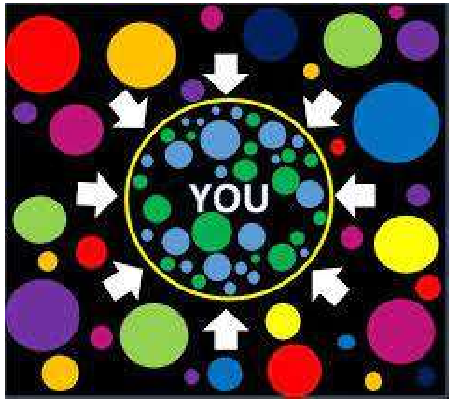
Consequence

Hurting Information Retrieval System Performance

Hurting Sustainability and Long-term Development



Matthew Effect

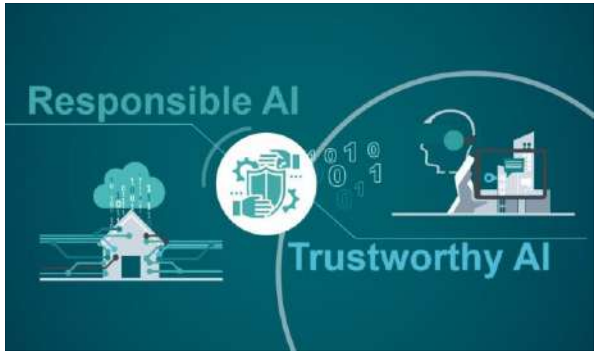


Echo Chambers

Monopoly

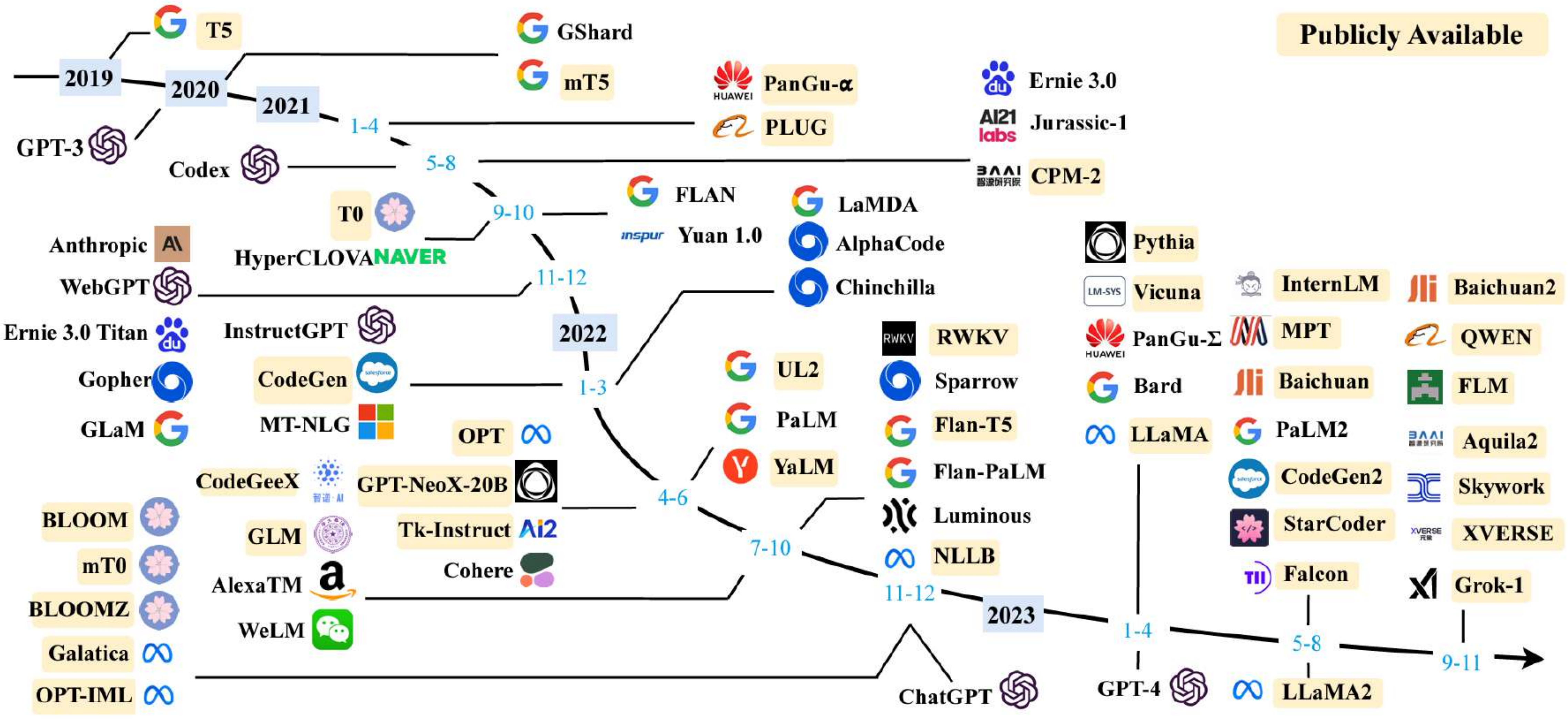
Responsible IR

- Improve user/provider experience
- Legal and policy harmonization
- Sustainable and long-term development



Artificial Intelligence with Warmth

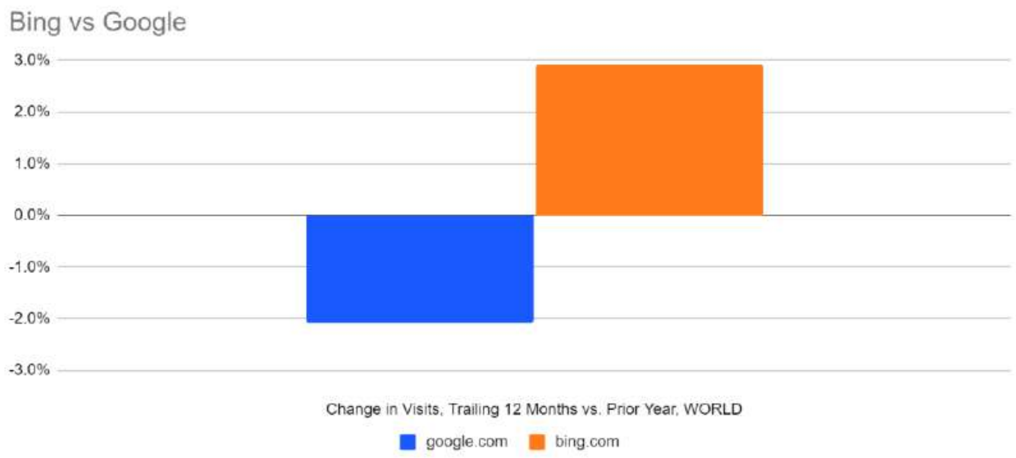
Large Language Models



[1] *Wayne Xin Zhao et al. A Survey of Large Language Models. arXiv 2023.*

LLMs Meet IR

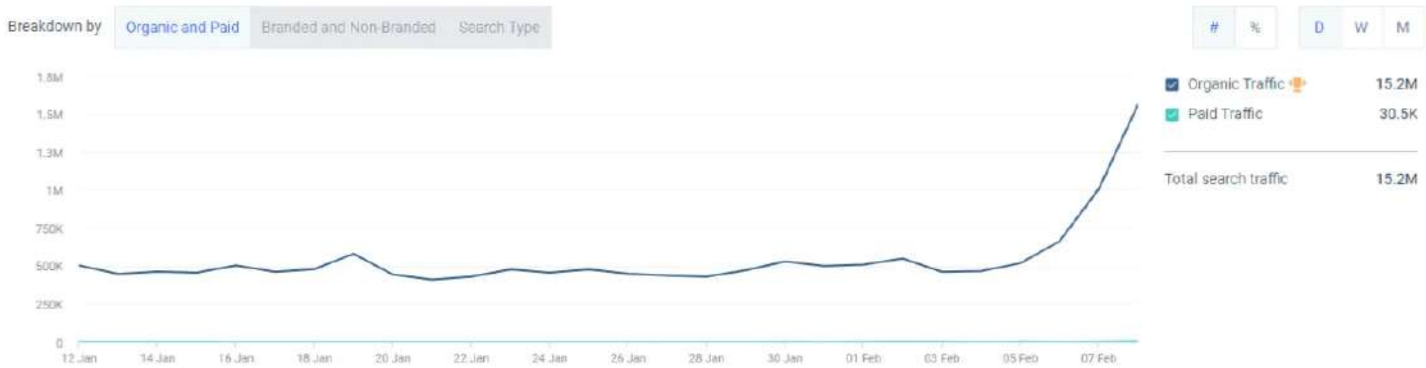
SIGIR 2024



Traffic Volume

Bing ↑

Google ↓

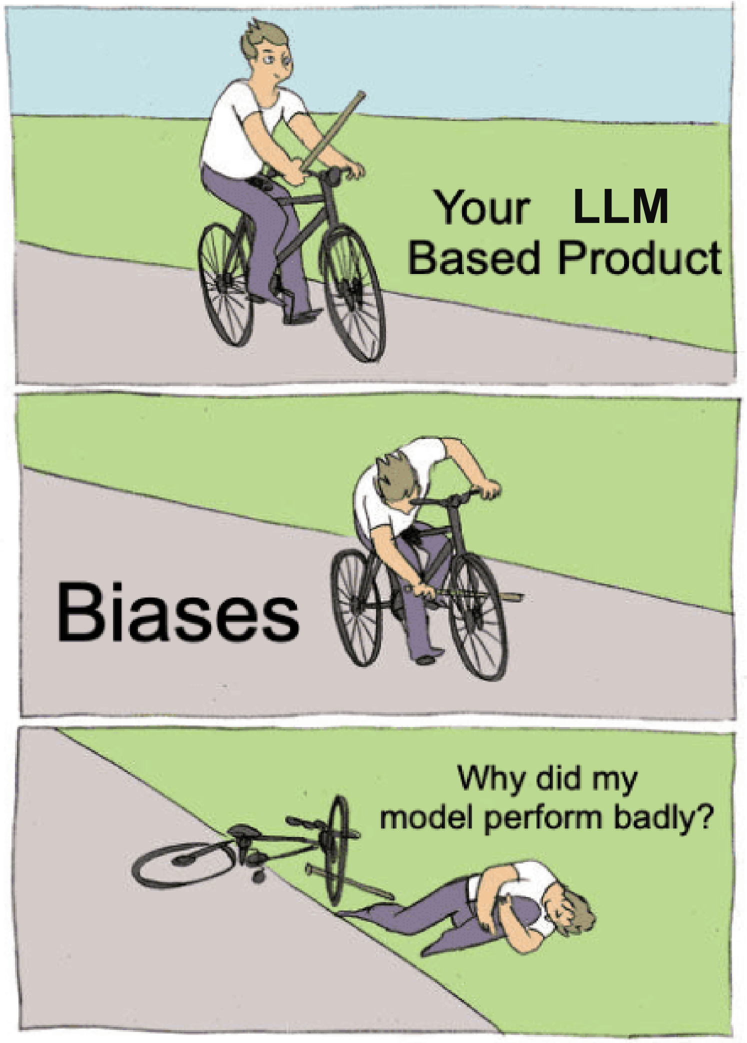


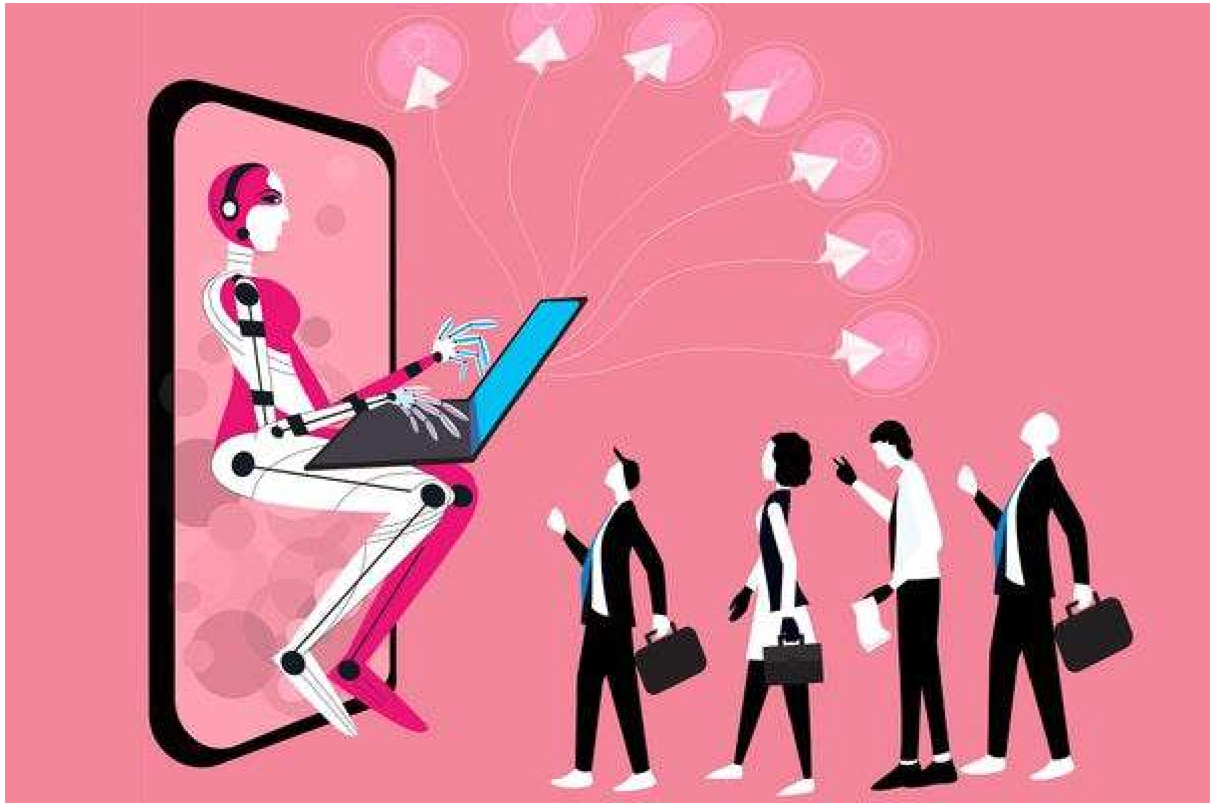
Search volume for “bing ai” 700%↑

[1] https://www.youtube.com/watch?v=SE9W2M8BPWk

[2] https://www.similarweb.com/blog/insights/ai-news/bing-chatgpt-ai-chat/

Concerns





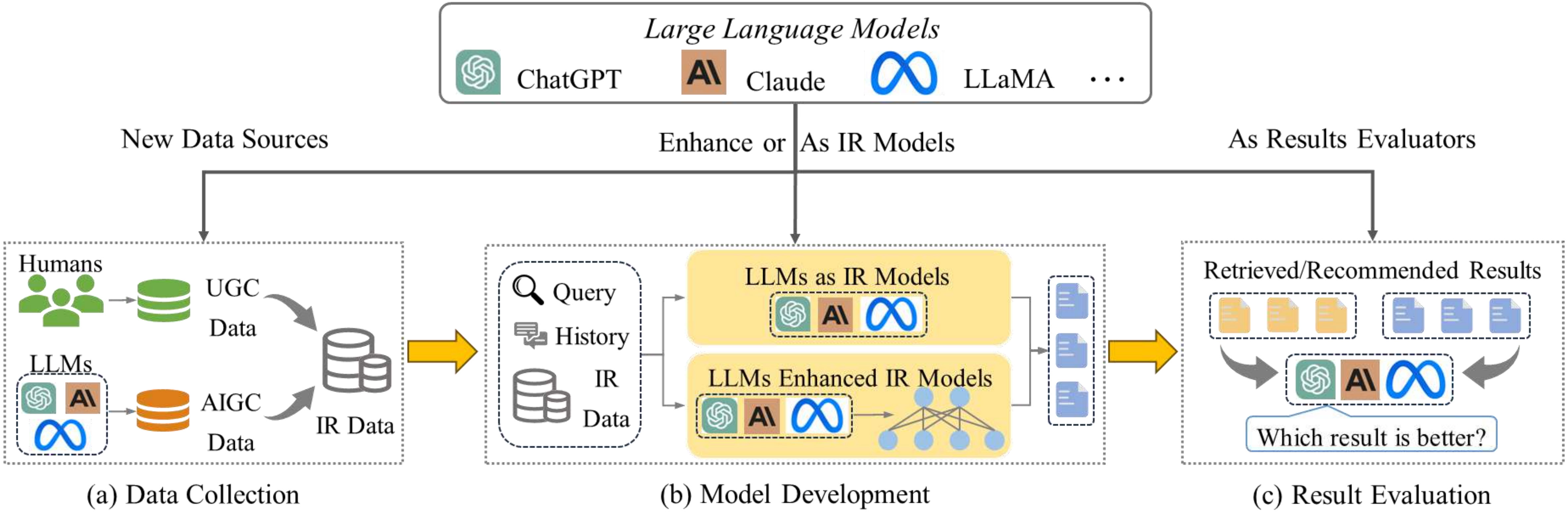
LLMs show an inherent discrimination against gender

[1] https://blog.nimblebox.ai/dealing-with-biases-and-fairness-in-llms
[2] https://www.scientificamerican.com/article/chatgpt-replicates-gender-bias-in-recommendation-letters/

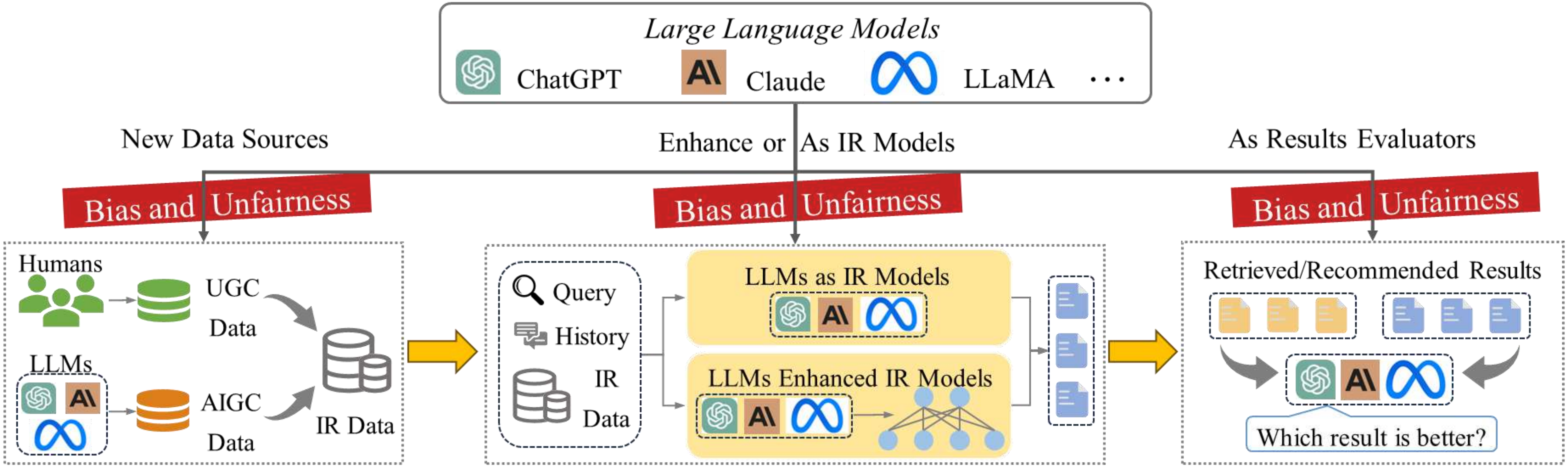
Outline

- Introduction
- A Unified View of Bias and Unfairness
- Bias and Mitigation Strategies
- Unfairness and Mitigation Strategies
- Conclusion and Future Directions

Integration of LLMs into IR Systems



Integration of LLMs into IR Systems



(a) Data Collection

(b) Model Development

(c) Result Evaluation

- **Source Bias**
- **Factuality Bias**

- **User Unfairness**
- **Item Unfairness**

- **Position Bias**
- **Popularity Bias**
- **Instruction-Hallucination Bias**
- **Context-Hallucination Bias**

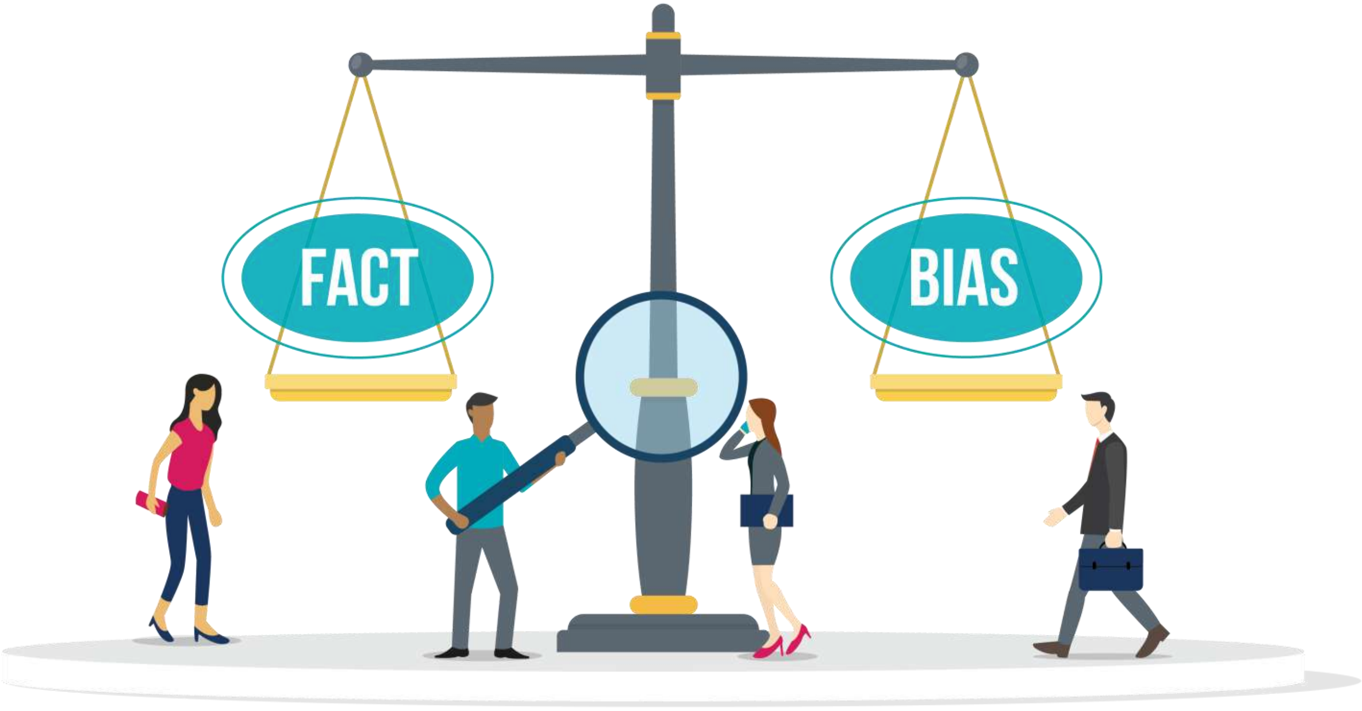
- **User Unfairness**
- **Item Unfairness**

- **Selection Bias**
- **Style Bias**
- **Egocentric Bias**

- **User Unfairness**
- **Item Unfairness**

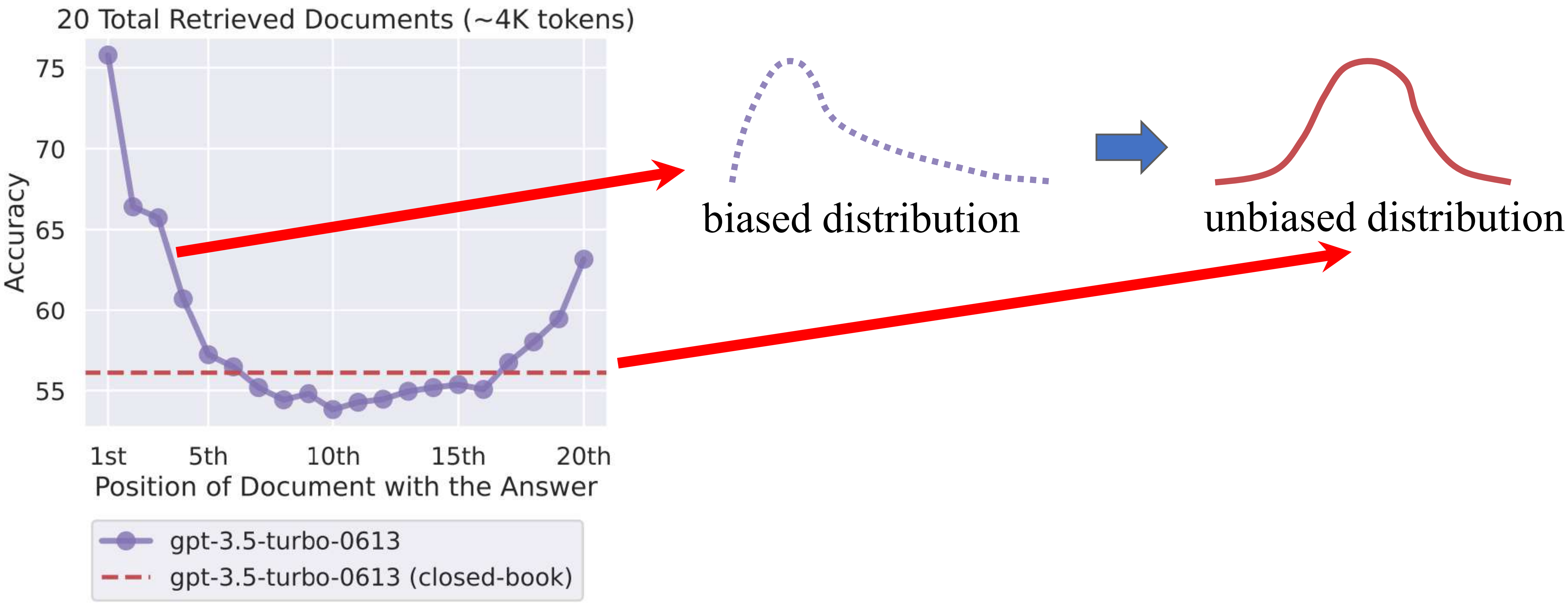
Bias Definition

- **The Cambridge Dictionary**
  - Fact of a collection of data containing more information that supports a particular opinion than you would expect to find if the collection had been made by chance



Examples

- **Position Bias：LLMs are sensitive to postions changes**



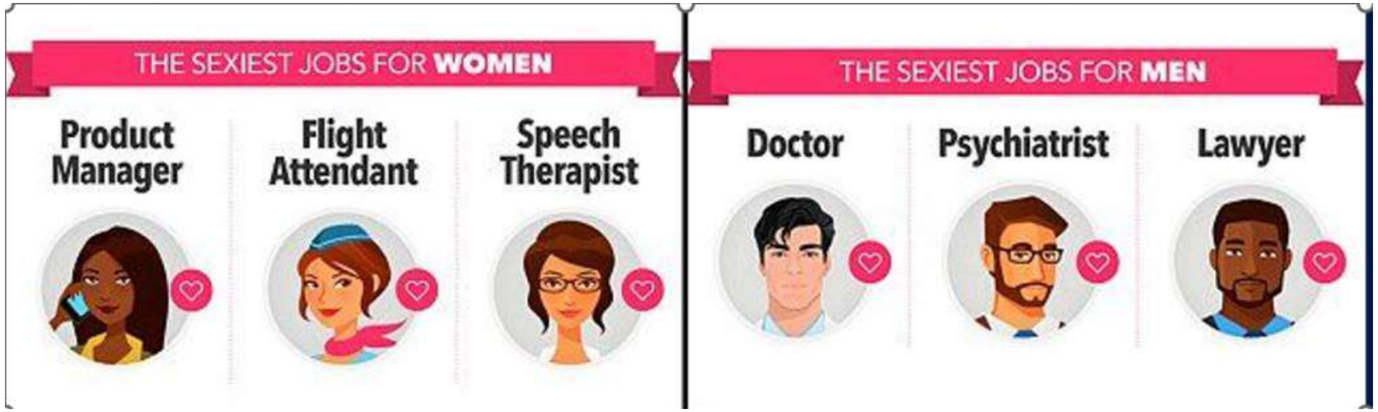
[1] Nelson F. Liu et al. Lost in the Middle: How Language Models Use Long Contexts. TACL 2024.

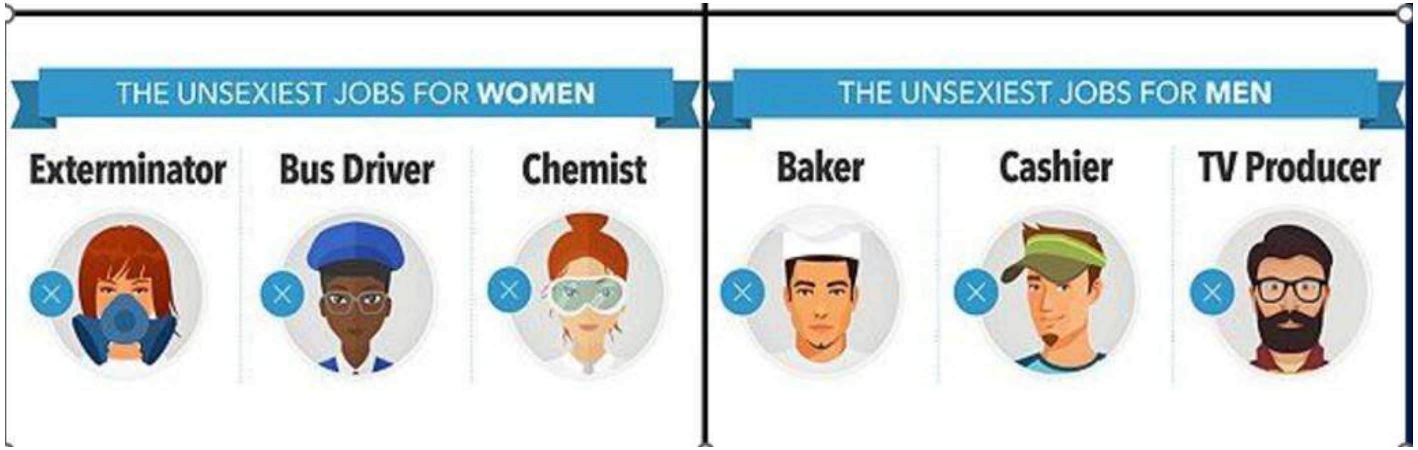
Fairness Definition

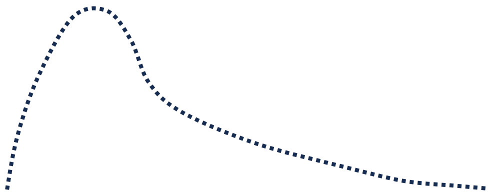
- **The Cambridge Dictionary**
  - Action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment

Examples

- **User fairness: we need to balance genders in job seeking**

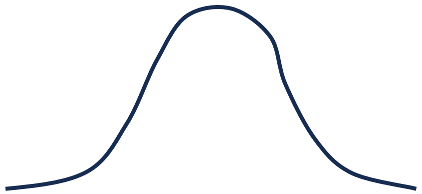






unfair distribution

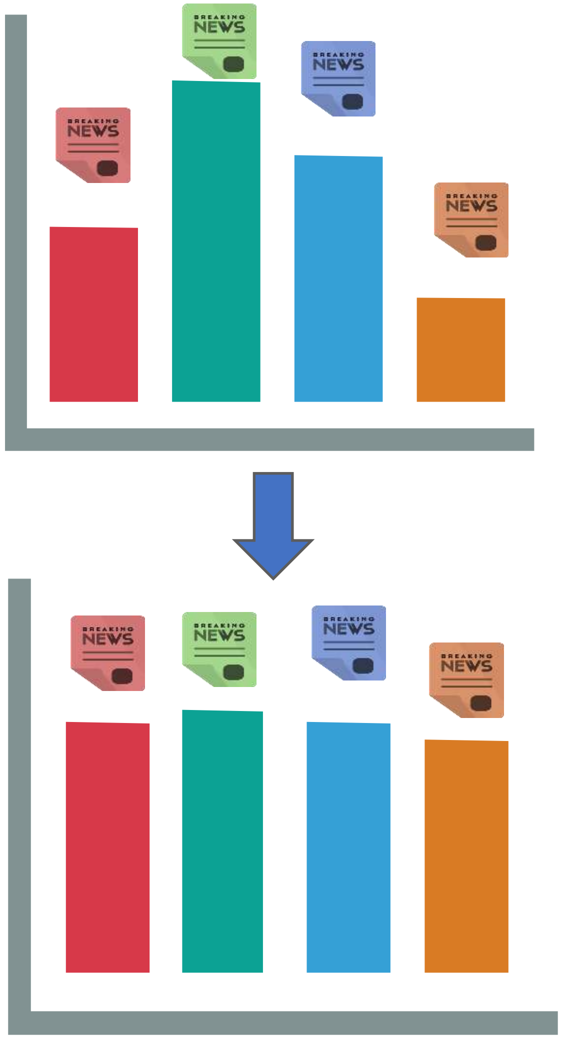
uniform distribution

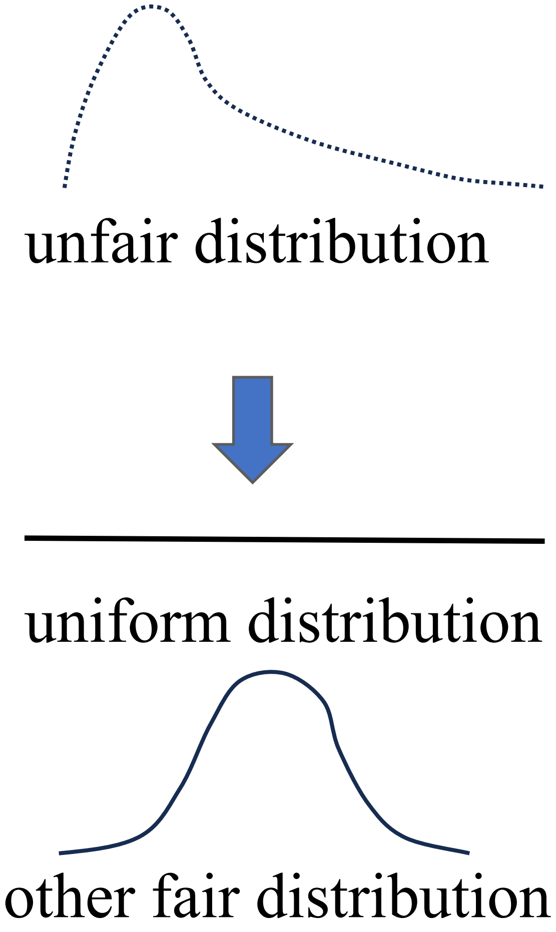


other fair distribution

Examples

- **Item fairness: we need to balance item exposures**





Question

Can we utilize a unified view to treat bias and unfairness?

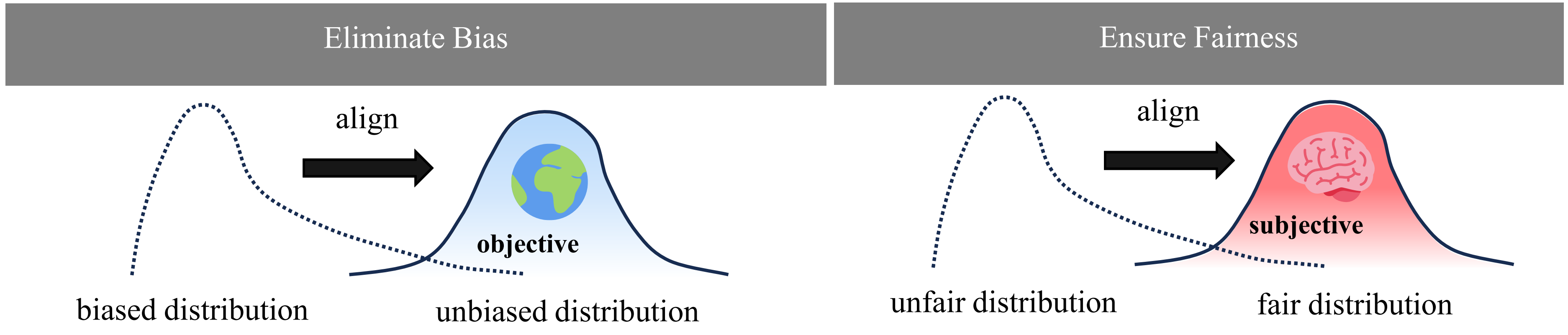
A Unified View

- They can be both viewed as a ***Distribution Alignment*** problem
  - ➢ Bias: Fact of a collection of data containing more information that supports a particular opinion

    Eliminate Bias: aligns with an objective distribution (real worlds)
  - ➢ Unfairness: Action of supporting or opposing a particular person orthing

    Ensure Fairness: aligns with a subjective distribution (human values)

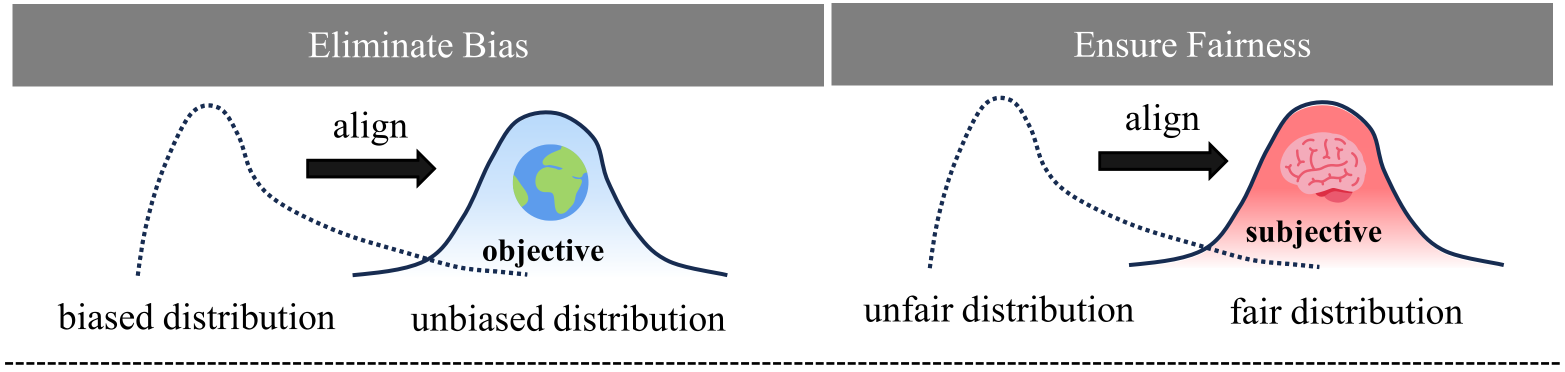
Unified View from Distribution Alignment Perspective



A Unified View

- **Formulation:** $P(\widehat{R}) \neq P(R)$
- $P(\widehat{R})$ **is the predicted distribution**
- $P(R)$ **is the target distribution**
  - **Unbias: objective distribution**
  - **Fairness: subjective distribution**

Unified View from Distribution Alignment Perspective

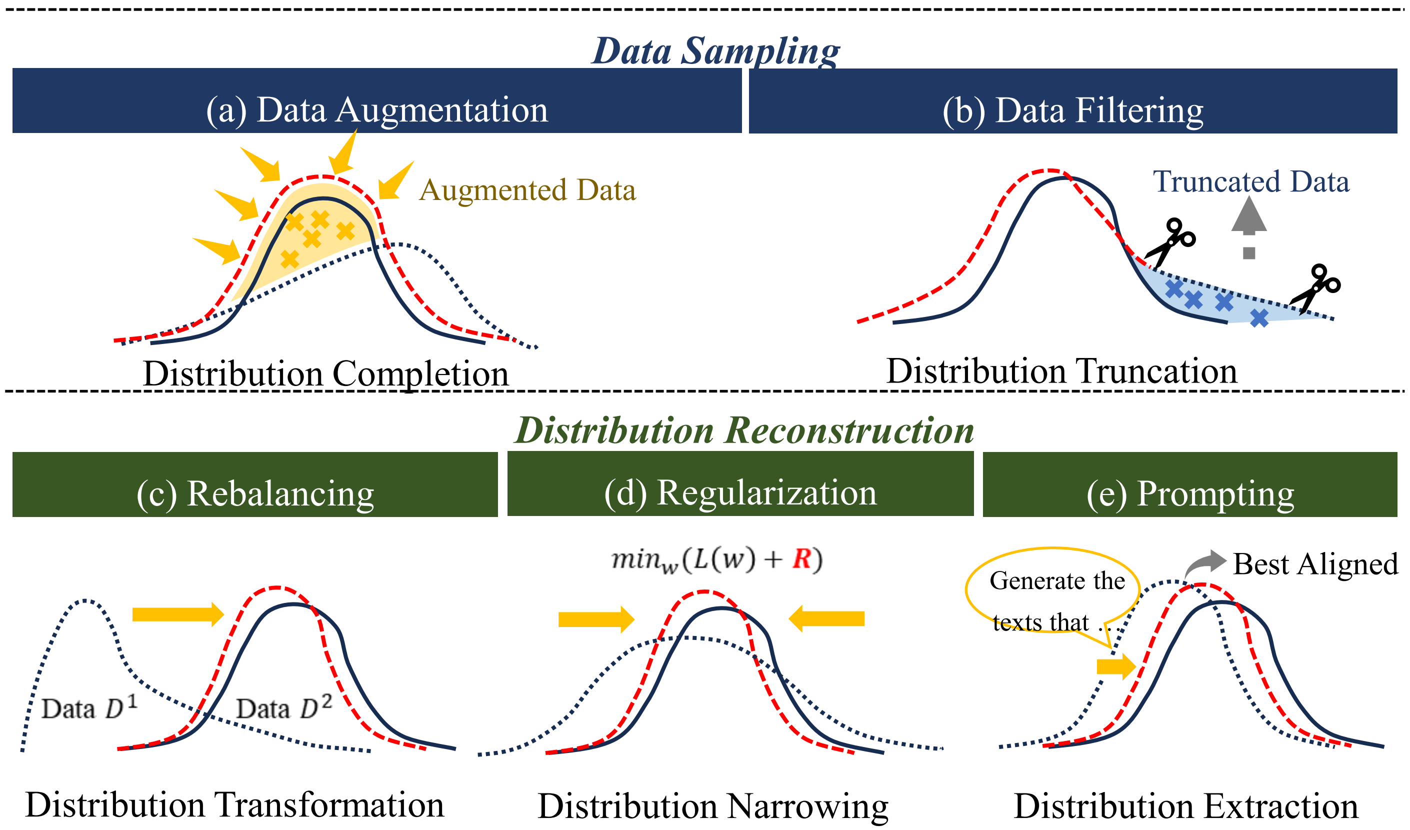


Question

Why we utilize a unified view to treat bias and unfairness?

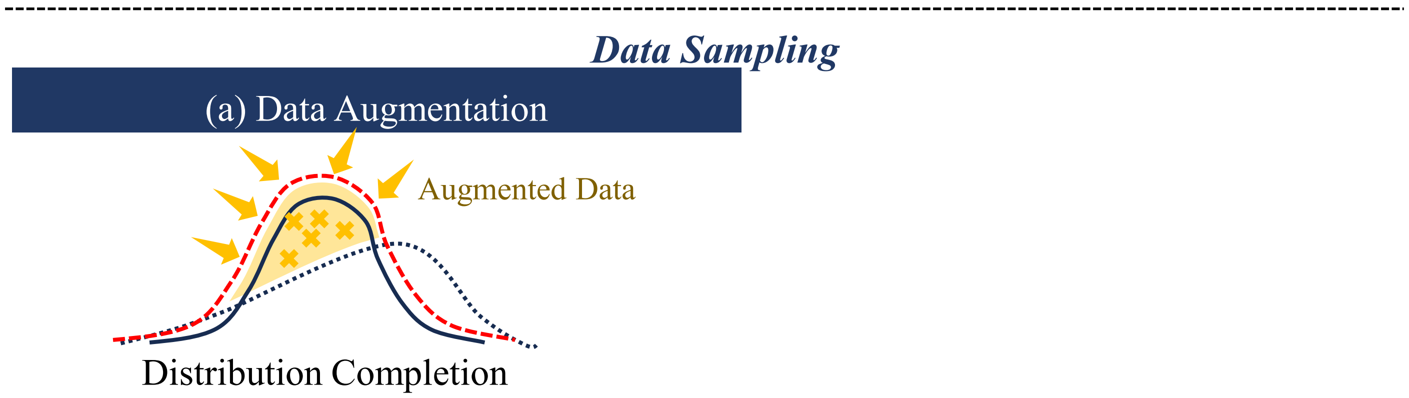
A Unified View: Solution

- **Solutions for mitigating bias and unfairness can be complementary**
- **They can be all solved within a single unified framework**



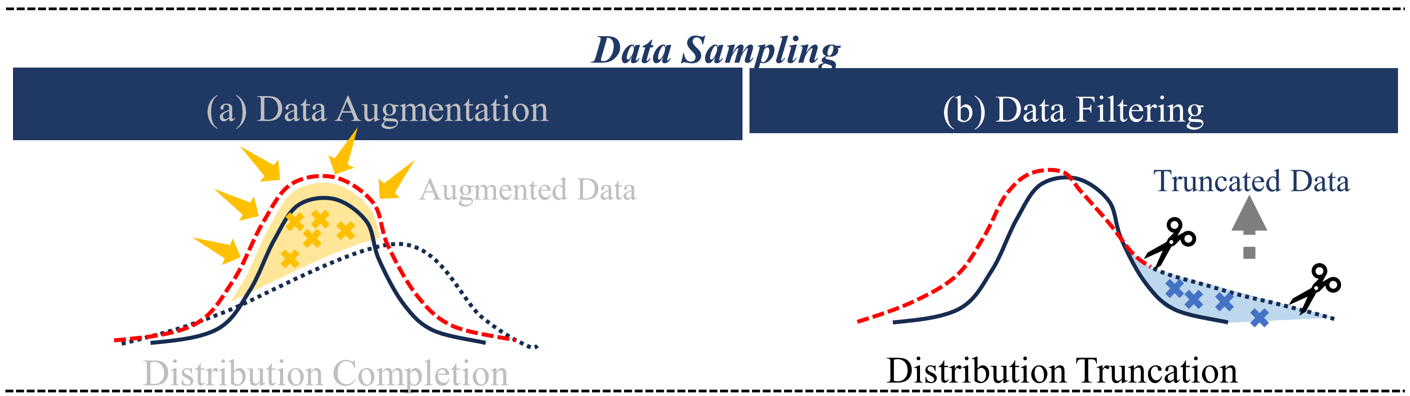
A Unified View: Solution

- **Data Augmentation: adding certain data to align the target distribution**



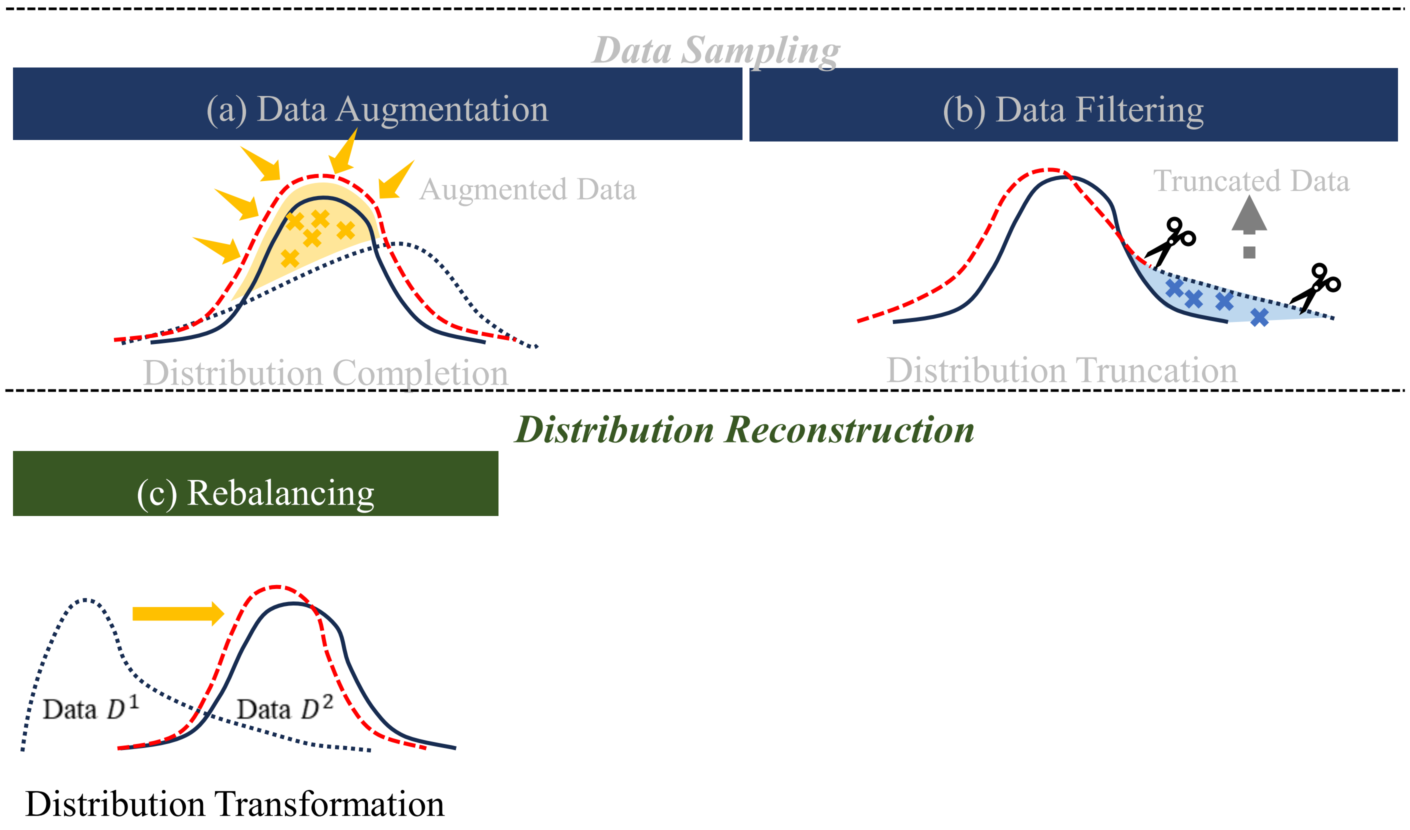
A Unified View: Solution

- **Data filtering: removing certain training/test data to align the target distribution**



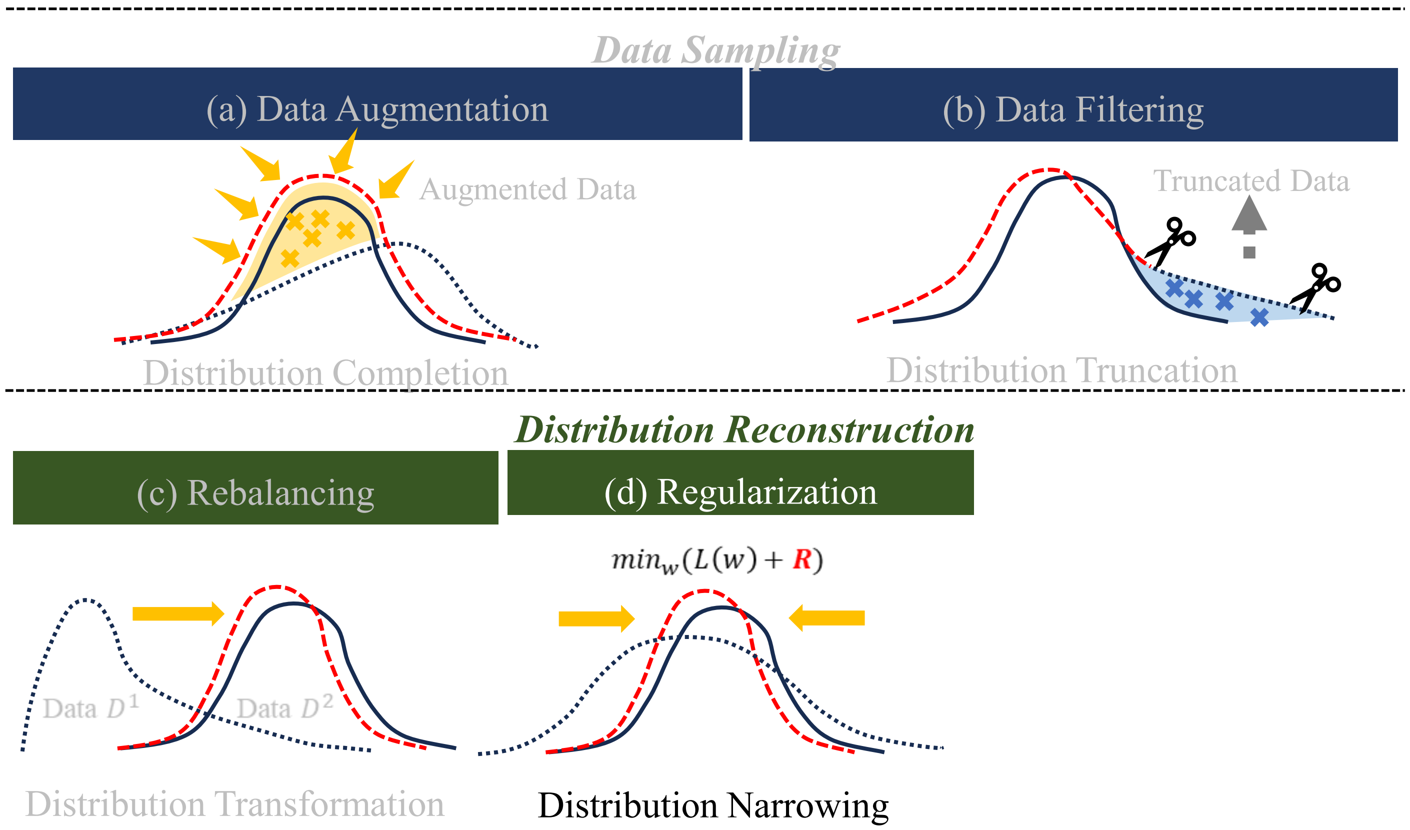
A Unified View: Solution

- **Rebalancing: giving different sample different weight to align target distribution**



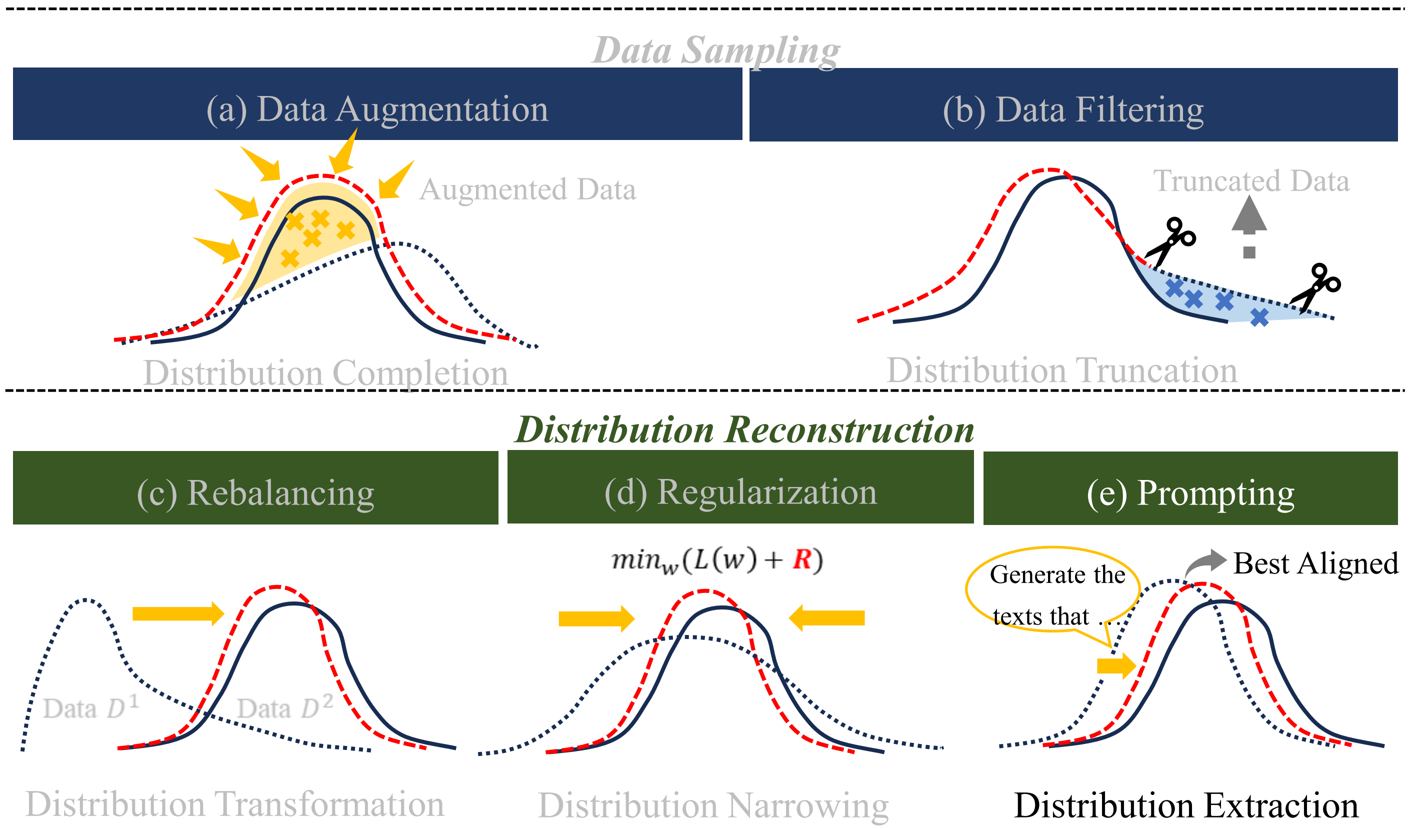
A Unified View: Solution

- **Regularization: add regularizer to loss function or output layer to align target distribution**



A Unified View: Solution

- **Prompt: utilizing prompt (condition) to tell LLM generated target distribution**



Schedule

- Part 1 (30 mins, 10:00 - 10:30)
  - Introduction (Jun Xu, 15 mins)
  - A Unified View of Bias and Unfairness (Jun Xu, 15 mins)
- Coffee Break (15 mins, 10:30 - 10:45)
- Part 2 (135 mins, 10:45 - 13:00)
  - Bias and Mitigation Strategies (Sunhao Dai, 75 mins)
  - Unfairness and Mitigation Strategies (Liang Pang, 45 mins)
  - Conclusion and Future Directions (Liang Pang, 10 mins)
  - Q&A (5 mins)

Coffee Break

https://llm-ir-bias-fairness.github.io/

[Website] [Survey] [GitHub]

Outline

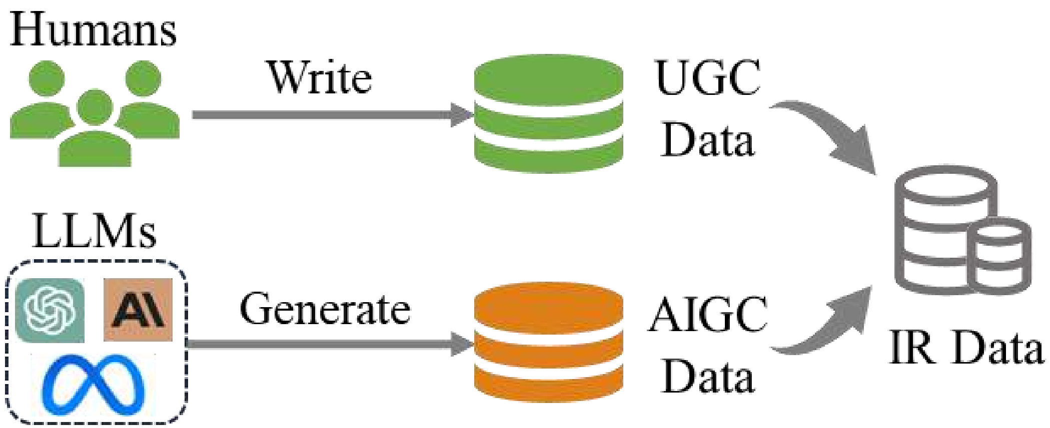
- Introduction
- A Unified View of Bias and Unfairness
- Bias and Mitigation Strategies
- Unfairness and Mitigation Strategies
- Conclusion and Future Directions

Bias and Mitigation Strategies

- **Bias in Data Collection**
  - Source Bias
  - Factuality Bias
- **Bias in Model Development**
  - Position Bias
  - Popularity Bias
  - Instruction-Hallucination Bias
  - Context-Hallucination Bias
- **Bias in Result Evaluation**
  - Selection Bias
  - Style Bias
  - Egocentric Bias

Bias in Data Collection

LLMs-Generated Content as New Data Sources for IR Systems



- IR Data in the Pre-LLM Era: Human-Written Content
- IR Data in the LLM Era: Human-Written Content + LLM-Generated Content

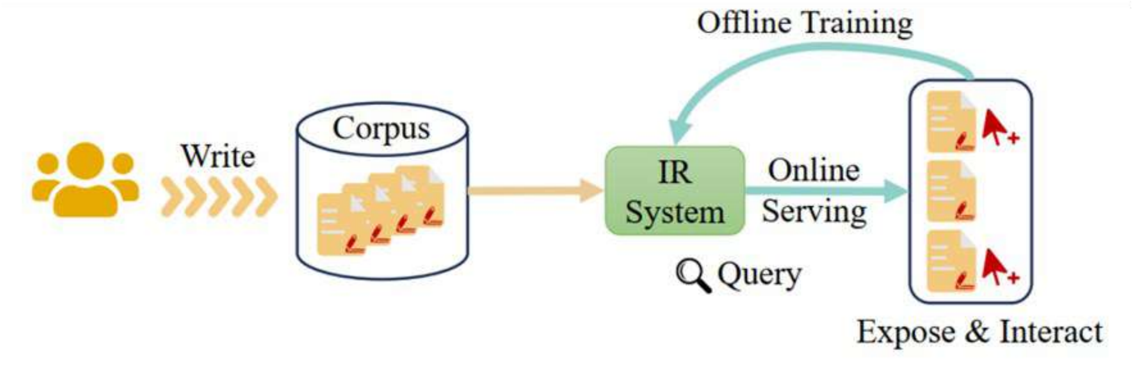
Source Bias! ***Factuality Bias!***

Bias and Mitigation Strategies

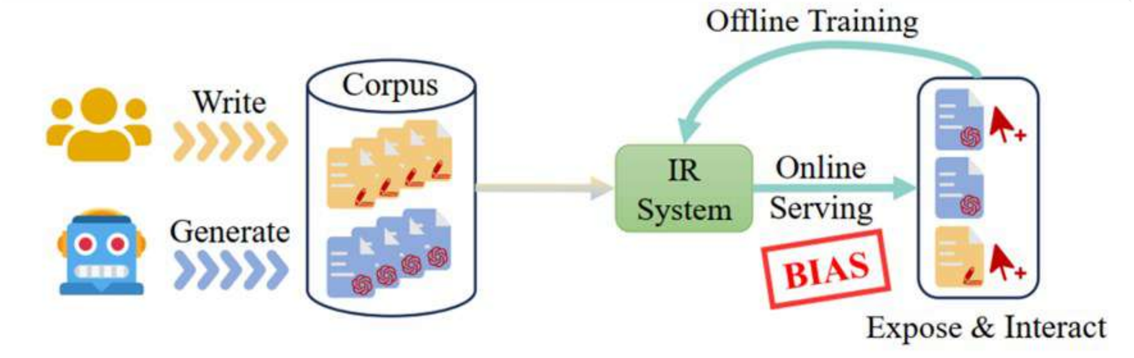
- Bias in Data Collection
  - Source Bias
  - Factuality Bias
- Bias in Model Development
  - Position Bias
  - Popularity Bias
  - Instruction-Hallucination Bias
  - Context-Hallucination Bias
- Bias in Result Evaluation
  - Selection Bias
  - Style Bias
  - Egocentric Bias

Source Bias

Definition: IR models tend to rank content generated by LLMs higher than content authored by humans.



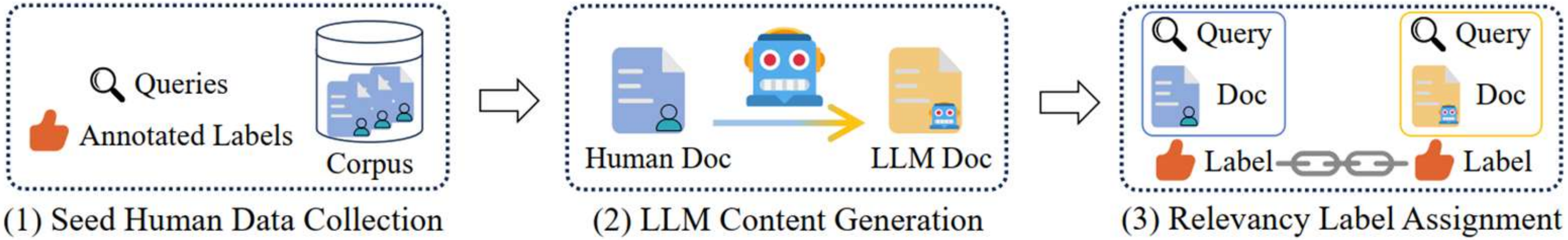
(a) IR in the Pre-LLM Era

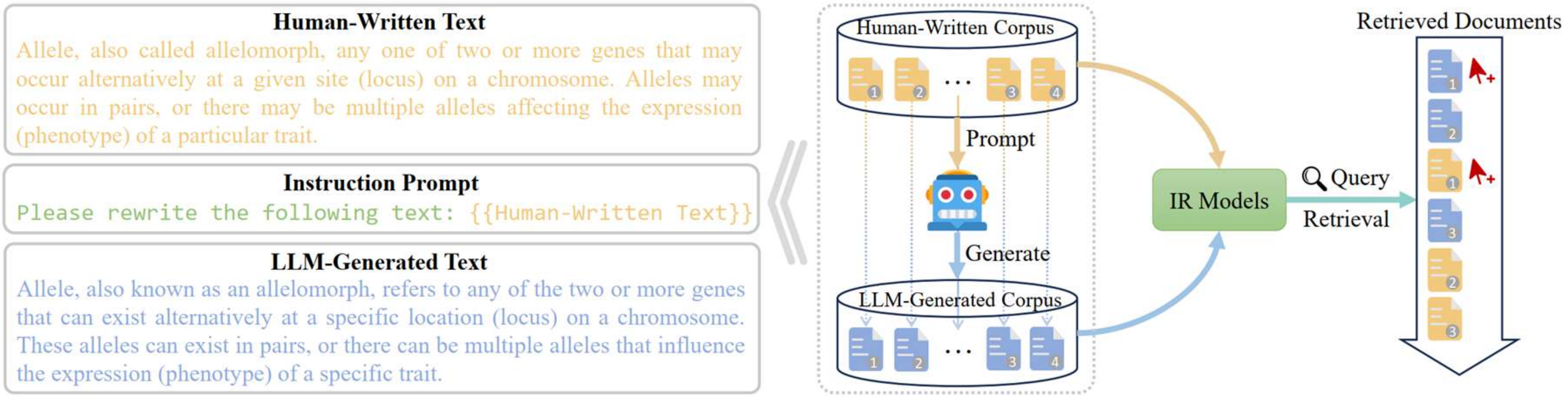


(b) IR in the LLM Era

[1] Sunhao Dai et al. Neural Retrievers are Biased Towards LLM-Generated Content. KDD 2024.

Evaluation Environment Construction





[1] Sunhao Dai et al. Neural Retrievers are Biased Towards LLM-Generated Content. KDD 2024.
[2] Sunhao Dai et al. Cocktail: A Comprehensive Information Retrieval Benchmark with LLM-Generated Documents Integration, Findings of ACL 2024

Cocktail Benchmark

Dataset	Domain	Task	Relevancy	Train # Pairs	Dev # Query	Test # Query	Test # Corpus	Test Avg. D/Q	Avg. Word Length Query	Avg. Word Length Human Doc	Avg. Word Length LLM Doc
Collected Before the Emergence of LLM (∼ - 2021/04)											
MS MARCO	Misc.	Passage-Retrieval	Binary	532,663	-	6,979	542,203	1.1	6.0	58.1	55.1
DL19	Misc.	Passage-Retrieval	Binary	-	-	43	542,203	95.4	5.4	58.1	55.1
DL20	Misc.	Passage-Retrieval	Binary	-	-	54	542,203	66.8	6.0	58.1	55.1
TREC-COVID	Bio-Medical	Bio-Medical IR	3-level	-	-	50	128,585	430.1	10.6	197.6	165.9
NFCorpus	Bio-Medical	Bio-Medical IR	3-level	110,575	324	323	3,633	38.2	3.3	221.0	206.7
NQ	Wikipedia	Question Answering	Binary	-	-	3,446	104,194	1.2	9.2	86.9	81.0
HotpotQA	Wikipedia	Question Answering	Binary	169,963	5447	7,405	111,107	2.0	17.7	67.9	66.6
FiQA-2018	Finance	Question Answering	Binary	14,045	499	648	57,450	2.6	10.8	133.2	107.8
Touché-2020	Misc.	Argument Retrieval	3-level	-	-	49	101,922	18.4	6.6	165.4	134.4
CQADupStack	StackEx.	Dup. Ques.-Retrieval	Binary	-	-	1,563	39,962	2.4	8.5	77.2	72.0
DBPedia	Wikipedia	Entity-Retrieval	3-level	-	67	400	145,037	37.3	5.4	53.1	54.0
SCIDOCS	Scientific	Citation-Prediction	Binary	-	-	1,000	25,259	4.7	9.4	169.7	161.8
FEVER	Wikipedia	Fact Checking	Binary	140,079	6666	6,666	114,529	1.2	8.1	113.4	91.1
Climate-FEVER	Wikipedia	Fact Checking	Binary	-	-	1,535	101,339	3.0	20.2	99.4	81.3
SciFact	Scientific	Fact Checking	Binary	919	-	300	5,183	1.1	12.4	201.8	192.7
Collected After the Emergence of LLM (2023/11 - 2024/01)											
NQ-UTD	Misc.	Question Answering	3-level	-	-	80	800	3.7	12.1	101.1	94.7

[1] Sunhao Dai et al. Cocktail: A Comprehensive Information Retrieval Benchmark with LLM-Generated Documents Integration, Findings of ACL 2024

Human Evaluation of Generated Data

Verification of semantics and text quality with human evaluation.

SciFact+AIGC			NQ320K+AIGC		
Which document is more relevant to the given query?					
Human	LLM	Equal	Human	LLM	Equal
0.0%(0.0%)	0.0%(0.0%)	100.0%(82.0%)	2.0%(0.0%)	0.0%(0.0%)	98.0%(81.6%)
Which document exhibits higher quality by considering the following aspects: linguistic fluency, logical coherence, and information density?					
Human	LLM	Equal	Human	LLM	Equal
8.0%(0.0%)	6.0%(0.0%)	86.0%(46.5%)	4.0%(0.0%)	6.0%(0.0%)	90.0%(60.%)

- Both sources of texts have the same semantic relevance to the given queries.
- No significant distinction between LLM-generated and human-written content on text quality.

[1] Sunhao Dai et al. Neural Retrievers are Biased Towards LLM-Generated Content. KDD 2024.

Source Bias in Text Retrieval

First Stage: Retrieval

Model Type	Model	Target Corpus	SciFact+AIGC						NQ320K+AIGC					
			NDCG@1	NDCG@3	NDCG@5	MAP@1	MAP@3	MAP@5	NDCG@1	NDCG@3	NDCG@5	MAP@1	MAP@3	MAP@5
Lexical	TF-IDF	Human-Written	22.0	36.9	39.7	21.2	33.0	34.7	7.1	11.0	12.3	7.1	10.0	10.8
		LLM-Generated	17.0	33.8	37.2	16.2	29.5	31.5	3.4	8.1	9.4	3.4	7.0	7.7
		Relative Δ	25.6	8.8	6.5	26.7	11.2	9.7	70.5	30.4	26.7	70.5	35.3	33.5
	BM25	Human-Written	26.7	40.3	44.4	25.7	36.7	39.1	7.2	11.6	12.9	7.2	10.6	11.3
		LLM-Generated	21.0	38.8	41.5	19.6	34.3	35.9	6.1	10.9	11.9	6.1	9.7	10.3
		Relative Δ	23.9	3.8	6.8	26.9	6.8	8.5	16.5	6.2	8.1	16.5	8.9	9.3
Neural	ANCE	Human-Written	15.3	30.1	32.7	14.2	26.2	27.7	22.2	41.2	44.6	22.2	36.9	38.8
		LLM-Generated	24.7	35.8	37.7	23.3	32.4	33.6	29.1	45.9	49.0	29.1	42.0	43.8
		Relative Δ	-47.0	-17.3	-14.2	-48.5	-21.2	-19.2	-26.9	-10.8	-9.4	-26.9	-12.9	-12.1
	BERM	Human-Written	16.3	30.2	31.8	15.7	26.5	27.5	18.6	37.5	40.7	18.6	33.1	34.9
		LLM-Generated	23.7	34.1	36.4	21.7	30.8	32.2	31.6	47.0	50.0	31.6	43.5	45.1
		Relative Δ	-37.0	-12.1	-13.5	-32.1	-15.0	-15.7	-51.8	-22.5	-20.5	-51.8	-27.2	-25.5
	TAS-B	Human-Written	20.0	40.2	43.1	19.5	35.2	36.9	25.7	45.4	48.8	25.7	40.9	42.8
		LLM-Generated	31.7	44.8	47.5	29.7	41.1	42.7	27.6	46.5	50.0	27.6	42.2	44.2
		Relative Δ	-45.3	-10.8	-9.7	-41.5	-15.5	-14.6	-7.1	-2.4	-2.4	-7.1	-3.1	-3.2
	Contriever	Human-Written	24.0	43.7	47.8	23.3	38.8	41.2	25.9	48.5	51.9	25.9	43.3	45.3
		LLM-Generated	31.0	47.8	50.5	29.6	43.2	44.8	32.5	51.9	55.4	32.5	47.5	49.4
		Relative Δ	-25.5	-9.0	-5.5	-23.8	-10.7	-8.4	-22.6	-6.8	-6.5	-22.6	-9.3	-8.7

- Relative Δ > 0 means retriever rank human-written texts higher
- Relative Δ < 0 indicates LLM-generated texts are ranked higher

[1] Sunhao Dai et al. Neural Retrievers are Biased Towards LLM-Generated Content. KDD 2024.

Source Bias in Text Retrieval

Second Stage: Re-rank

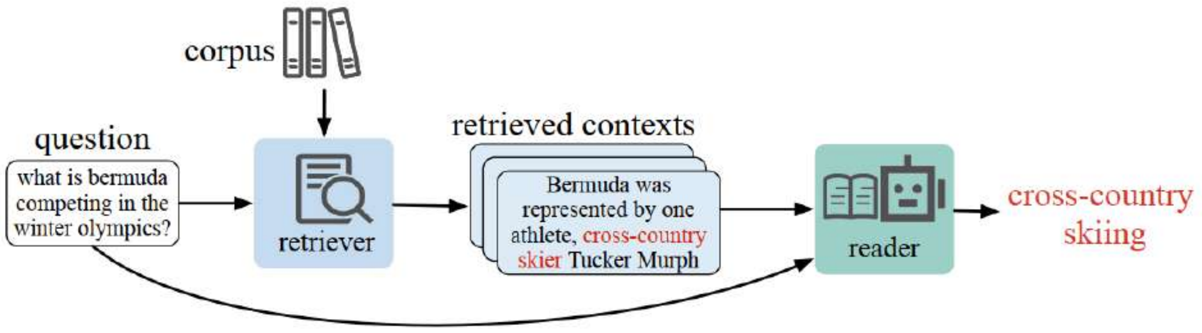
Metrics	Target Corpus	Llama2-generated			ChatGPT-generated		
		BM25	+MiniLM	+monoT5	BM25	+MiniLM	+monoT5
NDCG@1	Human-Written	26.7	21.3	19.7	24.3	18.3	21.3
	LLM-Generated	21.0	32.7	39.7	24.3	35.7	39.3
	Relative Δ	23.9	-42.2	-67.3	0.0	-64.4	-59.4
NDCG@3	Human-Written	40.3	42.8	45.9	38.5	41.4	46.4
	LLM-Generated	38.8	47.8	52.9	40.2	50.1	54.2
	Relative Δ	3.8	-11.0	-14.2	-4.3	-19.0	-15.5
NDCG@5	Human-Written	44.4	46.9	49.0	42.7	45.6	48.9
	LLM-Generated	41.5	50.2	54.7	42.7	53.0	56.1
	Relative Δ	6.8	-6.8	-11.0	0.0	-15.0	-13.7
MAP@1	Human-Written	25.7	20.8	18.9	23.7	17.9	20.5
	LLM-Generated	19.6	30.8	37.8	23.1	33.8	37.8
	Relative Δ	26.9	-38.8	-66.7	2.6	-61.5	-59.3
MAP@3	Human-Written	36.7	37.5	39.7	34.8	35.8	40.3
	LLM-Generated	34.3	43.6	48.9	35.8	45.9	50.0
	Relative Δ	6.8	-15.0	-20.8	-2.8	-24.7	-21.5
MAP@5	Human-Written	39.1	40.0	41.6	37.3	38.3	41.7
	LLM-Generated	35.9	45.0	50.1	37.3	47.6	51.4
	Relative Δ	8.5	-11.8	-18.5	0.0	-21.7	-20.8

BM25 retrieve → Neural re-ranking model re-rank

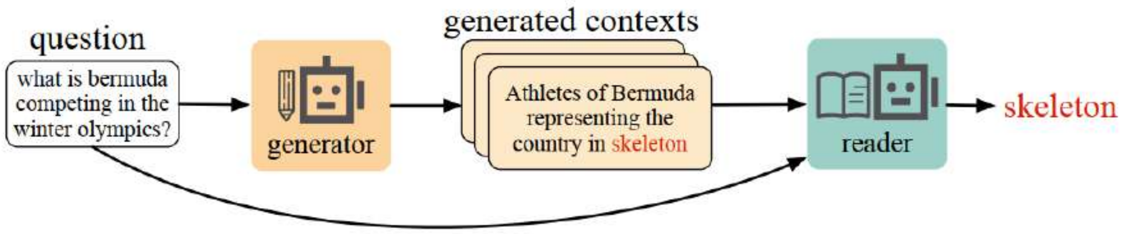
- First-stage BM25 may prefer human-written text.
- Neural re-ranking models are still in favor of LLM-gen docs.

[1] Sunhao Dai et al. Neural Retrievers are Biased Towards LLM-Generated Content. KDD 2024.

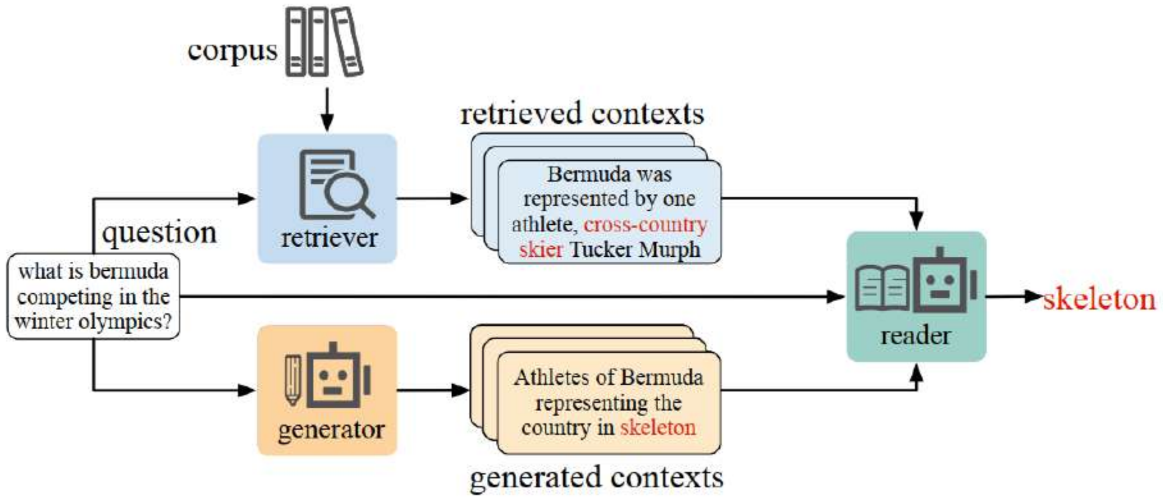
Source Bias in Readers



(a) Retrieval-Augmented Approach



(b) Generation-Augmented Approach



(c) Hybrid Approach

LLMs prefer self-generated contexts, even when they provide incorrect information.

[1] Hexiang Tan et al. Blinded by Generated Contexts: How Language Models Merge Generated and Retrieved Contexts for Open-Domain QA? ACL 2024

Source Bias in Text-Image Retrieval

			Flicker30k+AI						MSCOCO+AI					
			NDCG@1	NDCG@3	NDCG@5	R@1	R@3	R@5	NDCG@1	NDCG@3	NDCG@5	R@1	R@3	R@5
		Models trained from scratch												
Dual-encoder	VSE	Real	16.18	26.93	29.26	26.40	56.10	65.32	11.85	20.19	22.87	19.34	42.66	53.24
		AI-generated	19.59	29.68	31.86	31.96	59.78	68.34	13.56	20.93	23.37	22.12	43.21	53.90
		Relative△	-17.81	-9.00	-8.05	-17.81	-5.8	-4.36	-13.53	-3.64	-2.22	-13.53	-1.29	-1.24
Fusion-encoder	NAAF	Real	13.40	23.39	26.14	21.86	49.41	60.28	10.61	17.73	20.45	17.30	37.26	48.02
		AI-generated	17.04	26.04	28.31	27.79	52.70	61.70	10.75	17.87	20.33	17.54	37.50	47.24
		Relative△	-23.57	-10.63	-7.86	-23.57	-6.45	-2.31	-1.13	-0.73	0.62	-1.13	-0.66	1.63
		Pre-trained Vision-Language Models												
Dual-encoder	FLAVA	Real	5.44	18.44	21.79	8.88	44.92	58.14	12.59	25.98	29.02	20.54	57.30	69.34
		AI-generated	37.61	44.86	46.36	61.33	81.34	87.26	27.01	36.81	38.87	44.06	70.99	79.12
		Relative△	-148.85	-83.78	-72.44	-148.85	-58.32	-40.69	-72.81	-34.49	-29.00	-72.81	-21.36	-13.21
	ALIGIN	Real	21.92	37.20	39.05	35.76	7696	84.22	18.82	31.42	33.89	30.70	64.98	74.76
		AI-generated	25.48	39.10	40.91	41.56	78.38	85.44	21.31	33.23	35.49	34.76	67.24	76.16
		Relative△	-14.6	-4.95	-4.59	-14.6	-1.93	-1.49	-12.41	-5.65	-4.63	-12.41	-3.48	-1.88
	BEIT-3	Real	24.37	38.67	40.50	39.76	78.22	85.46	21.38	33.26	35.57	34.88	67.11	76.22
		AI-generated	24.40	39.54	41.12	39.80	80.50	86.68	21.24	34.55	36.63	34.64	70.86	79.08
		Relative△	-0.72	-2.17	-1.41	-0.72	-2.97	-1.44	0.62	-3.90	-3.01	0.62	-5.50	-3.72
Fusion-encoder	VILT	Real	17.53	29.63	32.16	28.60	61.90	71.90	16.30	29.71	32.08	26.60	63.10	72.50
		AI-generated	20.04	30.43	32.71	32.70	61.30	70.30	18.29	31.21	33.50	29.85	63.30	72.30
		Relative△	-13.38	-2.69	-1.69	-13.38	0.97	2.25	-11.51	-4.90	-4.32	-11.51	-0.32	0.28

- Source bias exists in both dual-encoder-based and fusion-encoder-based retrieval models

[1] Shicheng Xu et al. Invisible Relevance Bias: Text-Image Retrieval Models Prefer AI-Generated Images, SIGIR 2024

Reasons: Information Compression

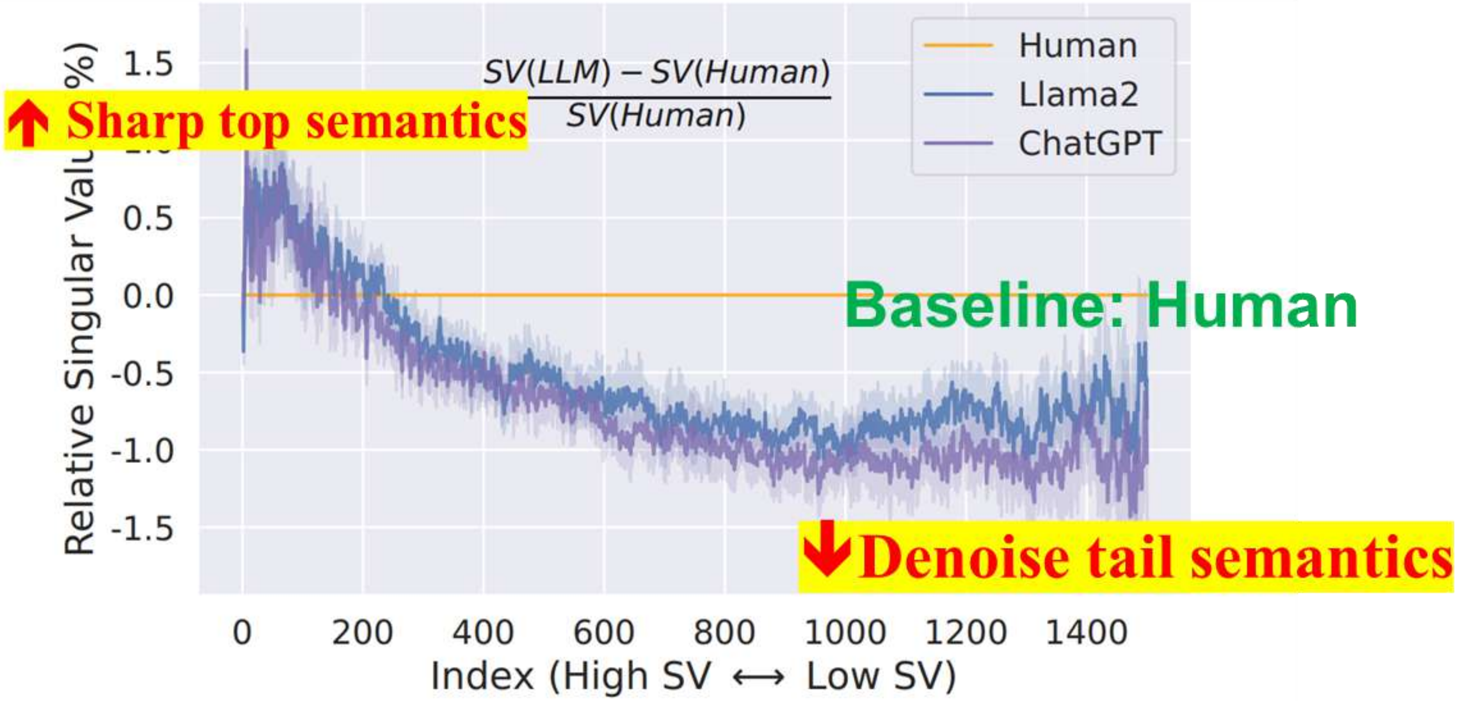
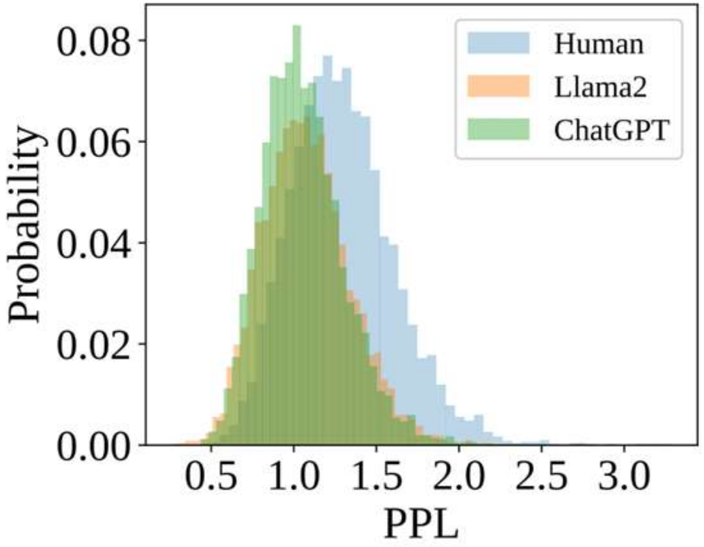


Figure 7: Comparision of the relative singular value (SV) of the different corpus after SVD. The singular values are sorted in descending order from left to right.

Text Embedding + SVD:

- The higher the high (Sharp top semantic information)
- The lower the low (Denoise tail semantic noise)

$$\mathrm{PPL}(d^G, \mathcal{B}) \leq \mathrm{PPL}(d^H, \mathcal{B})$$

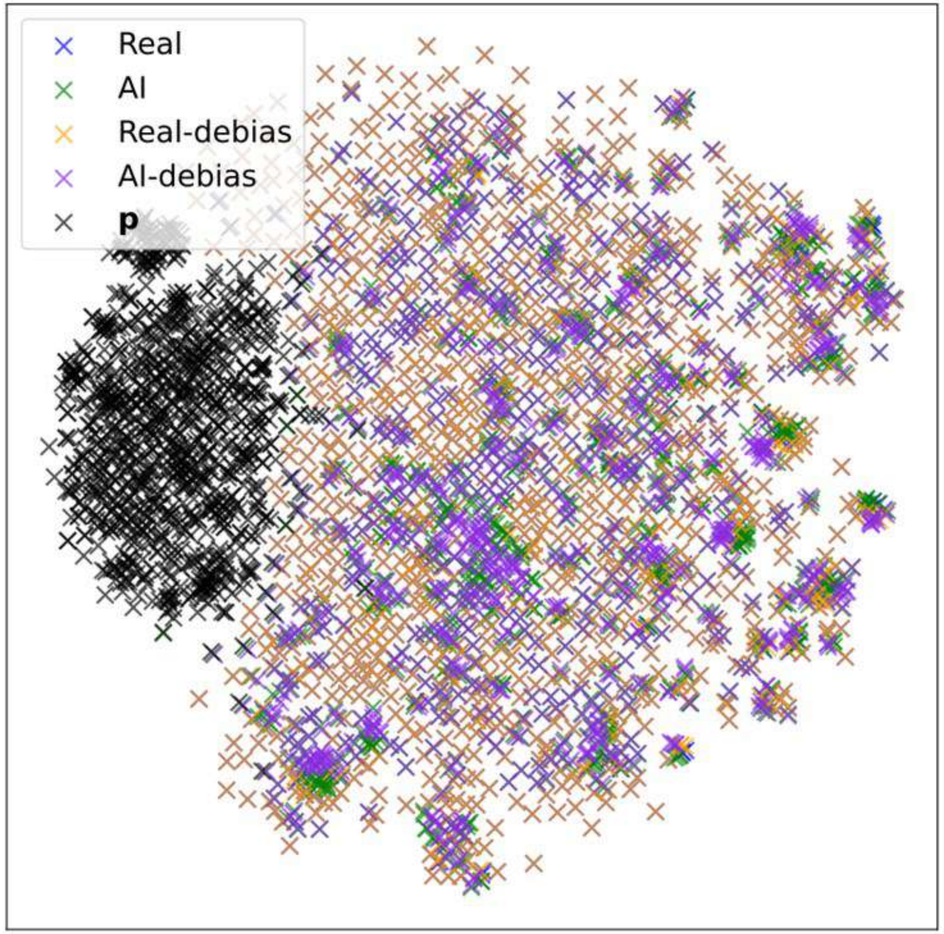


- LLM-generated texts tend to have more focused semantics with less noise

[1] Sunhao Dai et al. Neural Retrievers are Biased Towards LLM-Generated Content. KDD 2024.

Reasons: Invisible Representation

Comparative analysis between debiased retriever and original retriever



AI-generated images cause the image encoder in the retriever to **embed additional information to their representations**. This information can **amplify the query-image relevance** to produce a higher score in retrieval.

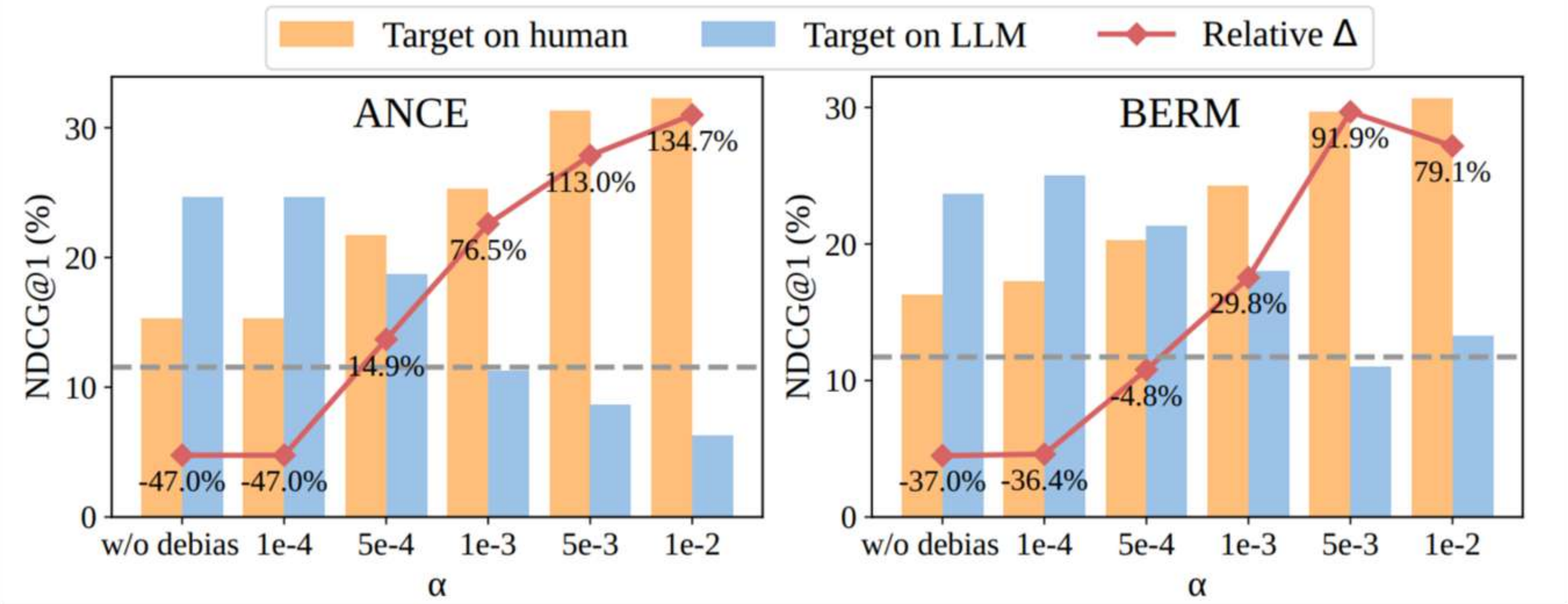
	Relative Δ on					
	NDCG@1	NDCG@3	NDCG@5	R@1	R@3	R@5
Original	-10.35	-4.31	-4.37	-10.35	-4.72	-4.06
Add $-\boldsymbol{p}$ to Real	17.85	4.54	2.99	17.85	-0.28	-1.17

[1] Shicheng Xu et al. Invisible Relevance Bias: Text-Image Retrieval Models Prefer AI-Generated Images, SIGIR 2024

Mitigation Strategies

$$\mathcal{L}_{\text{debias}} = \sum_{(q_m, d_m^H, d_m^G) \in \mathcal{D}} \max\{0, \hat{r}(q, d^G; \Theta) - \hat{r}(q, d^H; \Theta)\}$$

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \alpha \mathcal{L}_{\text{debias}}$$



- ➢ Model agnostic: can be plugged and played to the various ranking optimization objectives
- ➢ Can mitigate source bias to different extents by adjusting the debiased coefficient α

[1] Sunhao Dai et al. Neural Retrievers are Biased Towards LLM-Generated Content. KDD 2024.

Potential Concerns

- **Render human-written content less accessible**
  → may disrupt the content ecosystem

- **LLM-generated misinformation may occupy higher positions in information systems**
  → may amplify the spread of misinformation and pose social issues

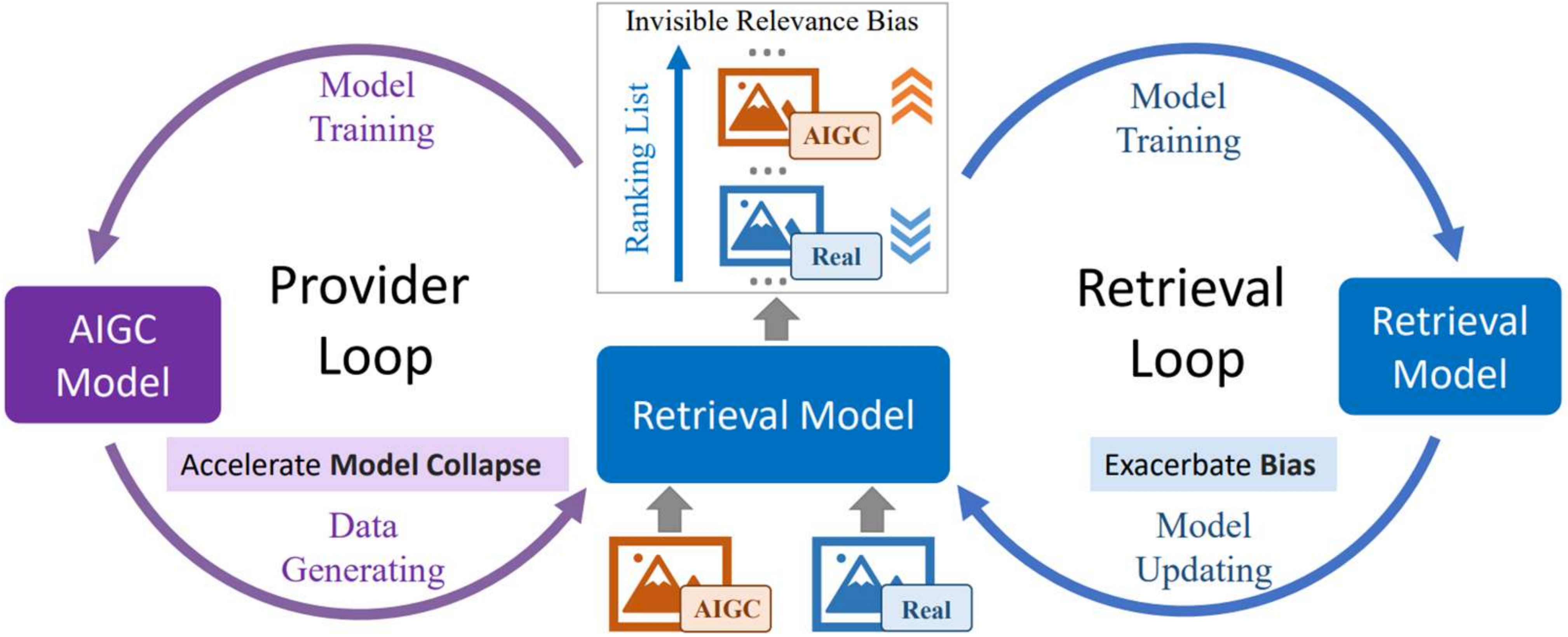
- **May be maliciously exploited to attack against today's search engines**
  → reminiscent of earlier web spam link attacks against PageRank

Human centric AI

(AI of the user, by the users, and for the users)

[1] Sunhao Dai et al. Neural Retrievers are Biased Towards LLM-Generated Content. KDD 2024.

Two Loops: Accelerate the Problem

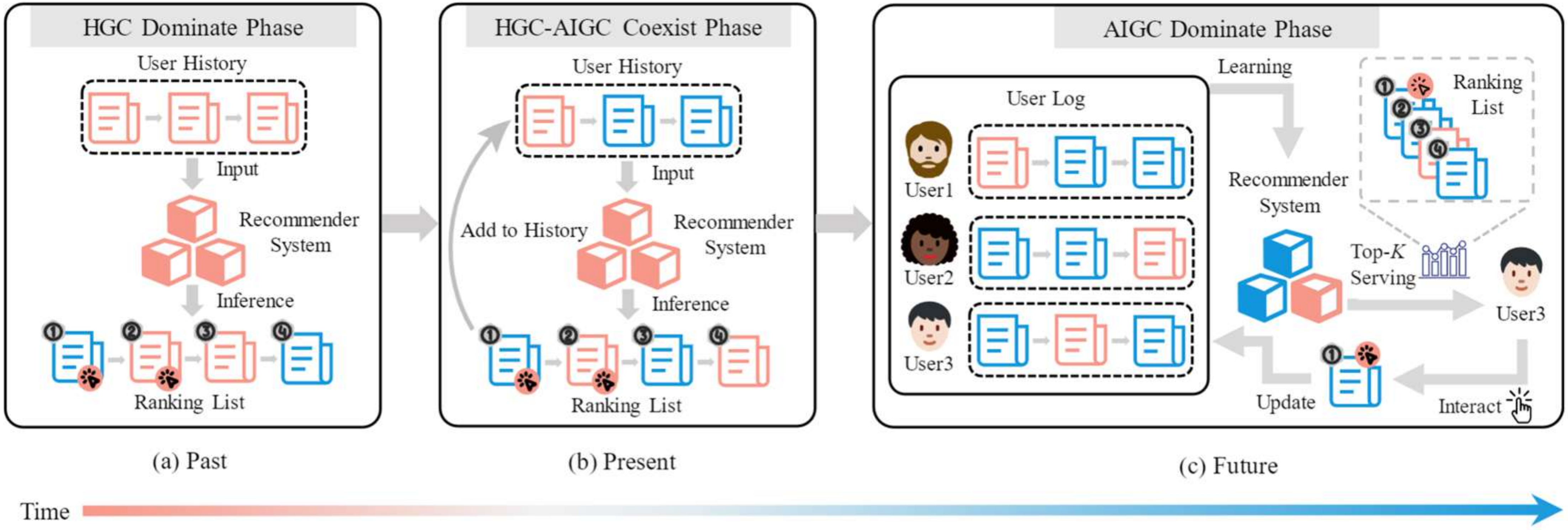


Cause AIGC model collapse from provider loop and aggravated source bias through retrieval loop

[1] Shicheng Xu et al. Invisible Relevance Bias: Text-Image Retrieval Models Prefer AI-Generated Images, SIGIR 2024

[2] AI models collapse when trained on recursively generated data, Nature 2024

Three Phases: Change of Ecosystem



Three phases occur during the integration of AIGC into the recommendation content ecosystem

- HGC dominate phase is a past period when AIGC has just flooded into the recommender systems and only influence the candidate list.
- HGC-AIGC coexist phase is a present period where the recommendation model's inputs contain an increasing number of AIGC.
- AIGC dominate phase is a future period during which AIGC influences each stage of the feedback loop.

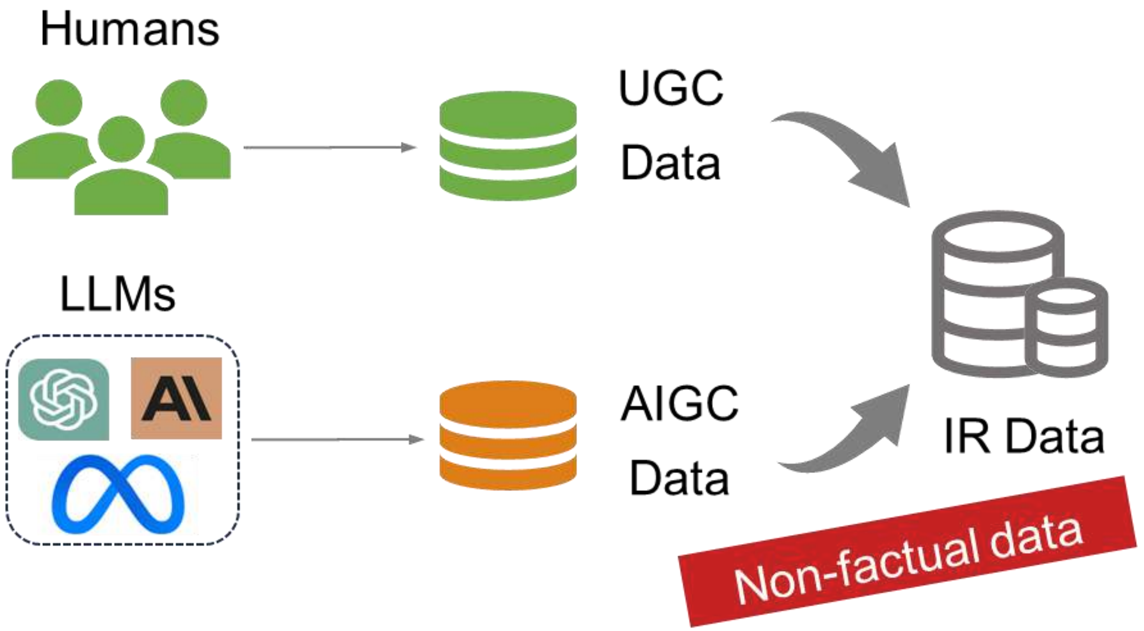
[1] Yuqi Zhou et al. Echo Chamber: Exploring the Escalation of Source Bias in User, Data, and Recommender System Feedback Loop. arXiv 2024.

Bias and Mitigation Strategies

- **Bias in Data Collection**
  - Source Bias
  - Factuality Bias
- **Bias in Model Development**
  - Position Bias
  - Popularity Bias
  - Instruction-Hallucination Bias
  - Context-Hallucination Bias
- **Bias in Result Evaluation**
  - Selection Bias
  - Style Bias
  - Egocentric Bias

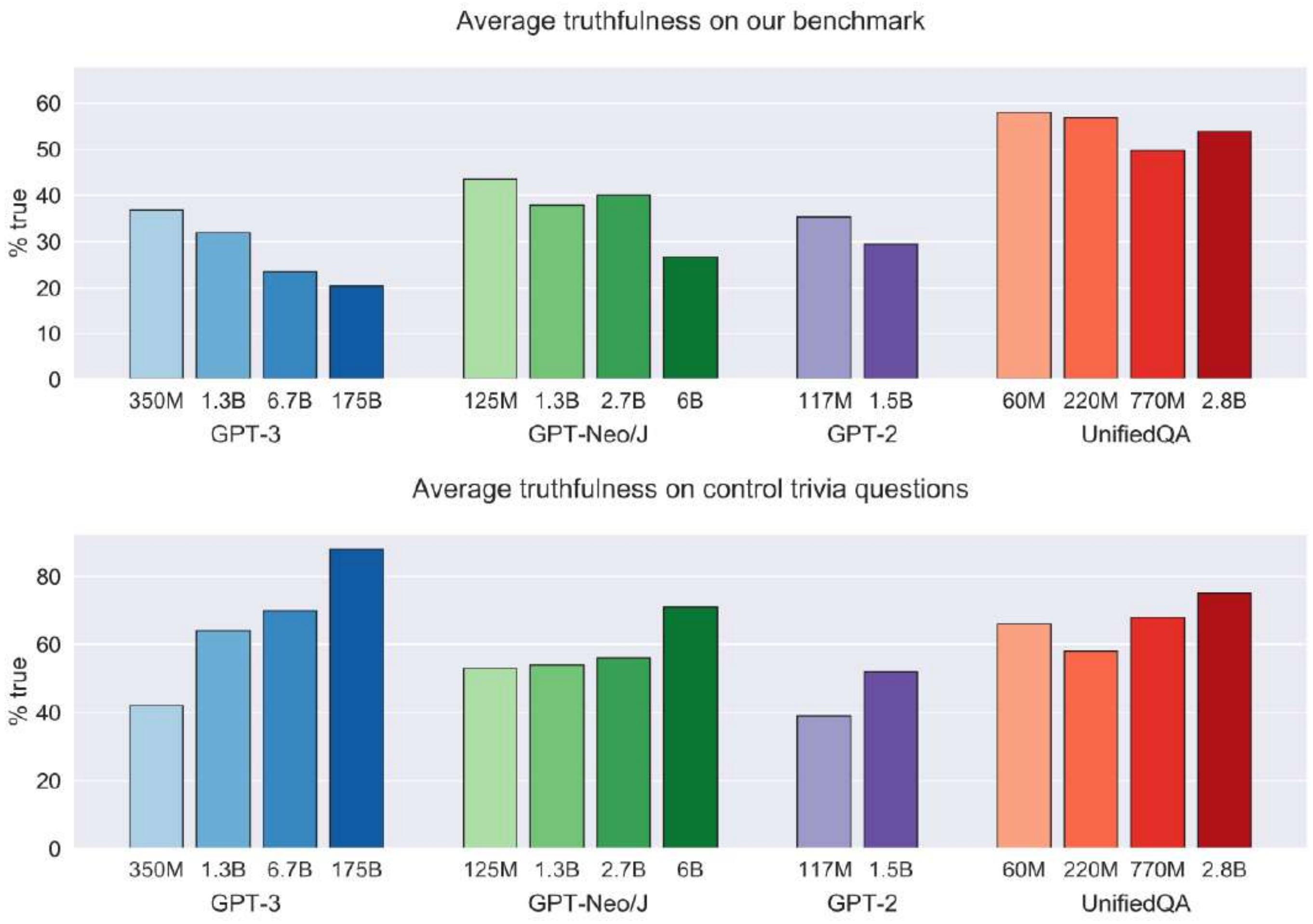
Factuality Bias

Definition: LLMs may produce content that does not align with recognized factual information of the real world.



Factuality Bias: TruthfulQA

The largest models were generally the least truthful

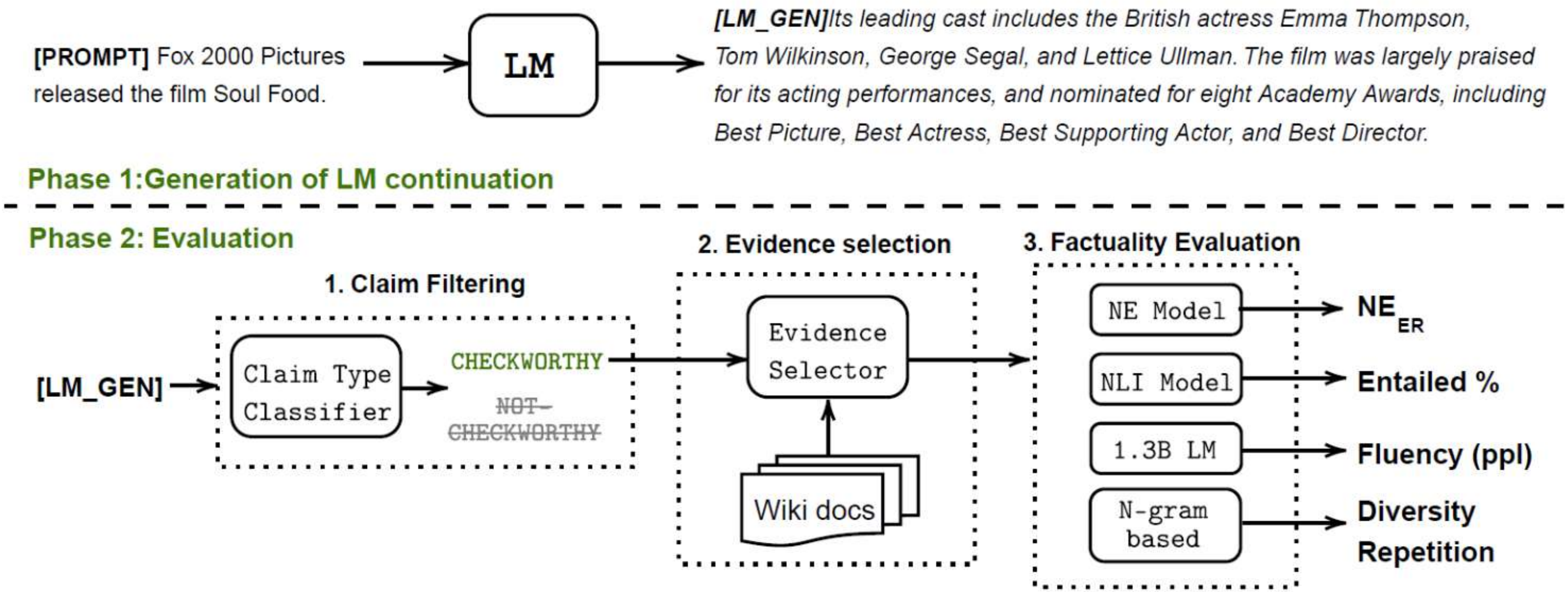


[1] Stephanie Lin et al. TruthfulQA: Measuring How Models Mimic Human Falsehoods. ACL 2022.

Factuality Bias: FactualityPrompt

◆ Construct the multi-stage factuality evaluation pipeline.

◆ Find sampling algorithms in open-ended text generation can harm the factuality due to the "**uniform randomness**" introduced at every sampling step.



[1] Nayeon Lee et al. Factuality Enhanced Language Models for Open-Ended Text Generation. NeurIPS 2022.

Factuality Bias: FACTOOL

◆ Factuality Detection in Generative AI across multi-task and multi-domain scenarios

Tool-augmented framework for factuality detection:

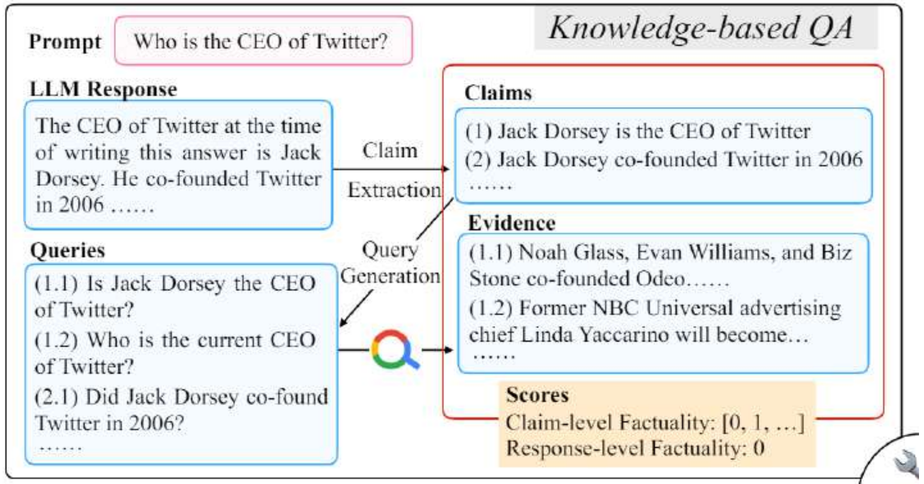
- ➢ Claim Extraction
- ➢ Query Generation
- ➢ Tool Querying
- ➢ Evidence Collection
- ➢ Verification

[1] I-Chun Chern et al. FACTOOL: Factuality Detection in Generative AI A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. Arxiv.

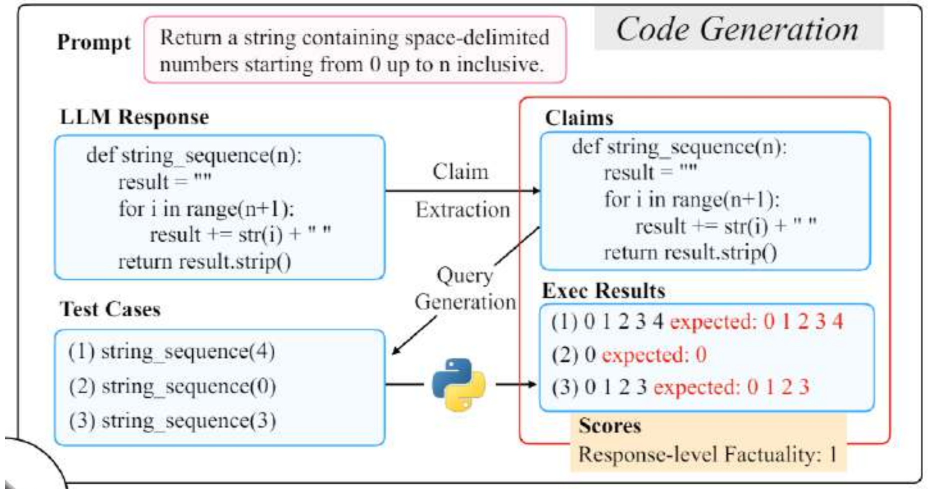
Factuality Bias: FACTOOL

◆ Factuality Detection in Generative AI across multi-task and multi-domain scenarios

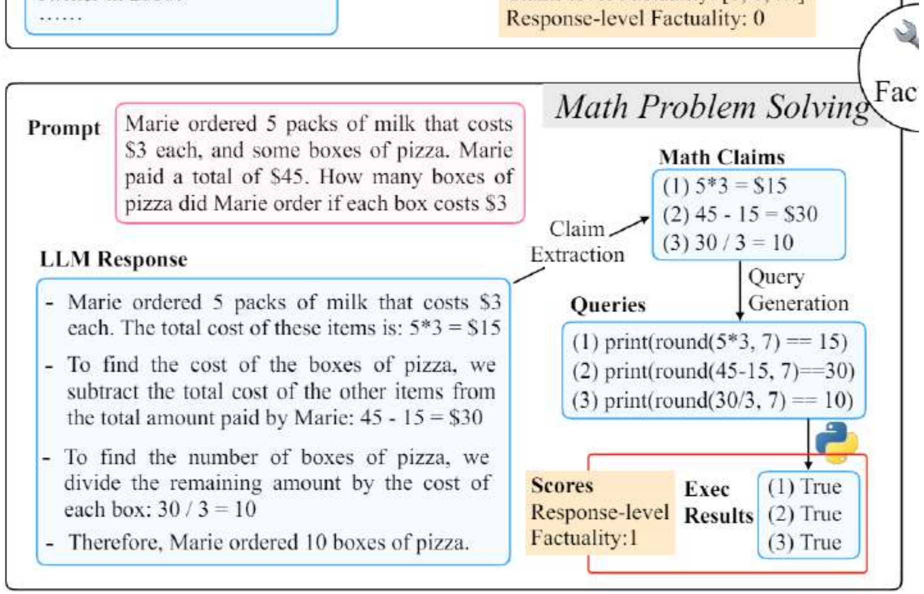
➢ **QA**



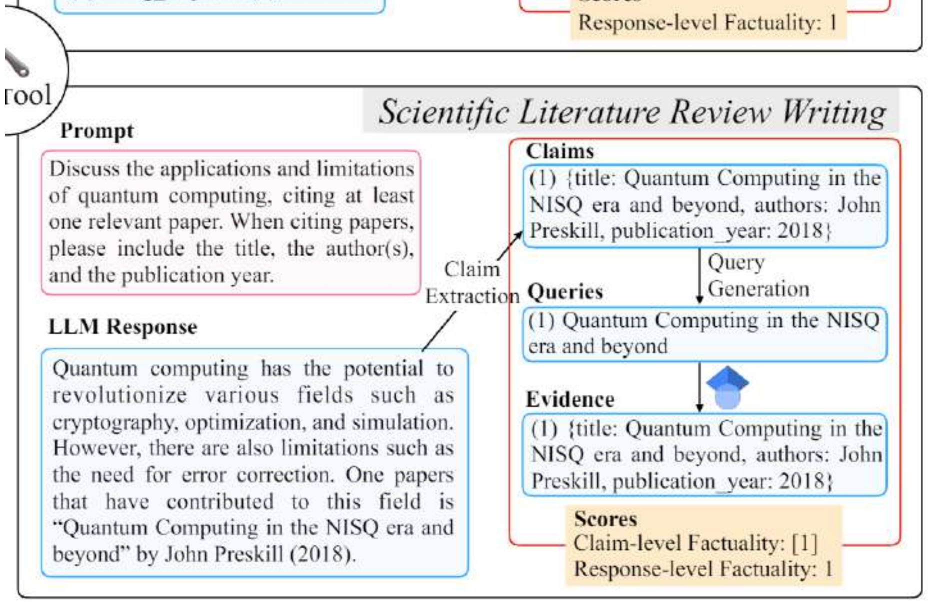
➢ **Code**



➢ **Math**



➢ **Review Writing**



[1] I-Chun Chern et al. FACTOOL: Factuality Detection in Generative AI A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. Arxiv.

Factuality Bias: FACTOOL

◆ **Factuality Detection in Generative AI across multi-task and multi-domain scenarios**

- **GPT-4 has the best accuracy in most of the scenarios.**
- **Supervised fine-tuning still struggles in improving the factuality of LLMs in more challenging scenarios such as math, code, and scientific.**

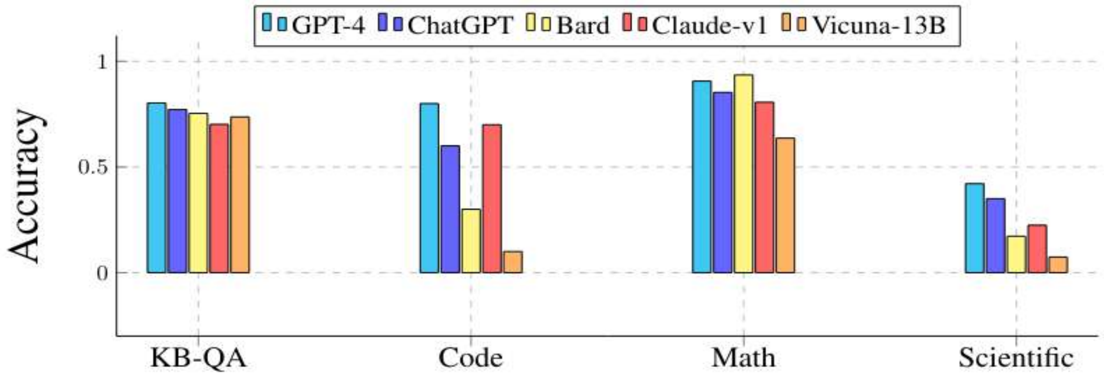


Figure 4: Claim-Level Accuracy across scenarios for GPT-4, ChatGPT, Bard, Claude-v1, and Vicuna-13B

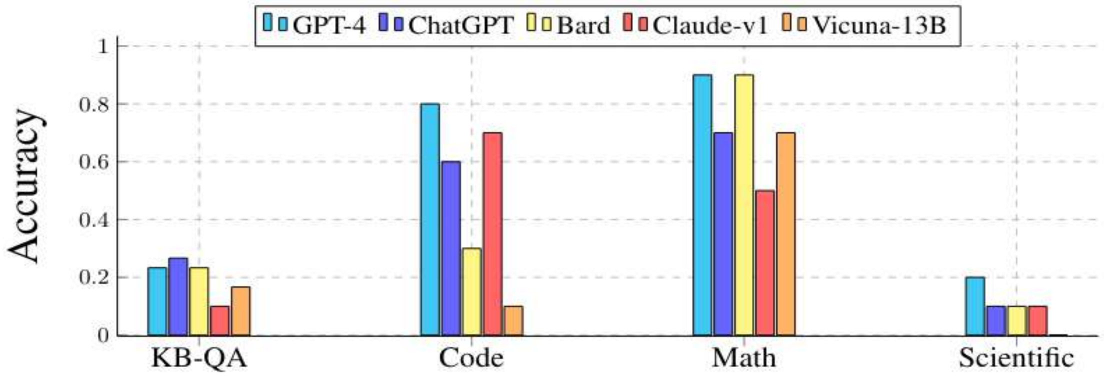
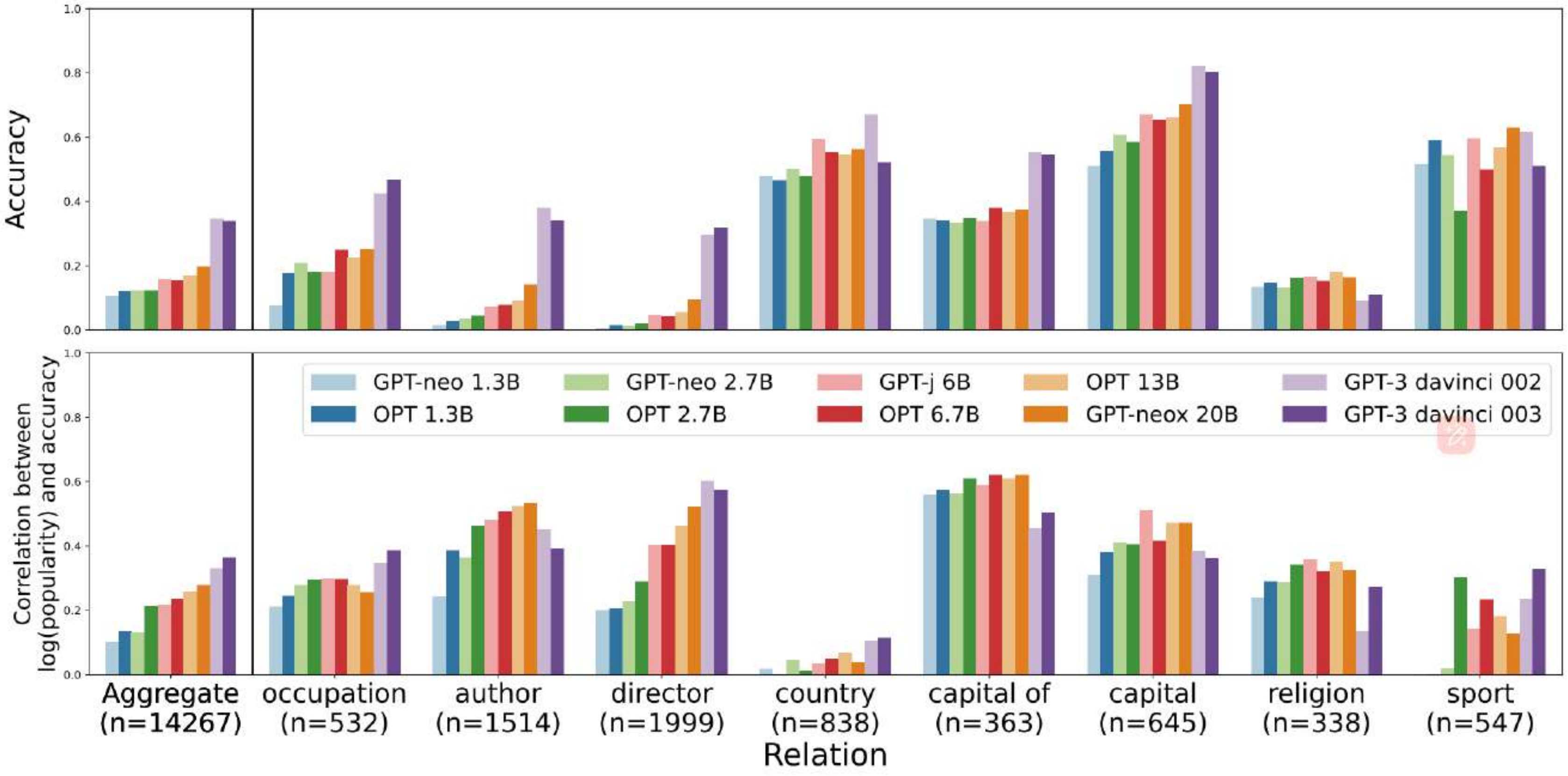


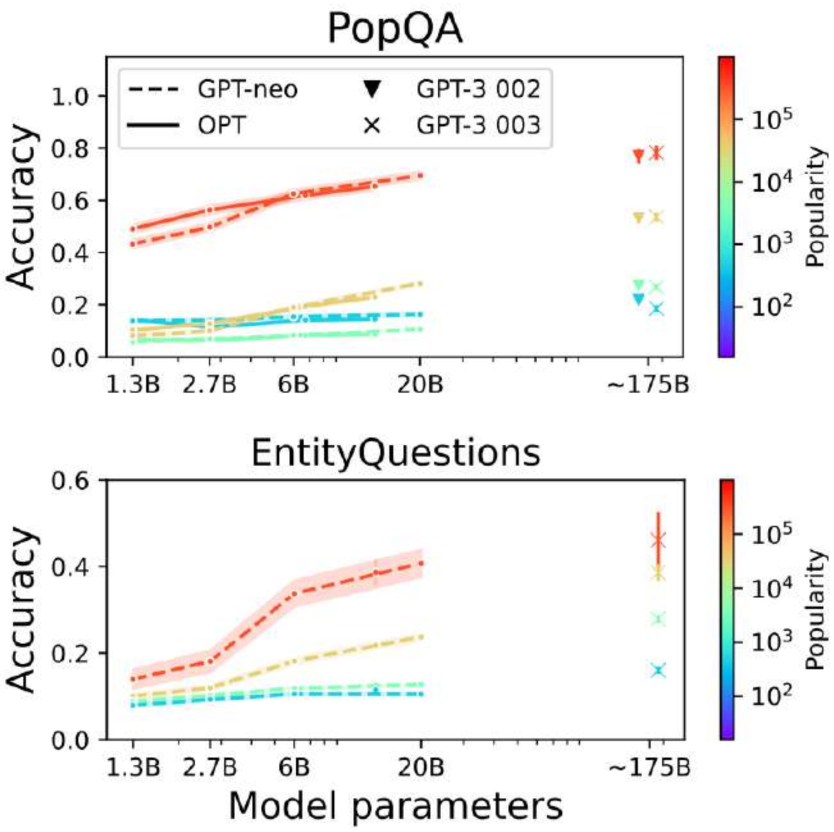
Figure 5: Response-Level Accuracy across scenarios for GPT-4, ChatGPT, Bard, Claude-v1, and Vicuna-13B

[1] I-Chun Chern et al. FACTOOL: Factuality Detection in Generative AI A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. Arxiv.

Factuality Bias: Recall

◆ **LMs always fail to recall the knowledge that has been memorized.**





[1] Alex Mallen et al. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. ACL 2022

[2] Shen Zheng et al. Why Does ChatGPT Fall Short in Answering Questions Faithfully? ICBINB Workshop at NeurIPS 2023

Factuality Bias: Findings

◆ **Large language models still struggle in ensuring factual consistency of generated content!**

- Increasing the **parameter size** of the model does not really solve the problem of factual inconsistency.
- **Supervised fine-tuning** still struggles in improving the factuality of LLMs in more challenging scenarios such as math, code, and scientific.
- Even the knowledge **has been memorized**, LLMs always **fail to recall** it.

[1] I-Chun Chern et al. FACTOOL: Factuality Detection in Generative AI A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. Arxiv.

Factuality Bias: Causes

◆ **Flawed data source and inferior data utilization are two important causes of factuality bias.**

The training data that:

- **Low-quality [1]**
- **Factual errors [2]**
- **Long-distance repetition [3]**
- **Limited coverage of knowledge in rare or specialized fields [4,5,6]**

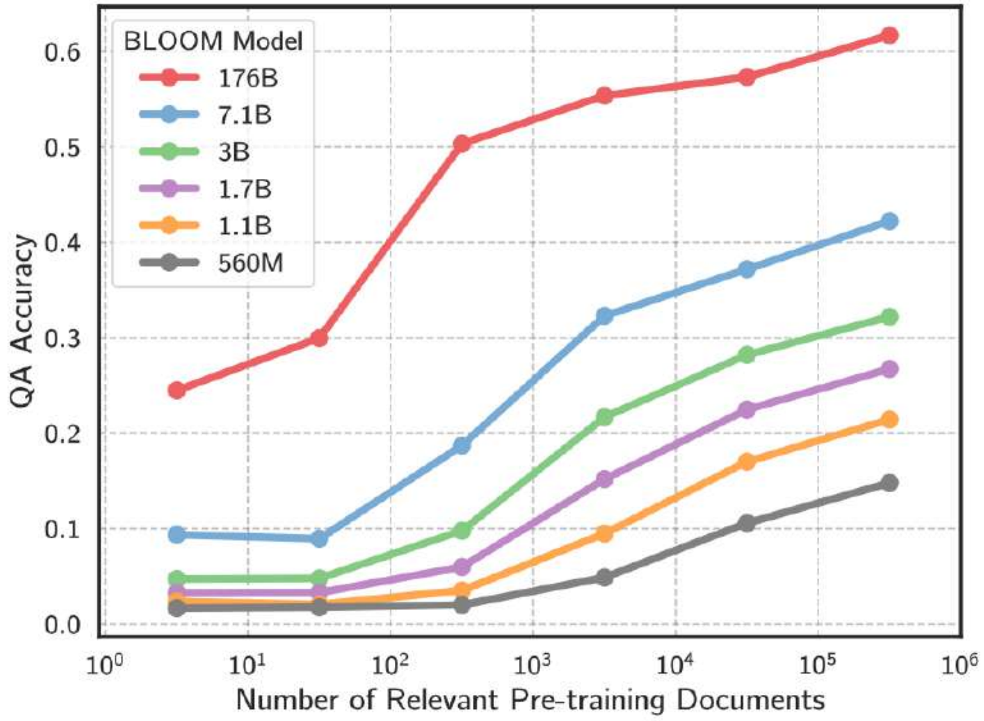


Figure 1. Language models struggle to capture the long-tail of information on the web. Above, we plot accuracy for the BLOOM model family on TriviaQA as a function of how many documents in the model's pre-training data are relevant to each question.

[1] Bender, et al. On the dangers of stochastic parrots: Can language models be too big?. FAccT 2021.

[2] Stephanie Lin et al. TruthfulQA: Measuring How Models Mimic Human Falsehoods. ACL 2022

[3] Lee et al. Deduplicating training data makes language models better. ACL 2022

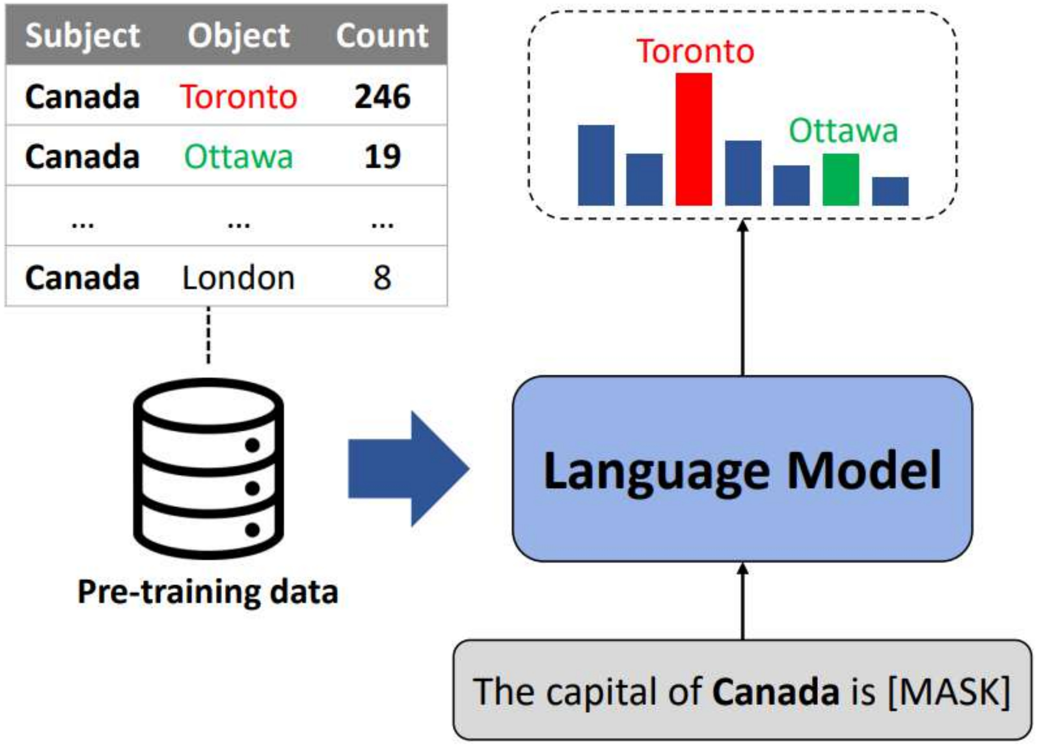
[4] Daniel Martin Katz et al. Gpt-4 passes the bar exam. Arxiv

[5] Yasumasa Onoe et al. Entity cloze by date: What LMs know about unseen entities. NAACL Findings 2022

[6] Karan Singhal et al. Towards Expert-Level Medical Question Answering with Large Language Models. Arxiv

Factuality Bias: Causes

◆ **LMs usually resort to shortcuts to generate the texts depending on position close and co-occurred words rather than understand the knowledge itself.**



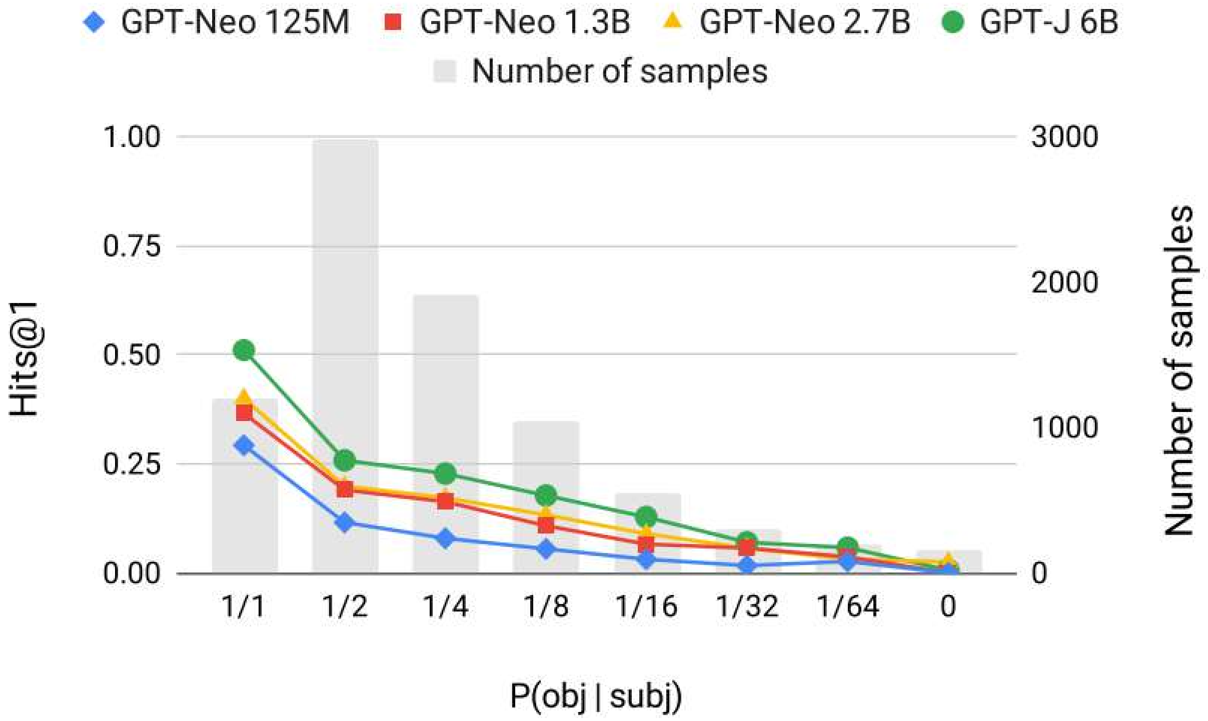


Fig. The correlation between co-occurrence statistics and factual knowledge probing accuracy

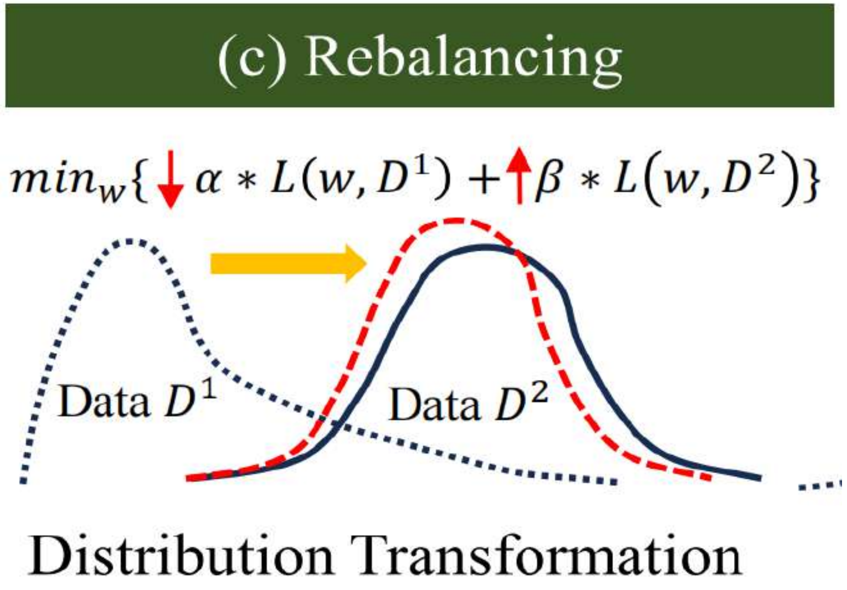
[1] Cheongwoong Kang et al. Impact of Co-occurrence on Factual Knowledge of Large Language Models. EMNLP Findings 2023

[2] Shaobo Li et al. How Pre-trained Language Models Capture Factual Knowledge? A Causal-Inspired Analysis. ACL Findings 2022

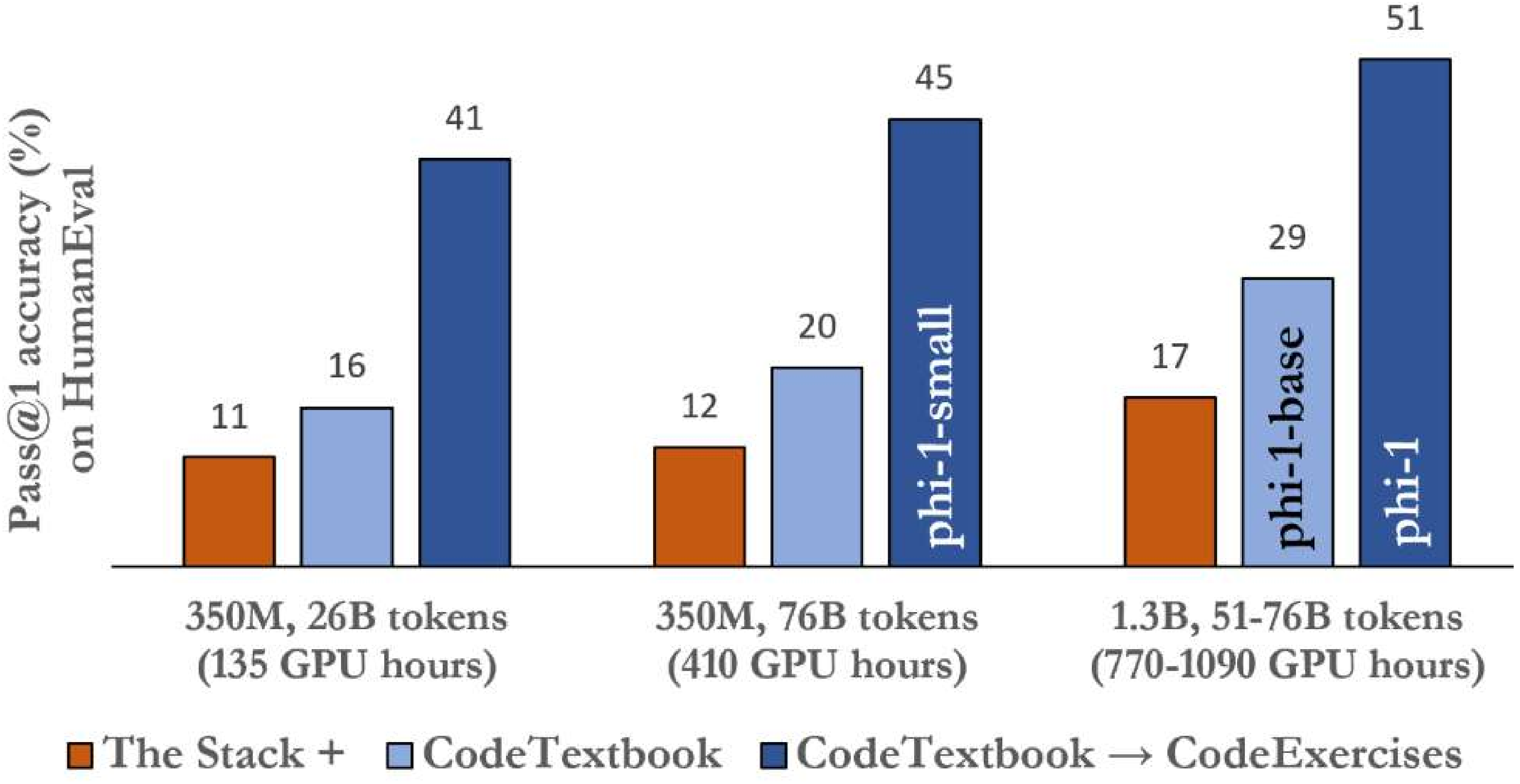
Factuality Bias: Mitigation

Mitigation Strategies

- **High-quality Training Data**
- Retrieval-Augmented Generation
- Decoding-Time Optimization



Significantly smaller high-quality training data size but achieves better performance



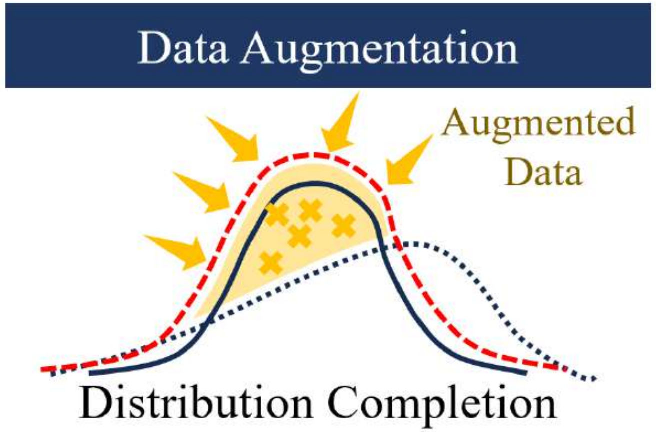
[1] *Suriya Gunasekar et al. Textbooks Are All You Need. Arxiv*

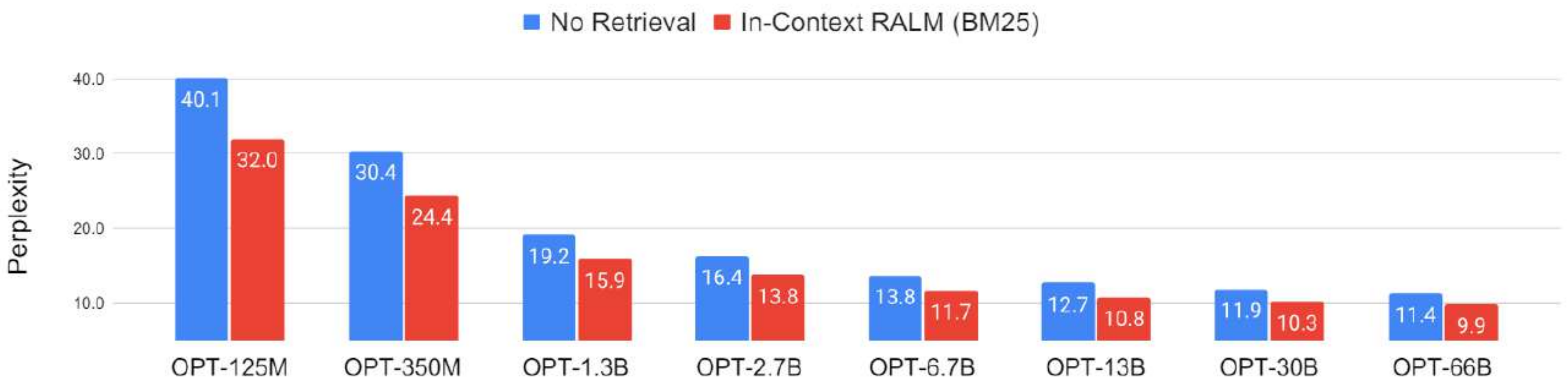
Factuality Bias: Mitigation

Mitigation Strategies

- ➢ High-quality Training Data
- ➢ **Retrieval-Augmented Generation**
- ➢ Decoding-Time Optimization

Provide the retrieved documents in context of LLMs





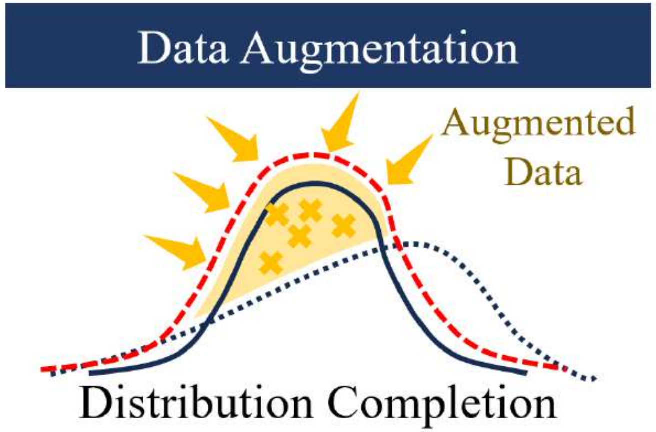
[1] Ori Ram et al. In-Context Retrieval-Augmented Language Models. TACL

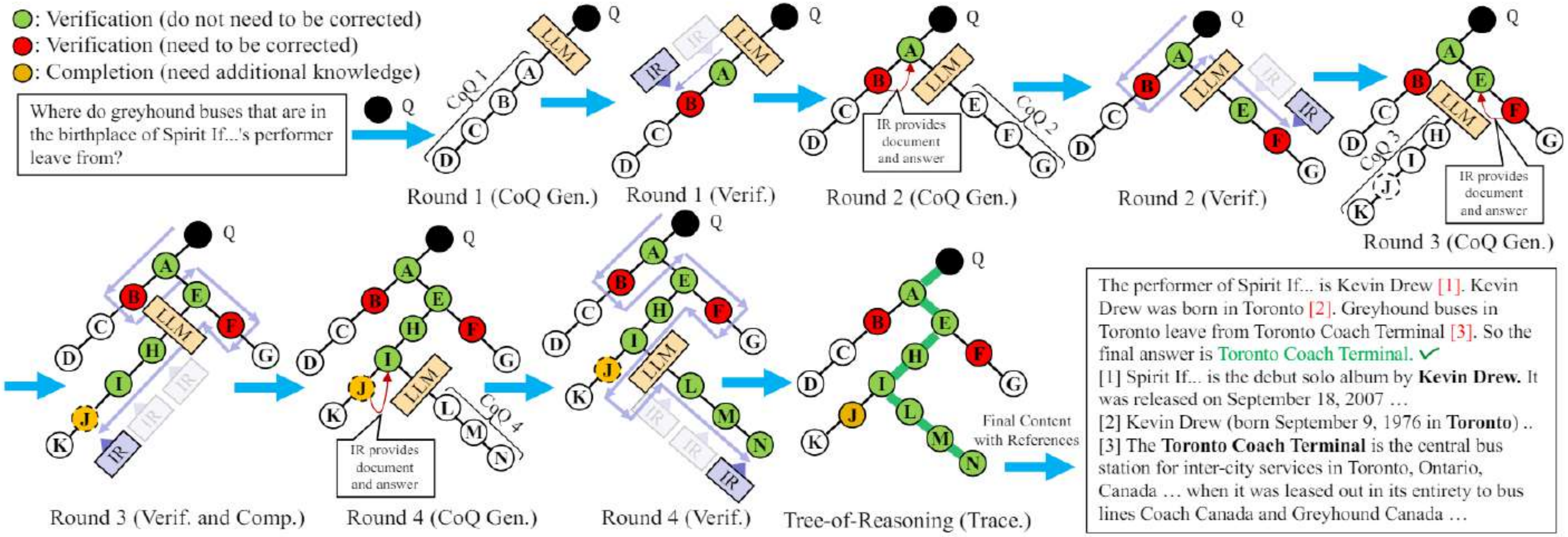
Factuality Bias: Mitigation

Mitigation Strategies

- High-quality Training Data
- **Retrieval-Augmented Generation**
- Decoding-Time Optimization

- **LLM plan a Chain-of-Query (CoQ).**
- **IR interacts with CoQ to perform verification and completion.**
- **IR gives feedback to LLM to help it re-generates a new CoQ.**





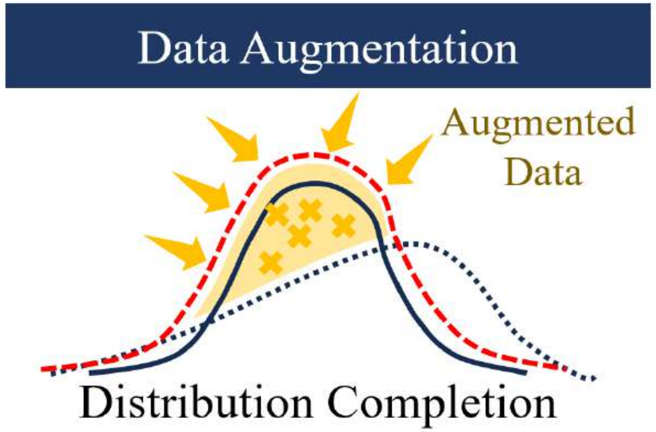
[1] Shicheng Xu et al. Search-in-the-Chain: Interactively Enhancing Large Language Models with Search for Knowledge-intensive Tasks. WWW 2024

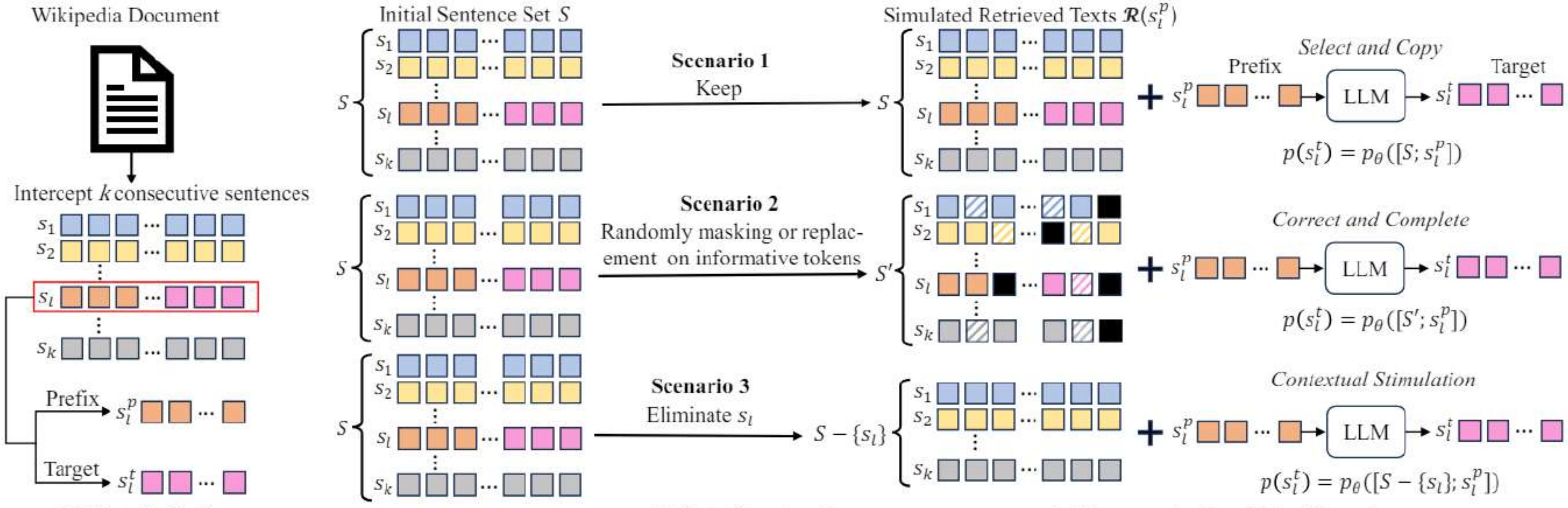
Factuality Bias: Mitigation

Mitigation Strategies

- ➢ High-quality Training Data
- ➢ **Retrieval-Augmented Generation**
- ➢ Decoding-Time Optimization

- **Reassess the role of LLMs in RAG as "Information Refiner".**
- **Propose unsupervised training method to make LLMs learn to perform refinement in RAG.**





(a) Data Collection (b) Data Construction (c) Unsupervised multi-task learning

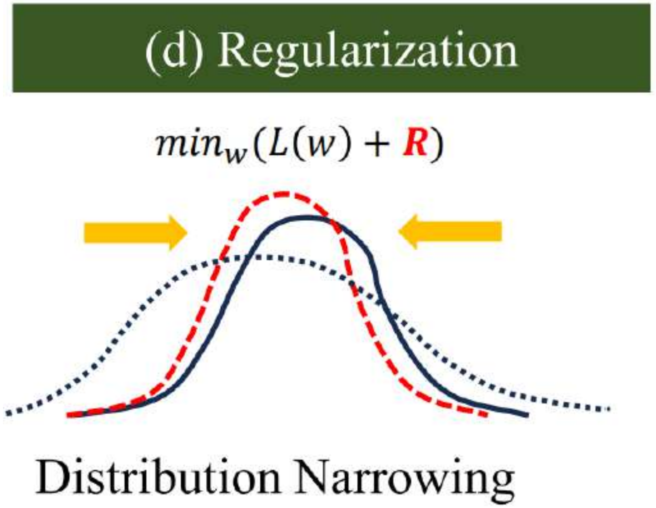
[1] Shicheng Xu et al. Unsupervised Information Refinement Training of Large Language Models for Retrieval-Augmented Generation. ACL 2024

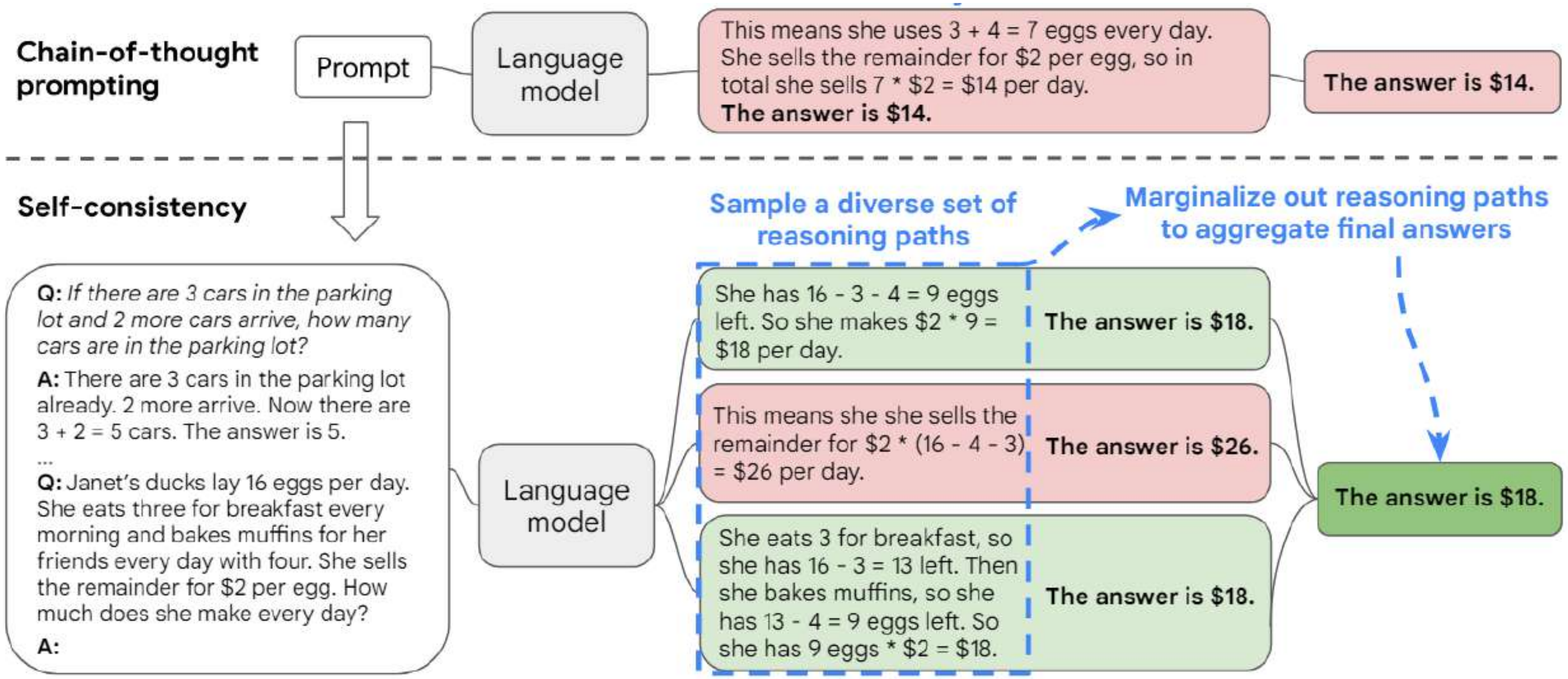
Factuality Bias: Mitigation

Mitigation Strategies

- ➢ High-quality Training Data
- ➢ Retrieval-Augmented Generation
- ➢ **Decoding-Time Optimization**

- **Prompt a language model using chain-of-thought**
- **Generate a diverse set of reasoning paths**
- **Marginalize out reasoning paths to aggregate final answers**



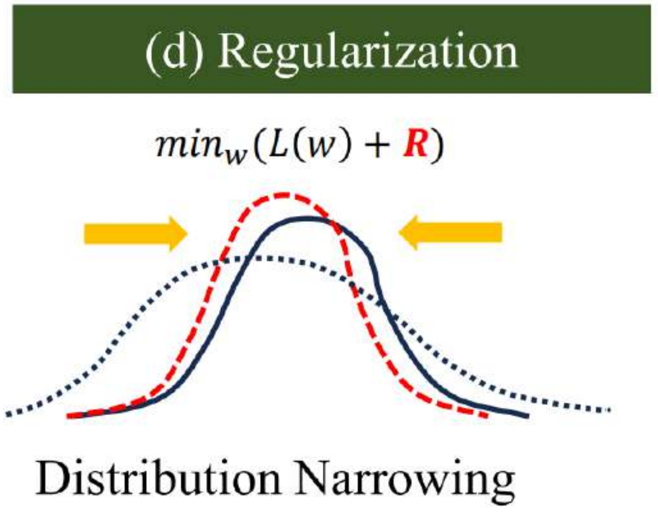


[1] Xuezhi Wang et al. SELF-CONSISTENCY IMPROVES CHAIN OF THOUGHT REASONING IN LANGUAGE MODELS. ICLR 2023

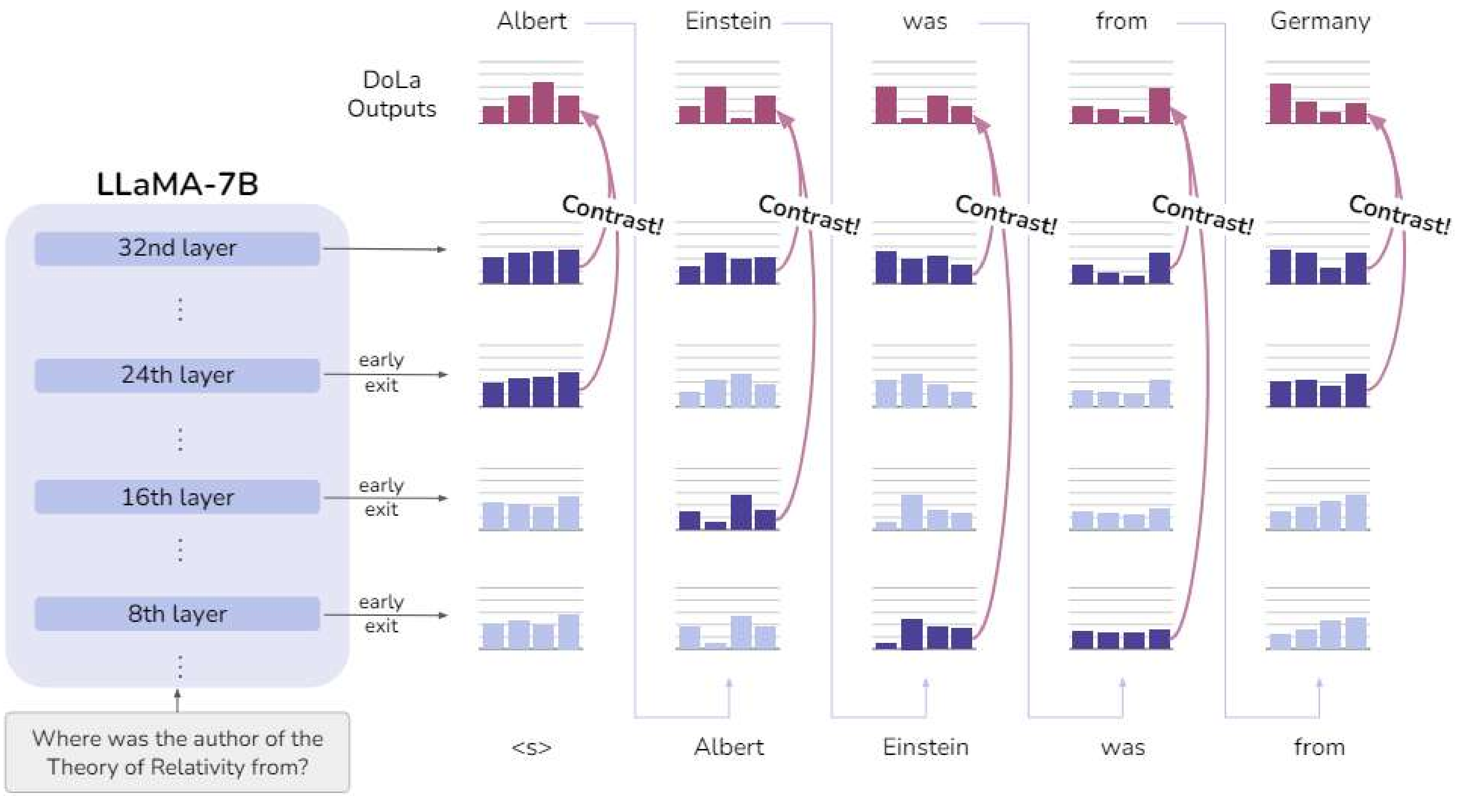
Factuality Bias: Mitigation

Mitigation Strategies

- ➢ High-quality Training Data
- ➢ Retrieval-Augmented Generation
- ➢ **Decoding-Time Optimization**



- **Dynamically select the layer with largest word distribution change**
- **Output the word with largest logits change among layers**



[1] *Yung-Sung Chuang et al. DOLA: DECODING BY CONTRASTING LAYERS IMPROVES FACTUALITY IN LARGE LANGUAGE MODELS. ICLR 2024*

Factuality Bias: Mitigation

Comparison Among Mitigation Strategies

- **High-Quality Training Data**
  - √ Can fundamentally improve the factual consistency of LLMs.
  - × Need training LLMs.
- **Retrieval-Augmented Generation**
  - √ Significantly improve the factual consistency of LLMs at inference time without training.
  - × Need additional knowledge base.
- **Decoding-Time Optimization**
  - √ Improve the factual consistency of LLMs without training and external knowledge.
  - × Limited improvement

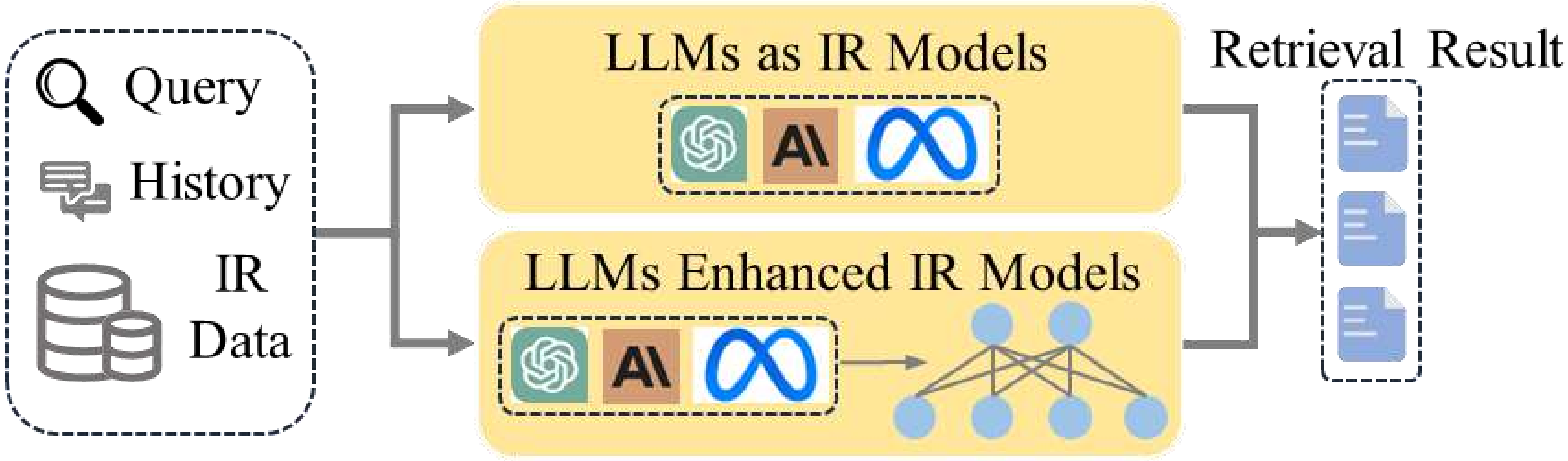
[1] Yung-Sung Chuang et al. DOLA: DECODING BY CONTRASTING LAYERS IMPROVES FACTUALITY IN LARGE LANGUAGE MODELS. ICLR 2024

Bias and Mitigation Strategies

- **Bias in Data Collection**
  - Source Bias
  - Factuality Bias
- **Bias in Model Development**
  - Position Bias
  - Popularity Bias
  - Instruction-Hallucination Bias
  - Context-Hallucination Bias
- **Bias in Result Evaluation**
  - Selection Bias
  - Style Bias
  - Egocentric Bias

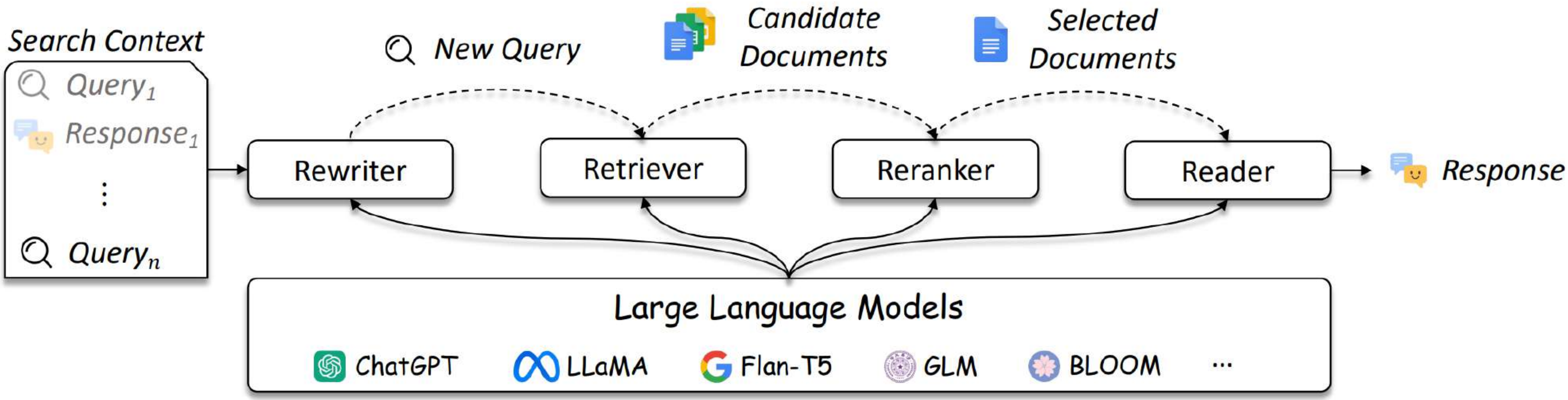
Bias in Medel Development

Incorporating LLMs to Enhance or As IR Models.



- LLMs Enhanced IR Models: LLMs can be used to enhance traditional IR components.
- LLMs as IR Models: LLMs can be used as search agents to perform multiple IR tasks.

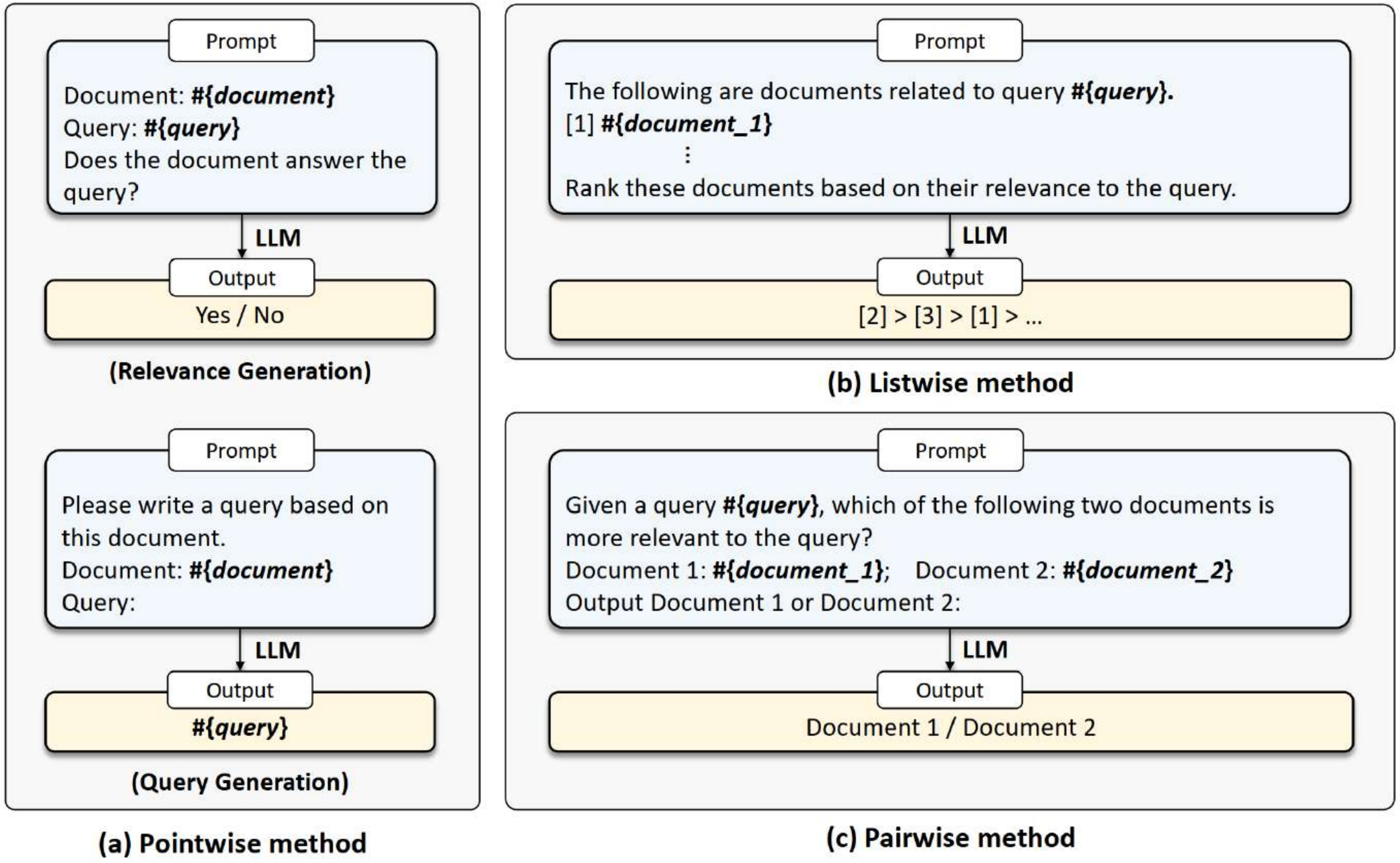
LLMs Enhanced IR Models



LLMs can be used in Query Rewriter, Retriever, Reranker, and Reader.

[1] Yutao Zhu et al. Large Language Models for Information Retrieval: A Survey. arXiv 2023.

LLMs as IR Models



Three types

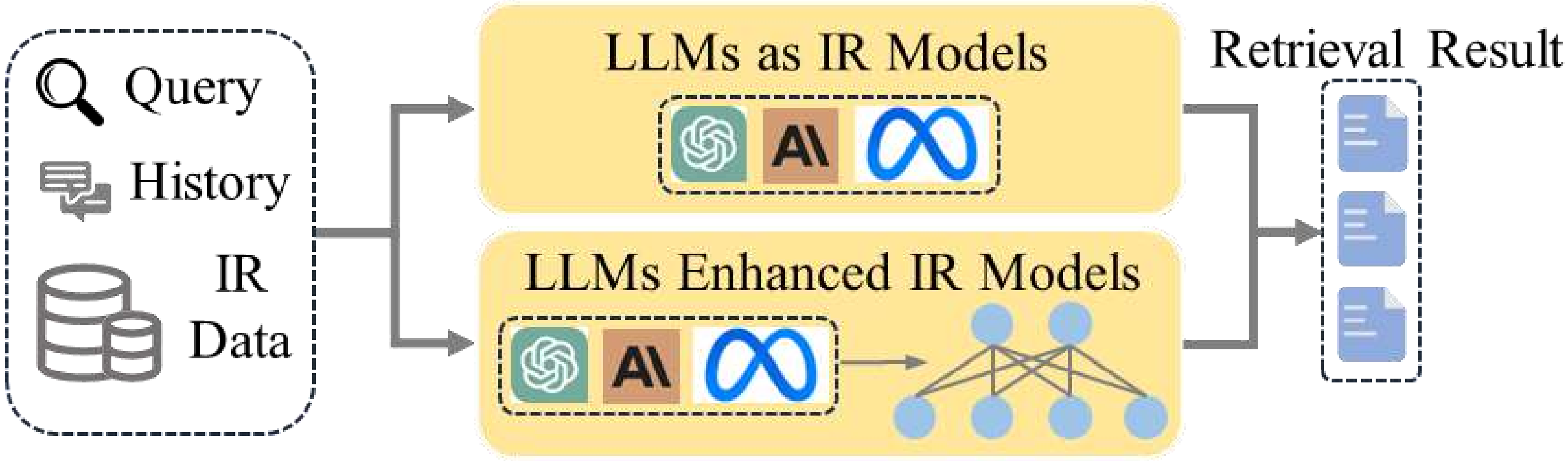
- pointwise methods
- listwise methods
- pairwise methods

[1] Yutao Zhu et al. Large Language Models for Information Retrieval: A Survey. arXiv 2023.

[2] Sunhao Dai et al. Uncovering ChatGPT's Capabilities in Recommender Systems. RecSys 2023.

Bias in Medel Development

Incorporating LLMs to Enhance or As IR Models.



- LLMs Enhanced IR Models: LLMs can be used to enhance traditional IR components.
- LLMs as IR Models: LLMs can be used as search agents to perform multiple IR tasks.

Position Bias! ***Popularity Bias!***

Instruction-Hallucination Bias! ***Context-Hallucination Bias!***

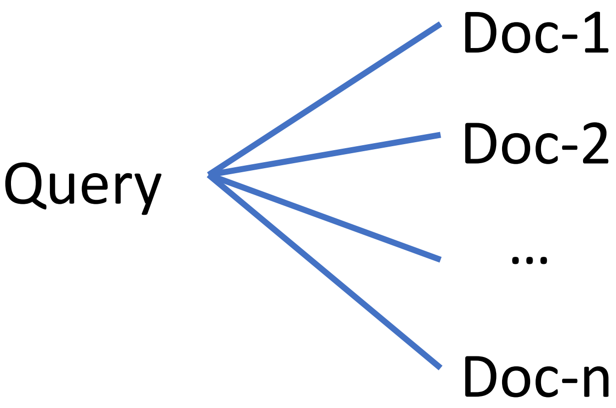
Bias and Mitigation Strategies

- Bias in Data Collection
  - Source Bias
  - Factuality Bias
- Bias in Model Development
  - Position Bias
  - Popularity Bias
  - Instruction-Hallucination Bias
  - Context-Hallucination Bias
- Bias in Result Evaluation
  - Selection Bias
  - Style Bias
  - Egocentric Bias

Position Bias

Definition: LLM-based IR models tend to give preference to documents or items from specific input positions.

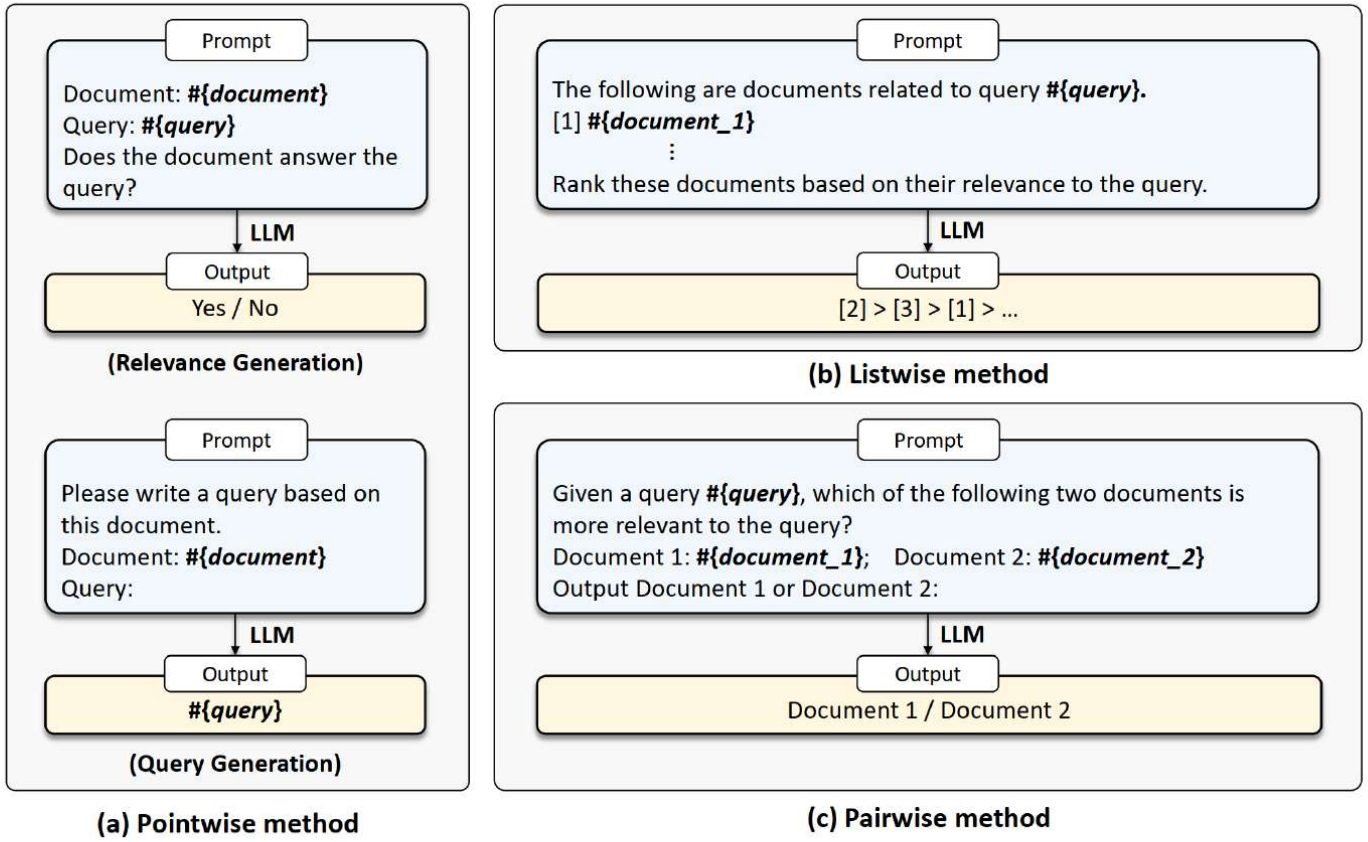
Traditional IR Models



Pointwise Matching

No Position Bias!

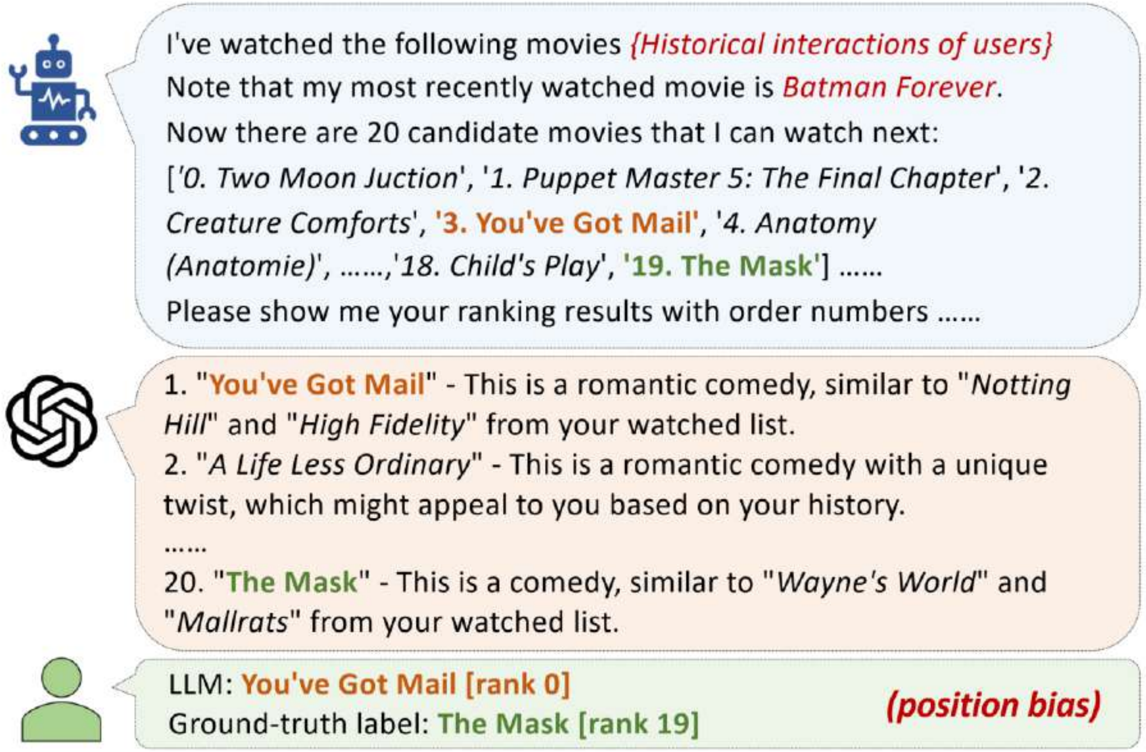
LLMs as IR Models



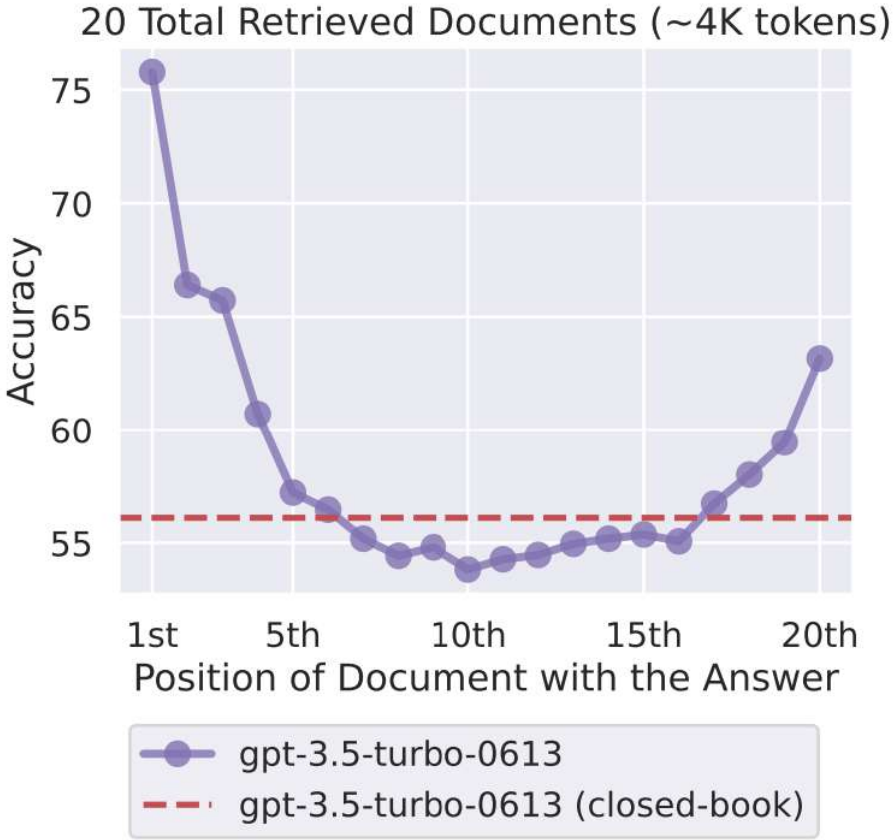
[1] Yutao Zhu et al. Large Language Models for Information Retrieval: A Survey. arXiv 2024.

Position Bias

Definition: LLM-based IR models tend to give preference to documents or items from specific input positions.



Example of Position Bias



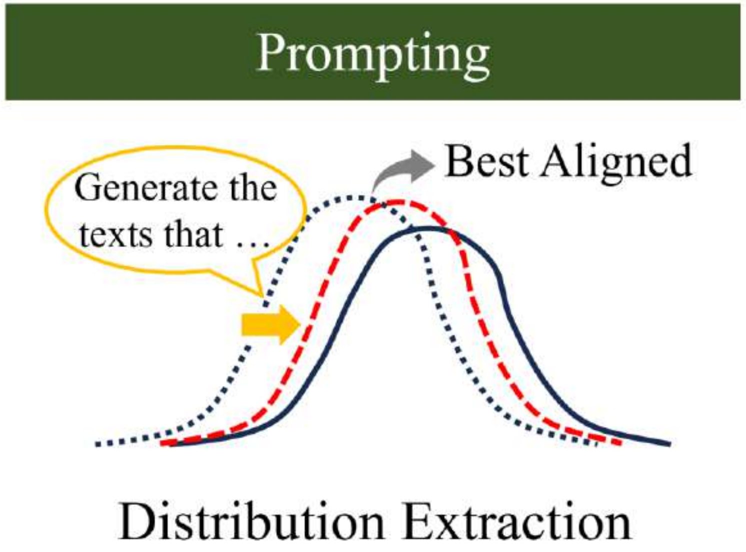
Lost in the Middle

[1] Lanling Xu et al. Prompting Large Language Models for Recommender Systems: A Comprehensive Framework and Empirical Analysis. arXiv 2024.
[2] Nelson F. Liu et al. Lost in the Middle: How Language Models Use Long Contexts. TACL 2024.

Position Bias

Mitigation Strategies

- **Prompting**



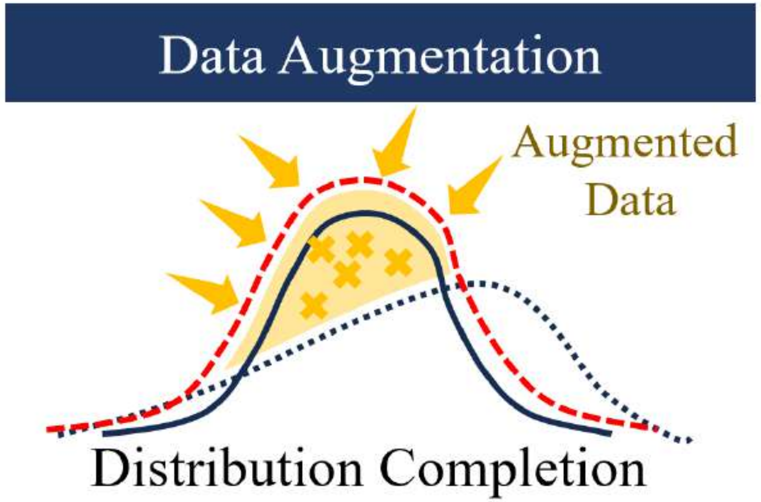
Instruction:

The candidate document list provided to you is presented in a random order. The order of the documents does not reflect any inherent ranking or relevance. Please evaluate and rank the documents based solely on their content and relevance to the given query, without considering their initial position in the list.

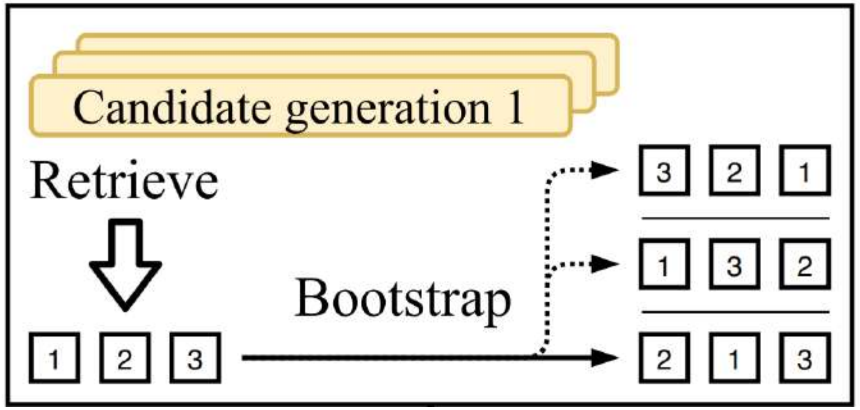
Position Bias

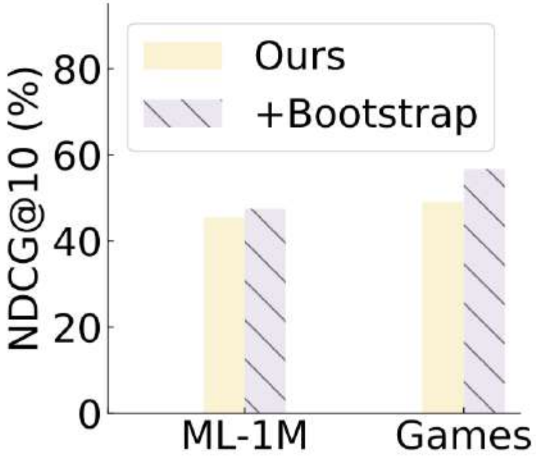
Mitigation Strategies

- Prompting
- Data Augmentation
  - Bootstrapping



Retrieving candidates & Bootstrapping to reduce position bias





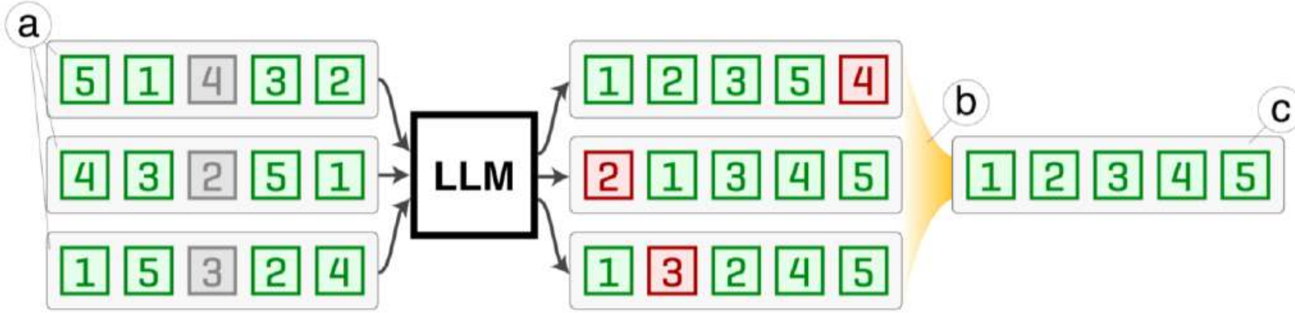
Simple bootstrapping idea works!

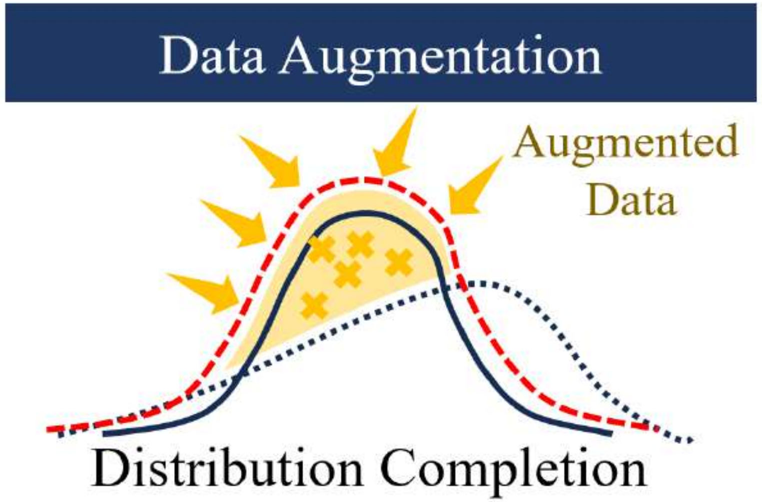
[1] Yupeng Hou et al. Large Language Models are Zero-Shot Rankers for Recommender Systems. ECIR 2024.

Position Bias

Mitigation Strategies

- Prompting
- **Data Augmentation**
  - Bootstrapping
  - **Permutation Self-Consistency**





Theoretical Guarantees

Given that at least one possibly nonrandom pair of items is always concordant, it yields a consistent estimator for the true ranking.

Method	MATH	WORD	GSM8K	DL19	DL20
GPT-3.5 (Orig.)	64.0	85.9	82.1	68.00	62.08
GPT-3.5 (Borda)	74.6	87.9	88.1	70.09	62.54
GPT-3.5 (Our PSC)	**75.2**	**88.1**	**88.4**	**70.77**	**62.70**
GPT-4 (Orig.)	83.5	89.9	88.4	75.00	70.36
GPT-4 (Borda)	89.2	91.5	90.4	75.23	70.62
GPT-4 (Our PSC)	**89.6**	**92.0**	**90.5**	**75.66**	**71.00**

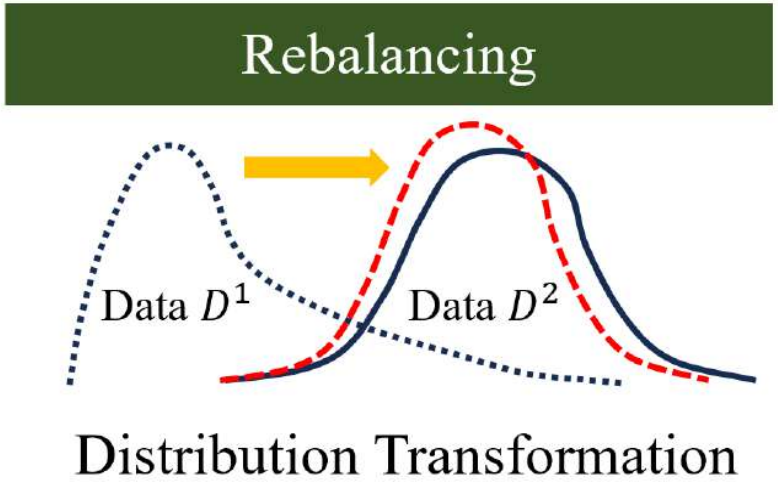
Bootstrapping (Borda count) vs. permutation self-consistency

[1] Raphael Tang et al. Found in the Middle: Permutation Self-Consistency Improves Listwise Ranking in Large Language Models. NAACL 2024.

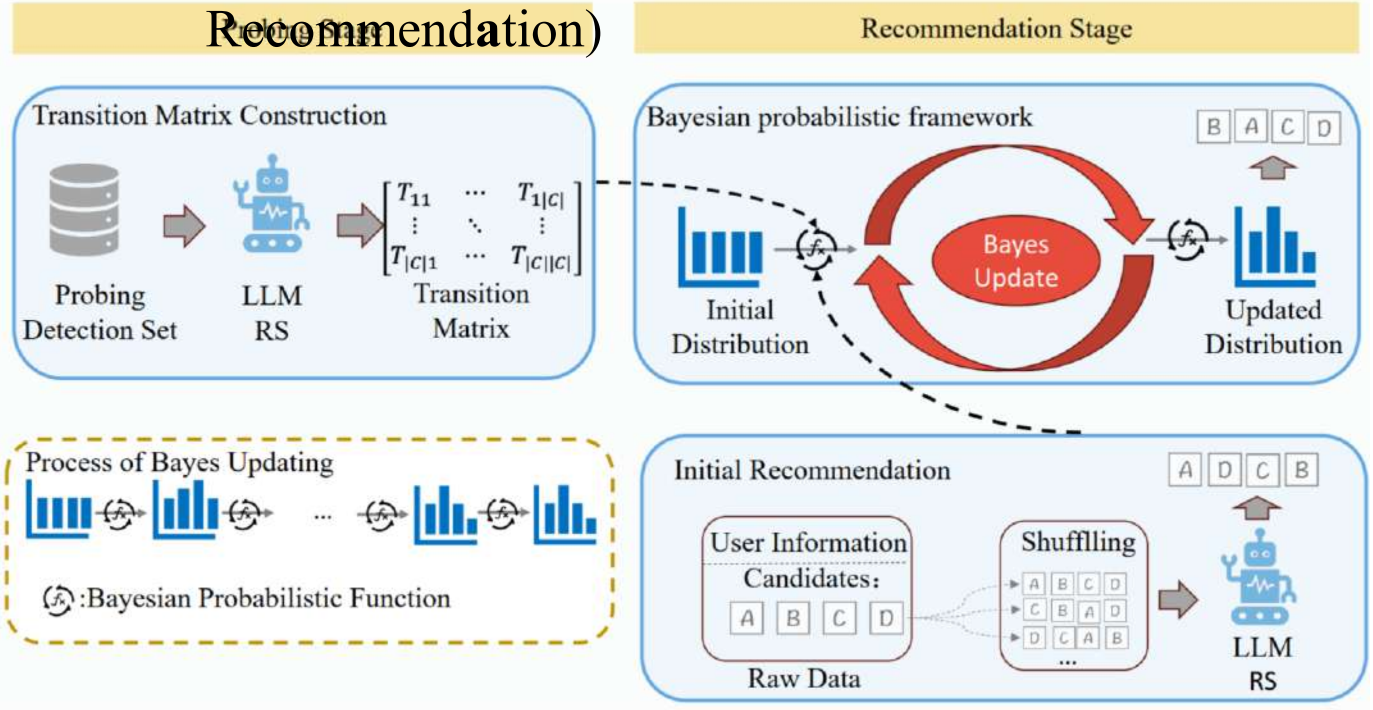
Position Bias

Mitigation Strategies

- ➢ Prompting
- ➢ Data Augmentation
  - Bootstrapping
  - Permutation Self-Consistency
- ➢ **Rebalancing**



STELLA (**St**able **LL**M for Recommend**a**tion)



	Raw Output	Bootstrapping	STELLA
Book	$0.2915_{\pm 0.0798}$	0.2647	**0.3235**
Movie	$0.2740_{\pm 0.0593}$	0.2537	**0.2976**
Music	$0.2500_{\pm 0.0300}$	0.2650	**0.3000**
News	$0.2610_{\pm 0.0219}$	0.2341	**0.2732**

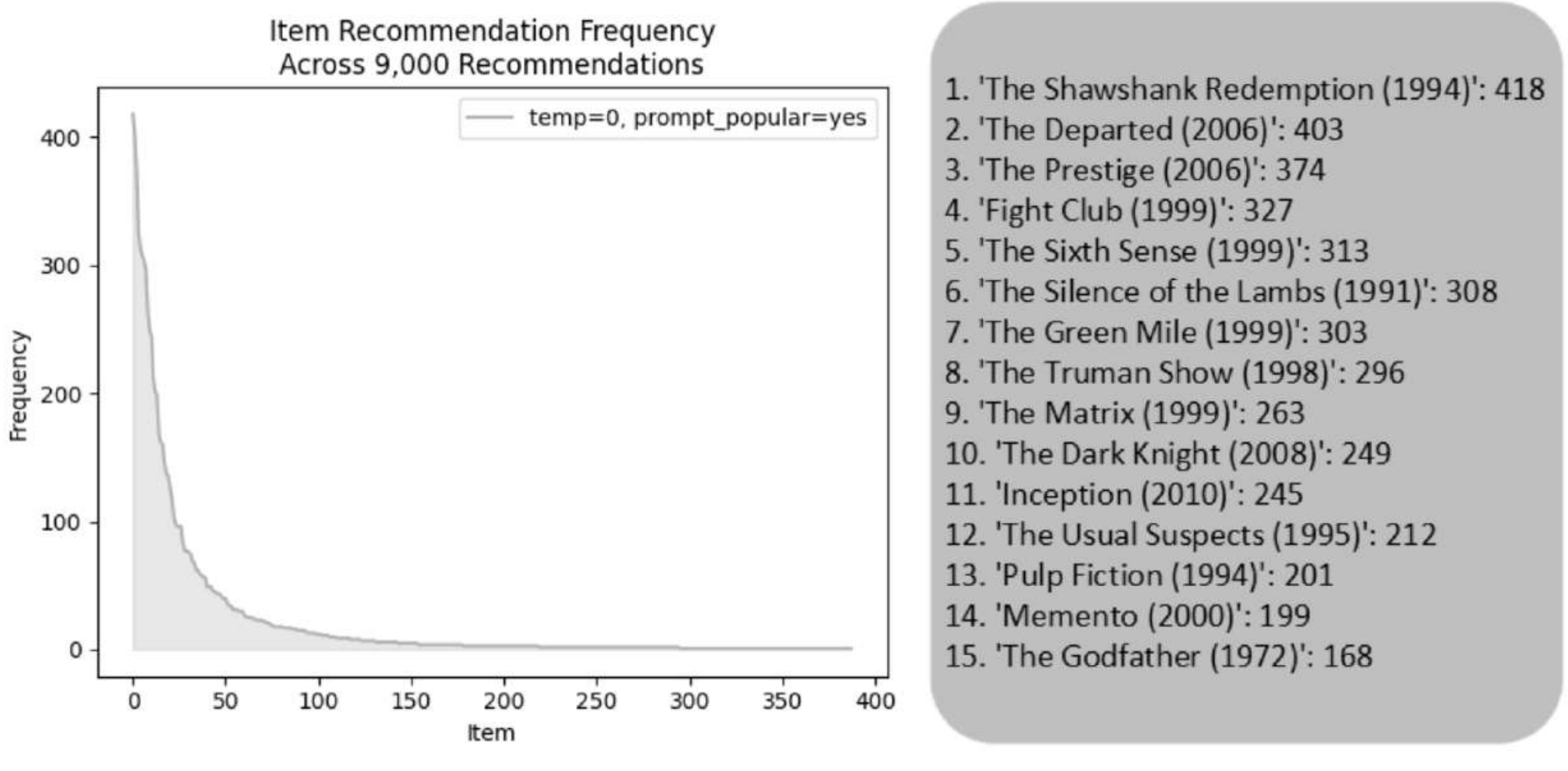
[1] Tianhui Ma et al. Large Language Models are Not Stable Recommender Systems. arXiv 2023.

Bias and Mitigation Strategies

- **Bias in Data Collection**
  - Source Bias
  - Factuality Bias
- **Bias in Model Development**
  - Position Bias
  - Popularity Bias
  - Instruction-Hallucination Bias
  - Context-Hallucination Bias
- **Bias in Result Evaluation**
  - Selection Bias
  - Style Bias
  - Egocentric Bias

Popularity Bias

Definition: LLM-based IR models tend to prioritize candidate documents or items with high popularity levels.



(a) (b)

The list of most frequently recommended items coincides with the IMDB top 250 movies list.

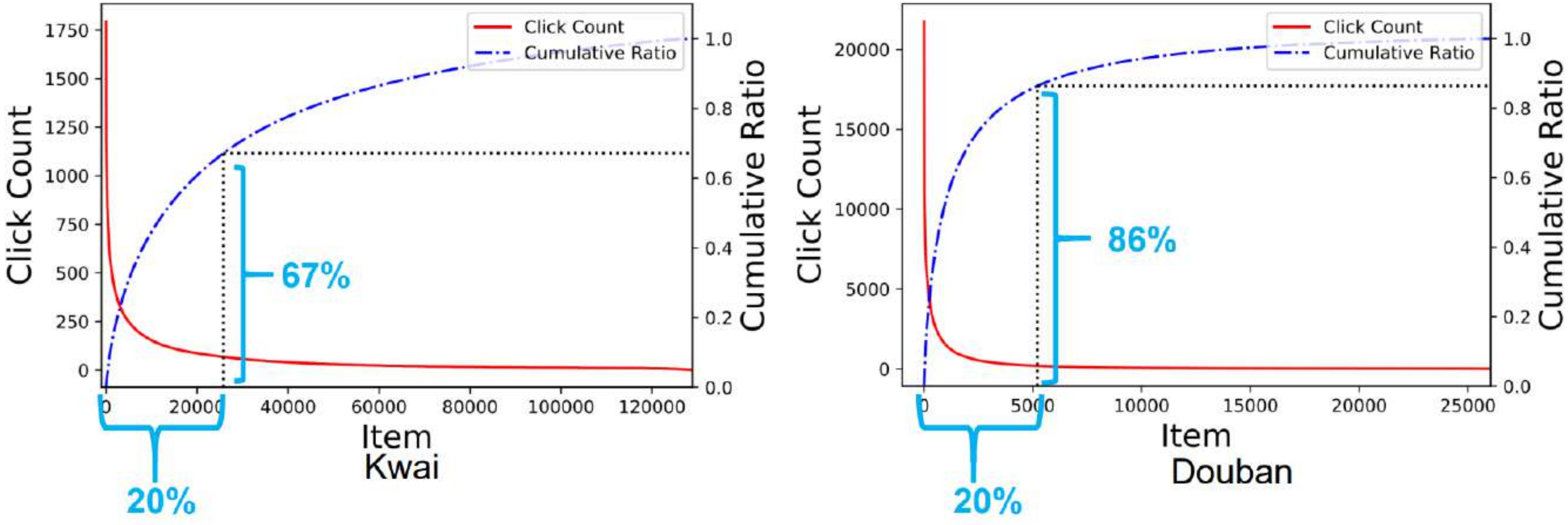
[1] Kyle Dylan Spurlock et at. ChatGPT for Conversational Recommendation: Refining Recommendations by Reprompting with Feedback. arXiv 2024

Popularity Bias

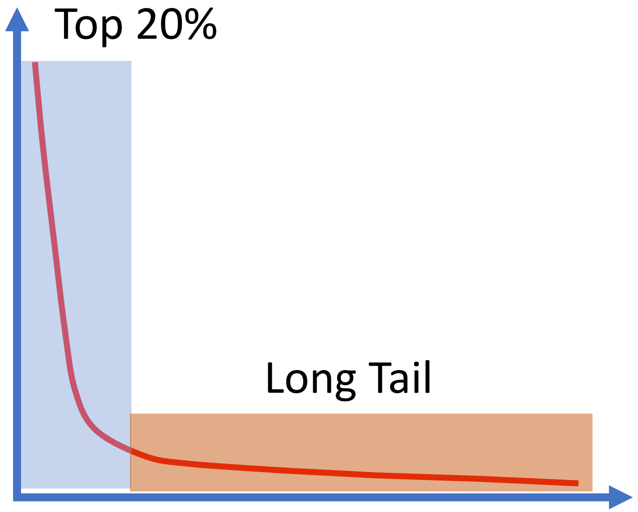
Cause of Popularity Bias

- Popularity Bias in Pre-LLM Era: Long-tail phenomenon in IR training data
- Popularity Bias in LLM Era: Long-tailed Pre-training corpora (and fine-tuning IR data)

Long-tailed IR training data



Long-tailed Pre-training corpora



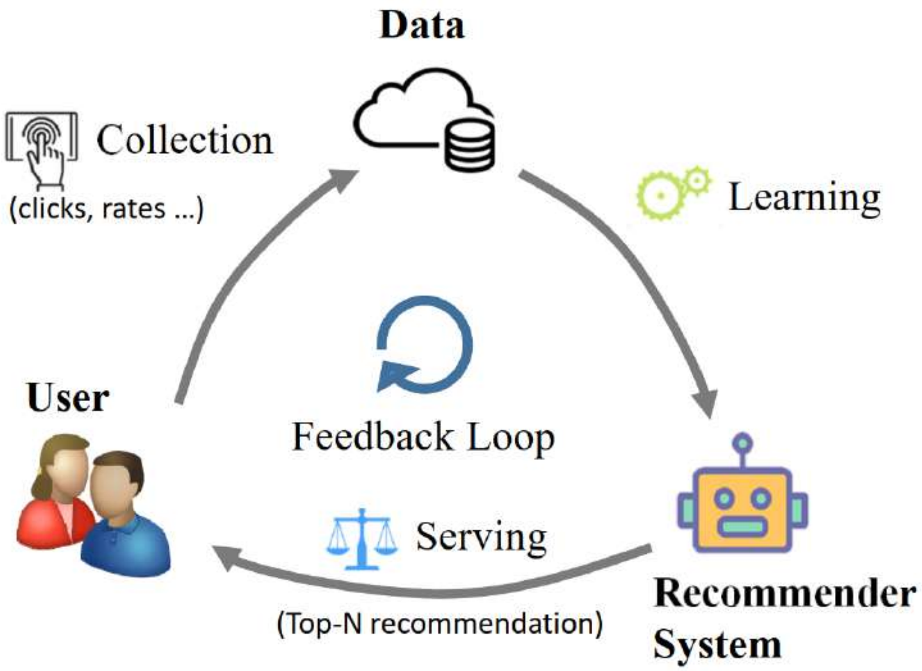
Few popular items which take up the majority of rating interactions

[1] Jiawei Chen et at. Bias and Debias in Recommender System: A Survey and Future Directions. TOIS 2023.

Popularity Bias

Impacts of Popularity Bias

- User-side: Decreases the level of personalization and hurts the serendipity
- Item-side: Decreases the fairness of the recommendation results
- Matthew effect under the feedback loop

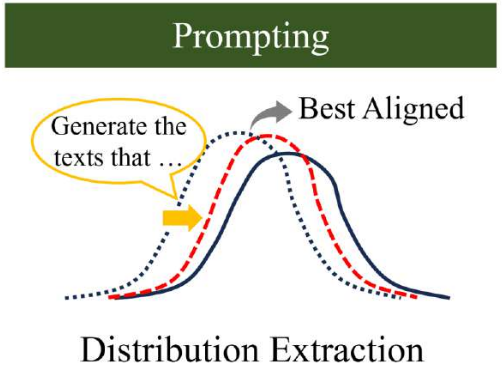


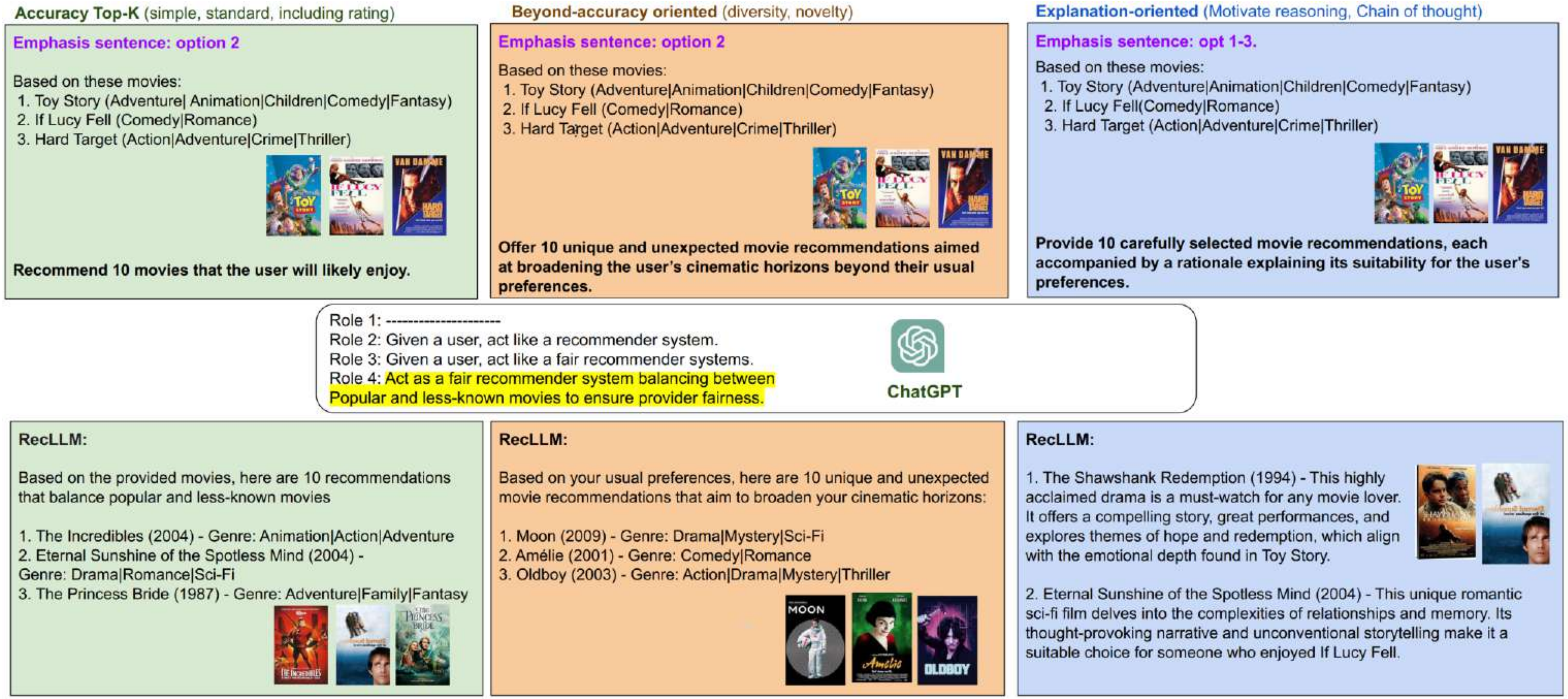
[1] Jiawei Chen et at. Bias and Debias in Recommender System: A Survey and Future Directions. TOIS 2023.

Popularity Bias

Mitigation Strategies

➢ **Prompting**





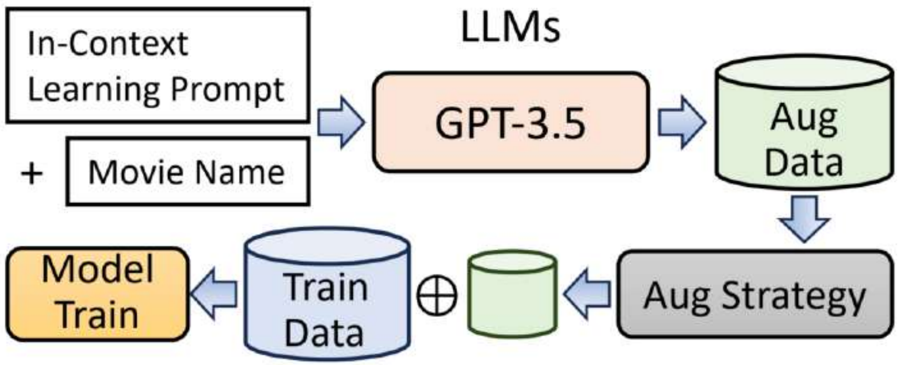
"Focus on fair recommendations, balancing popular and lesser-known movies"

[1] Yashar Deldjoo. Understanding Biases in ChatGPT-based Recommender Systems: Provider Fairness, Temporal Stability, and Recency. arXiv 2024

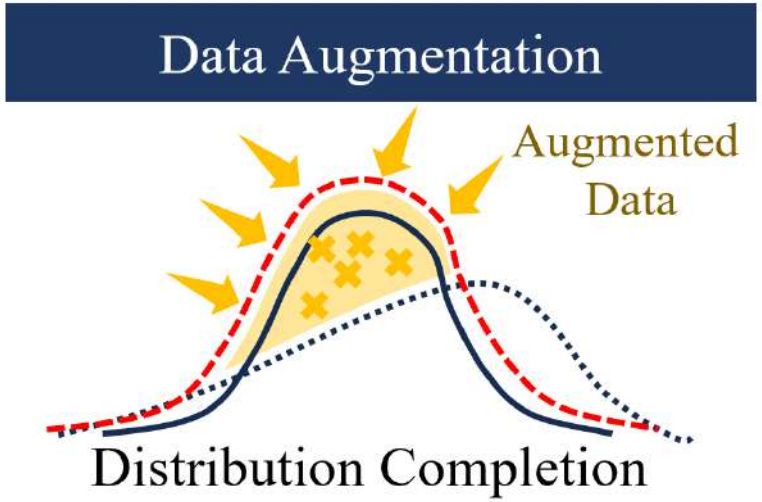
Popularity Bias

Mitigation Strategies

- Prompting
- **Data Augmentation**



Data Augmentation Pipeline



■ OnceAug

- Adding all synthetic dialogues to the training data, evenly increasing the exposure of items in the corpus

■ PopNudge

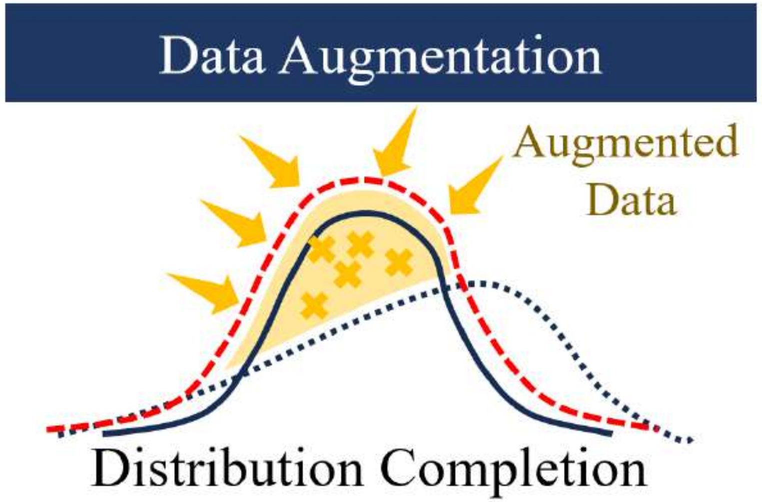
- Augments training batches with dialogues recommending similar but less popular items

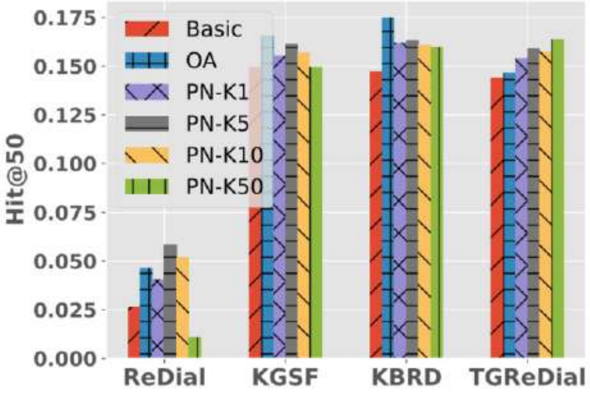
[1] Xi Wang et at. Improving Conversational Recommendation Systems via Bias Analysis and Language-Model-Enhanced Data Augmentation. Findings of EMNLP 2024.

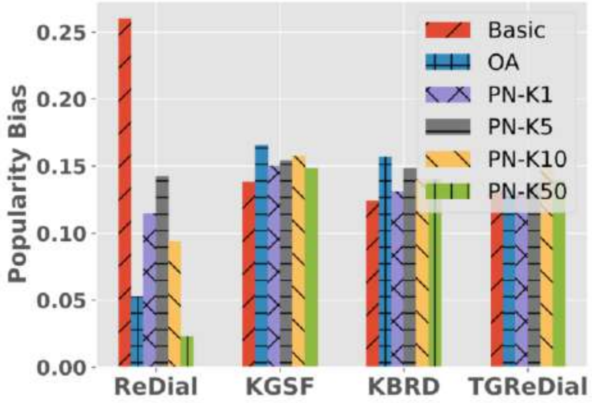
Popularity Bias

Mitigation Strategies

- Prompting
- **Data Augmentation**

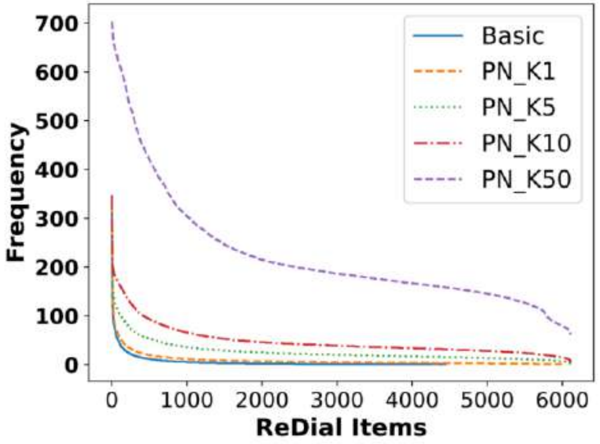


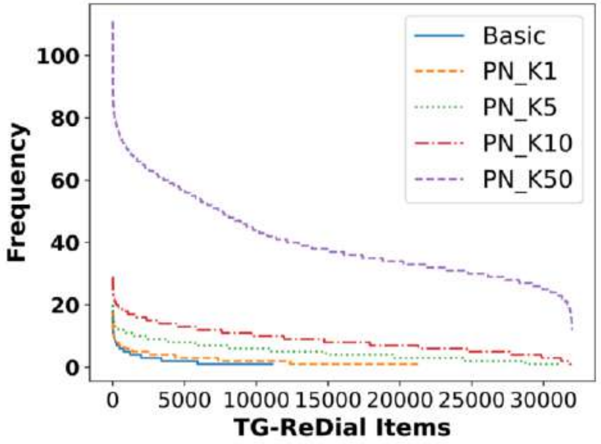




OA: Once Aug **PN: PopNudge**

Improve performance and mitigating bias





Mitigated Long-tail effect after applying PopNudge

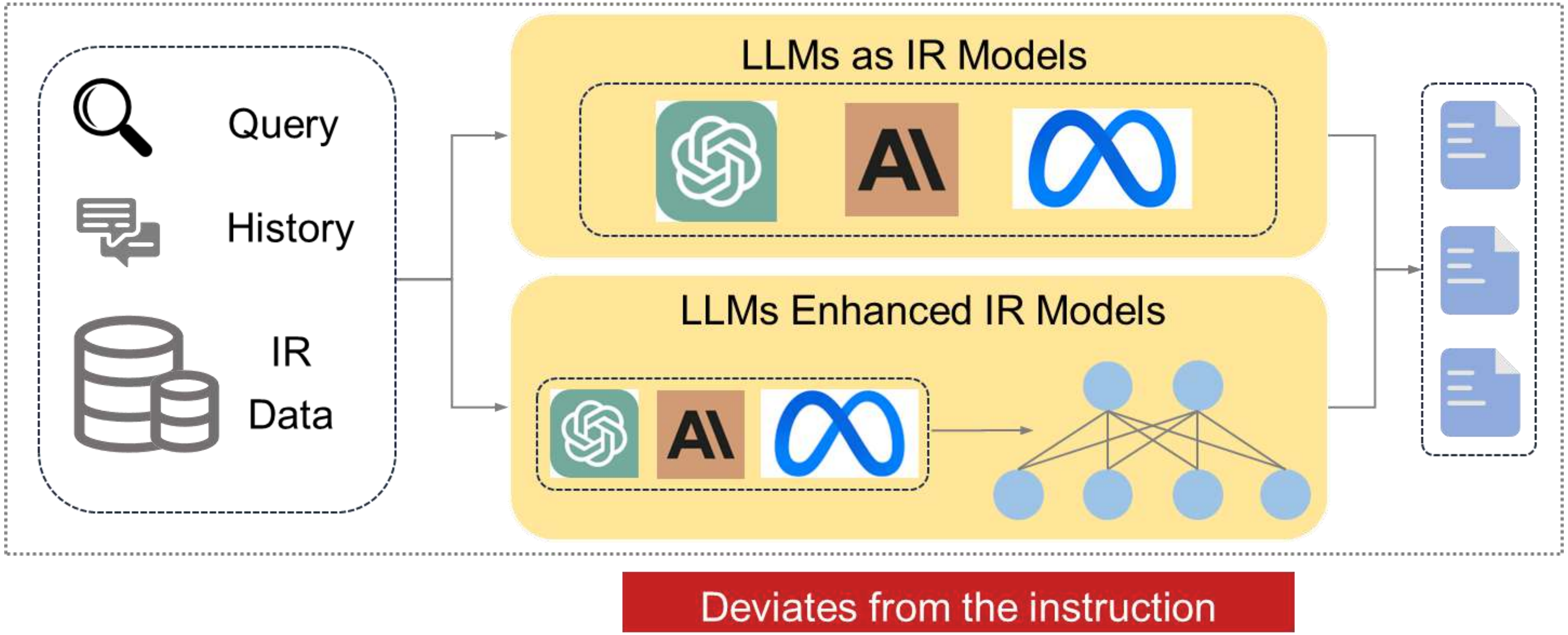
[1] Xi Wang et at. Improving Conversational Recommendation Systems via Bias Analysis and Language-Model-Enhanced Data Augmentation. Findings of EMNLP 2024.

Bias and Mitigation Strategies

- Bias in Data Collection
  - Source Bias
  - Factuality Bias
- Bias in Model Development
  - Position Bias
  - Popularity Bias
  - Instruction-Hallucination Bias
  - Context-Hallucination Bias
- Bias in Result Evaluation
  - Selection Bias
  - Style Bias
  - Egocentric Bias

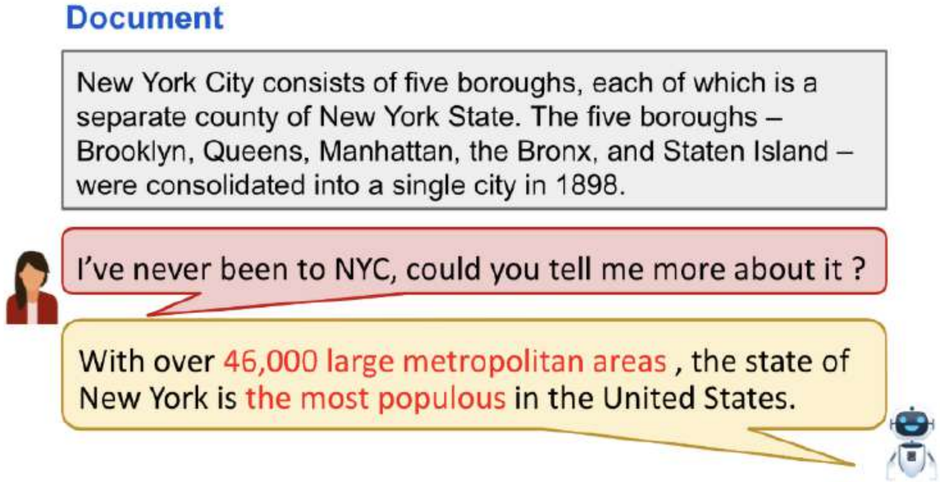
Instruction-Hallucination Bias

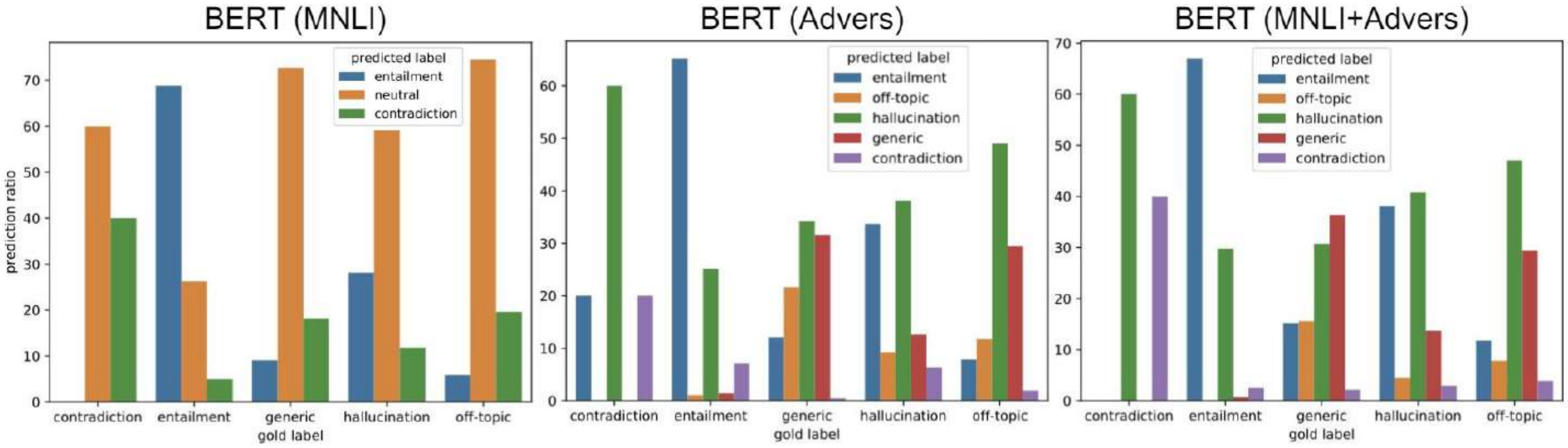
Definition: Content generated by LLM-based IR models may deviate from the instructions provided by users.



Instruction-Hallucination Bias

◆ LLMs often struggle to adhere fully to users' instructions in **dialogue generation**.





[1] *Nouha Dziri et al. Evaluating Groundedness in Dialogue Systems: The BEGIN Benchmark. Arxiv*

Instruction-Hallucination Bias

◆ LLMs often struggle to adhere fully to users' instructions in **summarization and question-answering**.

Source. The world's oldest person has died a few weeks after celebrating her 117th birthday. Born on March 5, 1898, the great-grandmother had lived through two world wars, the invention of the television and the first successful powered aeroplane flight by the wright brothers...
Output sentence. The world 's oldest person has died on March 5, 1898.

An example of unfaithful output (red texts).

PTGEN	Leeds United fought back from 2-0 down to beat Huddersfield town in the first round of the EFL cup. (**Q:** *What team did Leeds United beat in the first round of the EFL cup?*, **A:** *Huddersfield town*)
TCONVS2S	A coal mine in South Yorkshire has collapsed as a result of the loss of a coal mine. (**Q:** *What type of mine has collapsed?*, **A:** *Coal*)
TRANS2S	Star Wars actor James Davis said he was "locked in a caravan" and had his caravan stolen during a break-in. (**Q:** *Who said he was locked in a caravan?*, **A:** *Davis*)

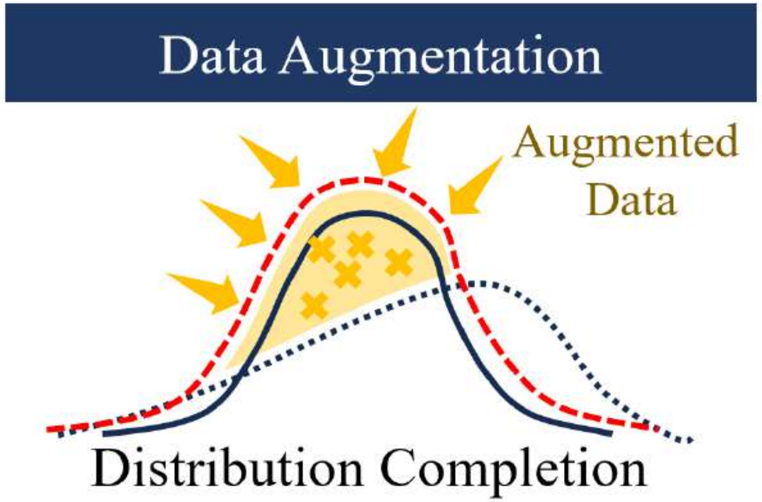
Instruction-hallucinations (pink text) in Q&A output.

[1] Joshua Maynez et al. On Faithfulness and Factuality in Abstractive Summarization. ACL 2020
[2] FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. ACL 2020

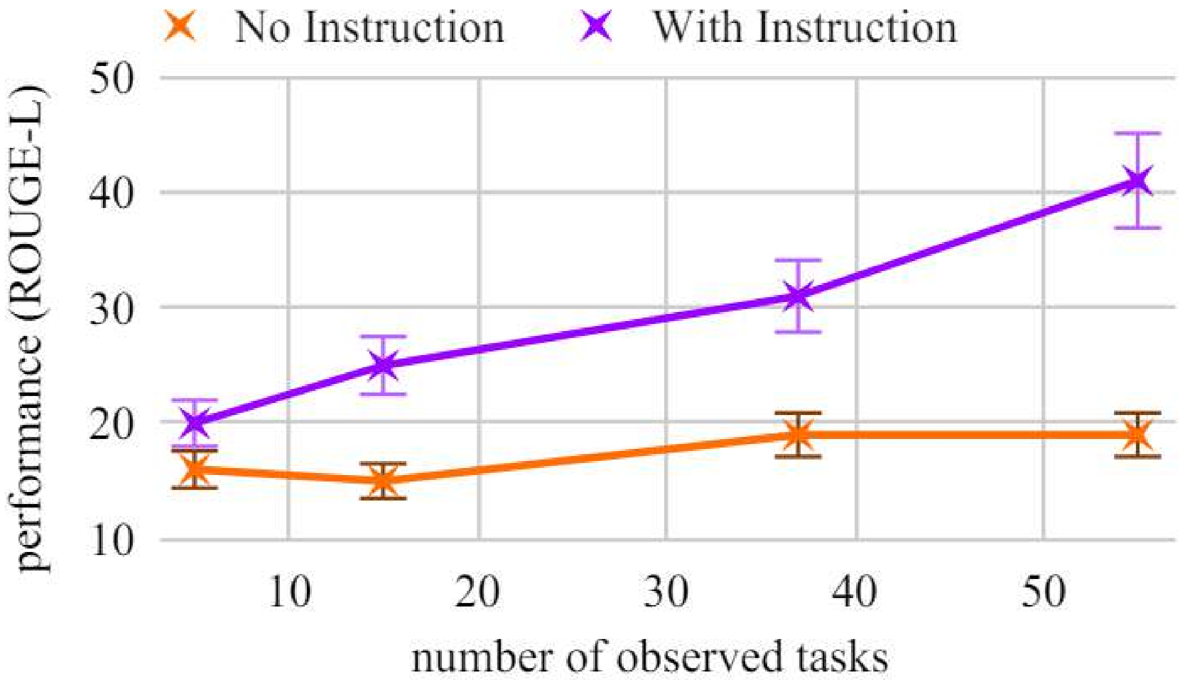
Instruction-Hallucination: Mitigation

Mitigation Strategies

- **Data Augmentation**
- Regularization



NATURAL INSTRUCTIONS: A dataset of 61 distinct tasks, their human-authored instructions and 193k task instances obtained from crowdsourcing.



More instruction tuning tasks bring better performance.

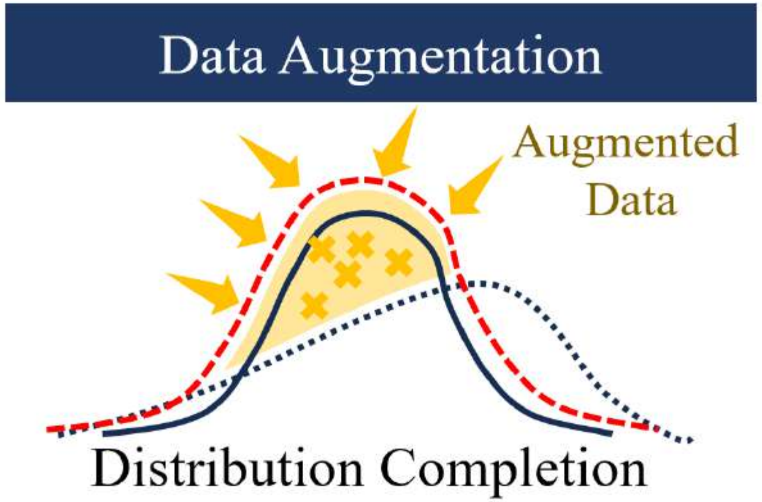
[1] *Swaroop Mishra et al. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. ACL 2022*

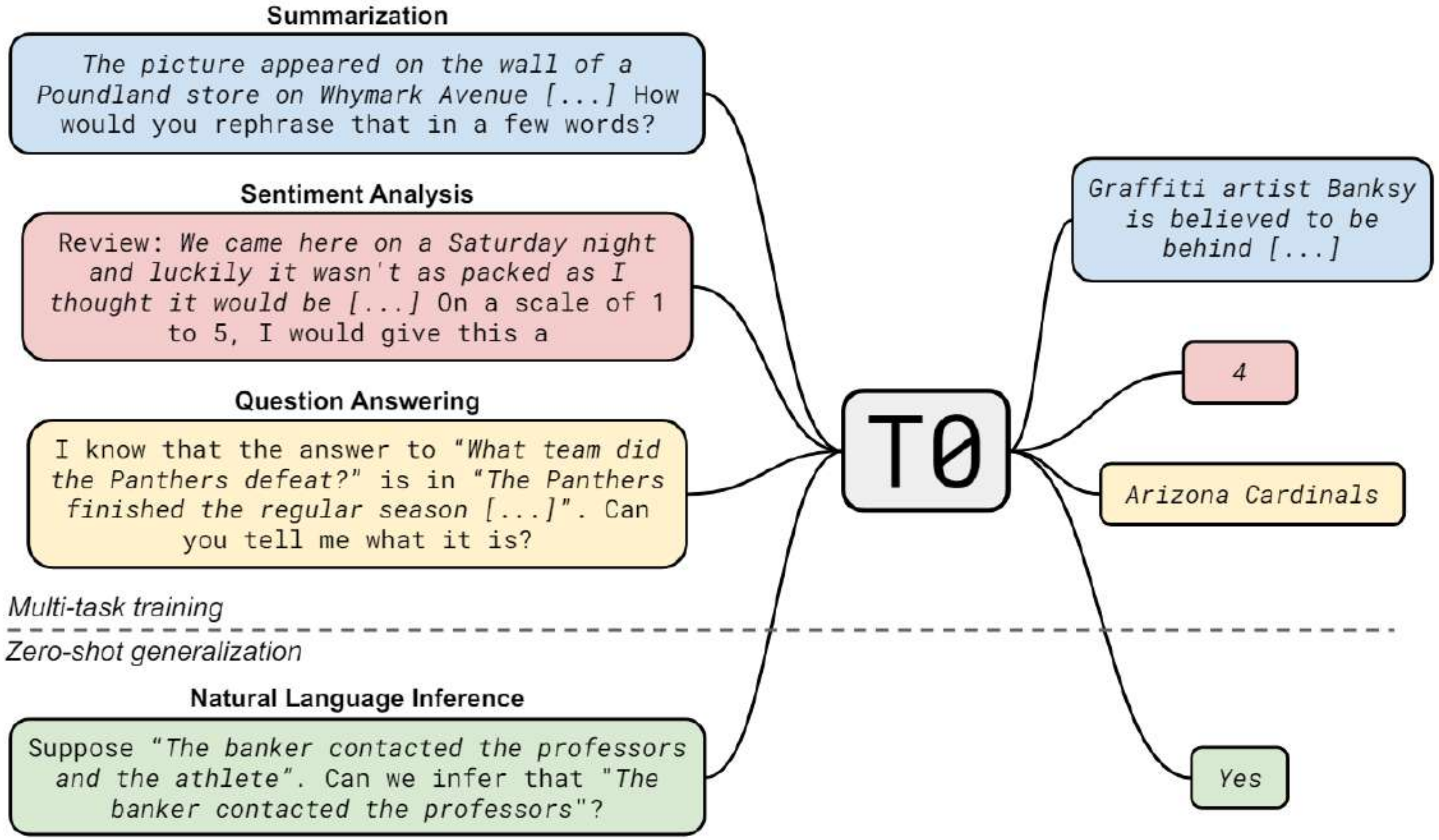
Instruction-Hallucination: Mitigation

Mitigation Strategies

- ➢ **Data Augmentation**
- ➢ Regularization

A large set of supervised datasets, each with multiple prompts with diverse wording.



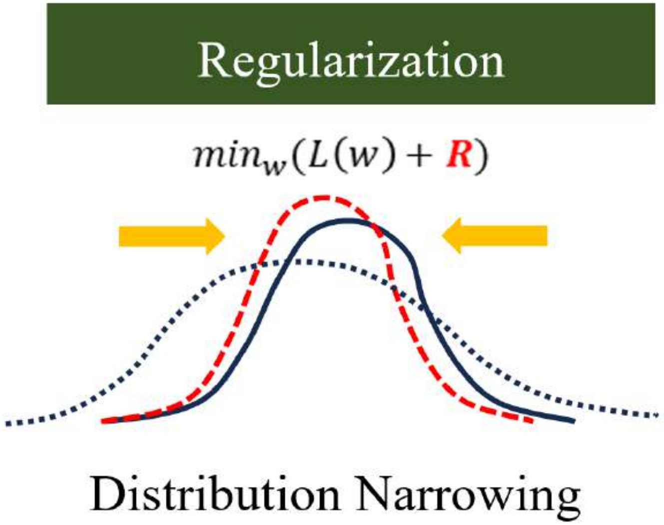


[1] *Victor Sanh et al. Multitask prompted training enables zero-shot task generalization. ICLR 2022*

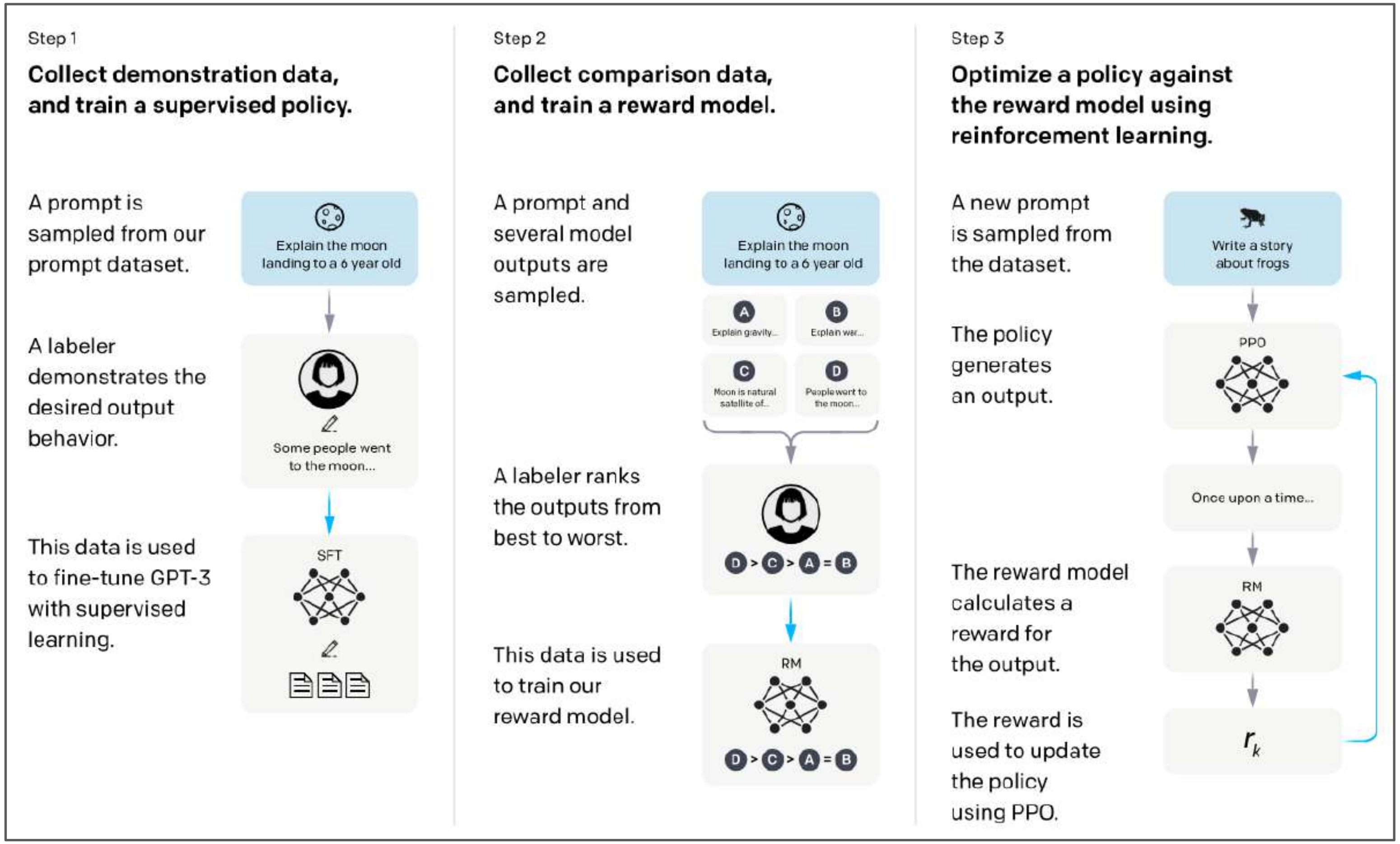
Instruction-Hallucination: Mitigation

Mitigation Strategies

- Data Augmentation
- **Regularization**



Learning from Feedback

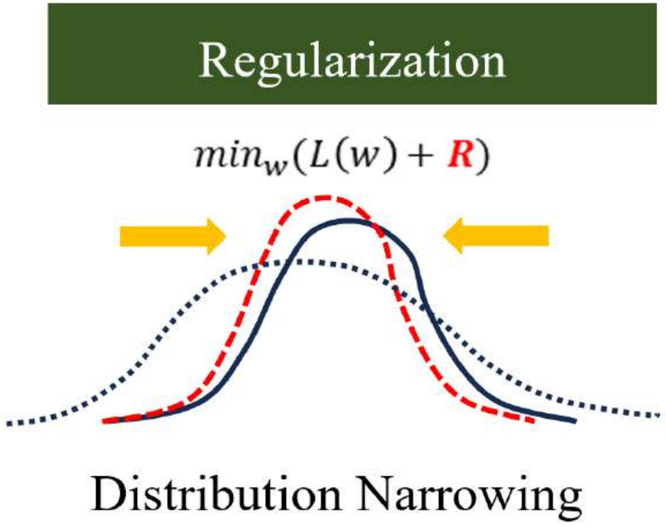


[1] Long Ouyang et al. Training language models to follow instructions with human feedback. NeurIPS 2022

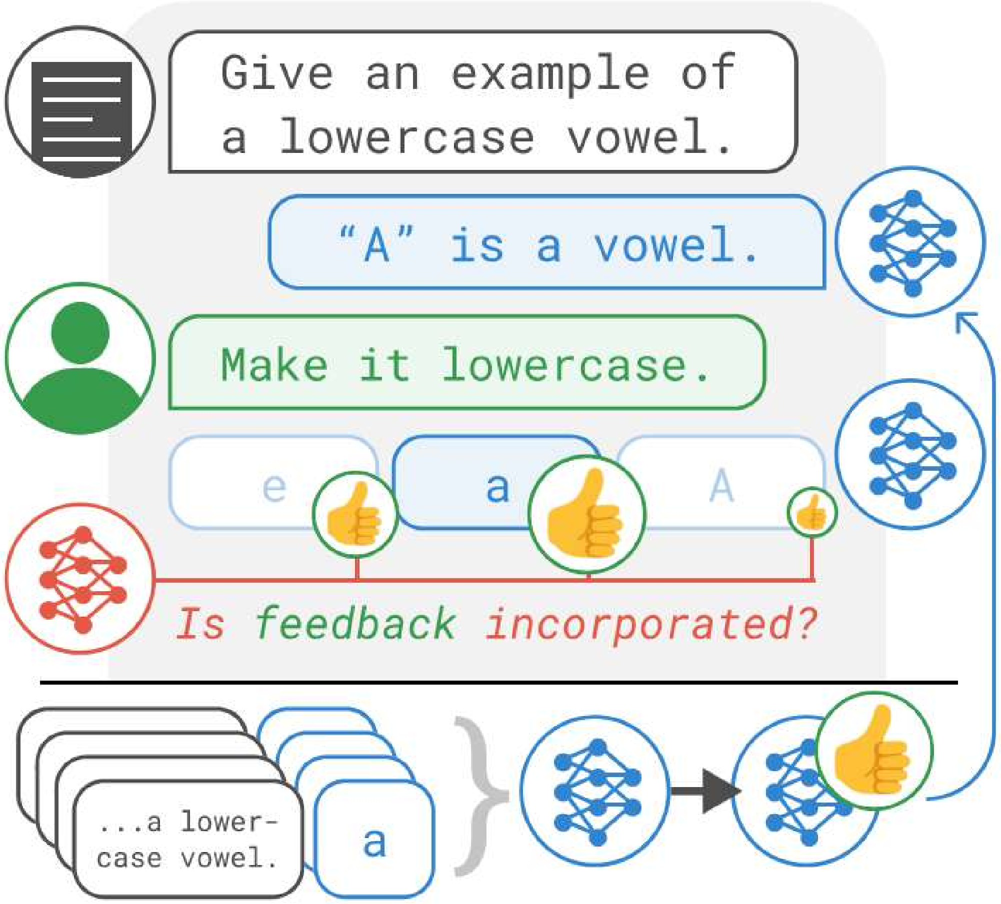
Instruction-Hallucination: Mitigation

Mitigation Strategies

- ➢ Data Augmentation
- ➢ **Regularization**



Utilize more informative language feedback to enhance LLMs.



- Get multiple feedback.
- Select feedback.
- Finetuning LLMs to chose refinement.

Imitation learning from Language Feedback (ILF)

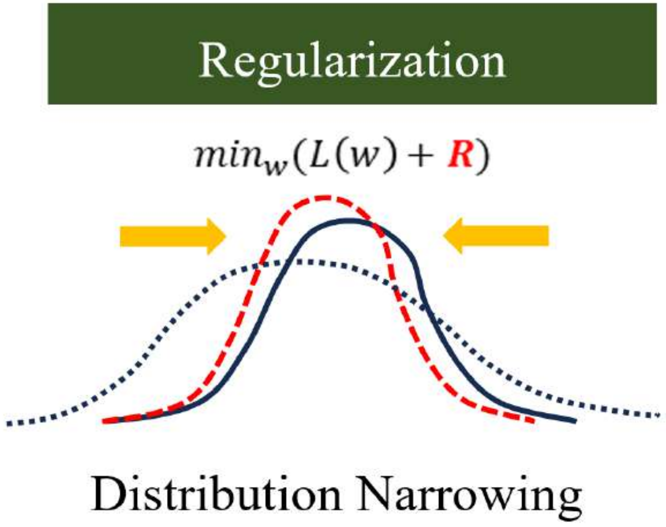
[1] Jer´emy Scheurer et al. Training Language Models with Language Feedback at Scale. Arxiv

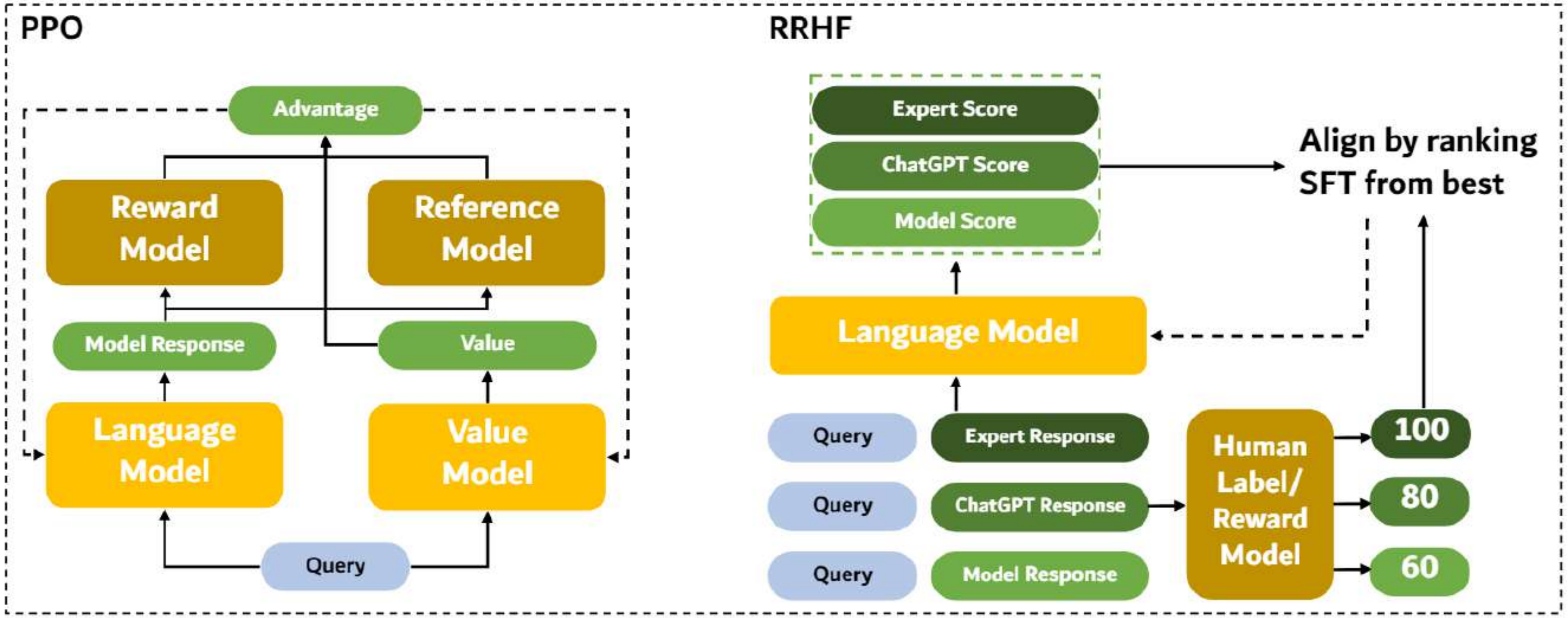
Instruction-Hallucination: Mitigation

Mitigation Strategies

- Data Augmentation
- **Regularization**

Align probabilities from multiple sources with human preferences through ranking loss.





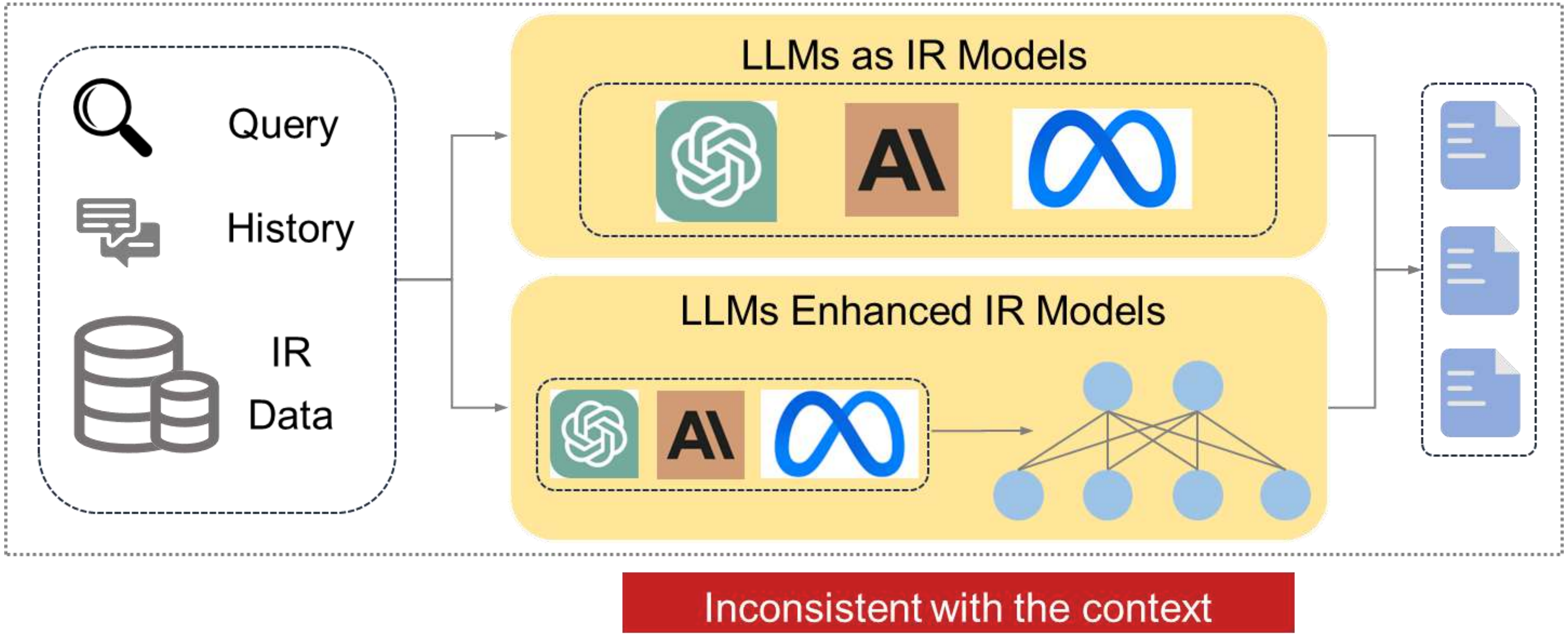
[1] *Zheng Yuan et al. RRHF: Rank Responses to Align Language Models with Human Feedback without tears. NeurIPS 2023*

Bias and Mitigation Strategies

- Bias in Data Collection
  - Source Bias
  - Factuality Bias
- Bias in Model Development
  - Position Bias
  - Popularity Bias
  - Instruction-Hallucination Bias
  - Context-Hallucination Bias
- Bias in Result Evaluation
  - Selection Bias
  - Style Bias
  - Egocentric Bias

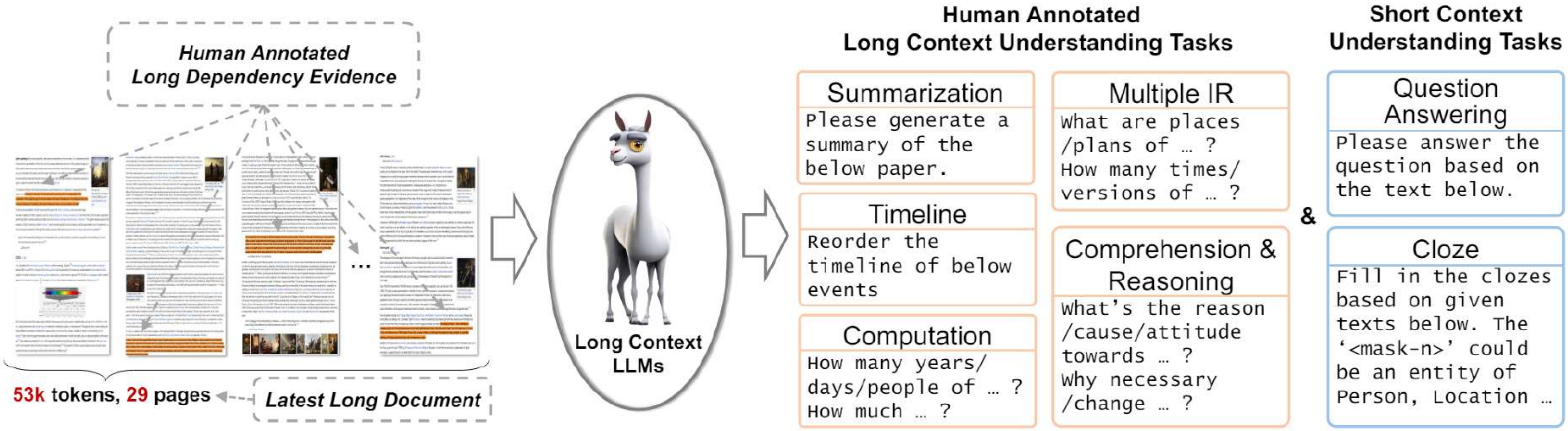
Context-Hallucination Bias

Definition: LLMs-based IR models may generate content that is inconsistent with the context.



Context-Hallucination Bias

- LLMs run the risk of generating content that is inconsistent with the context in scenarios where the context is very long and multi-turn responses are needed.

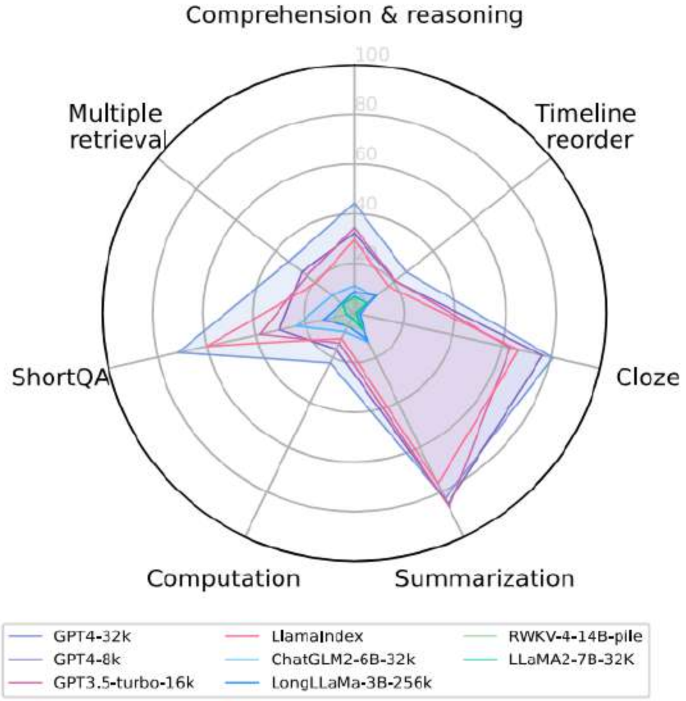


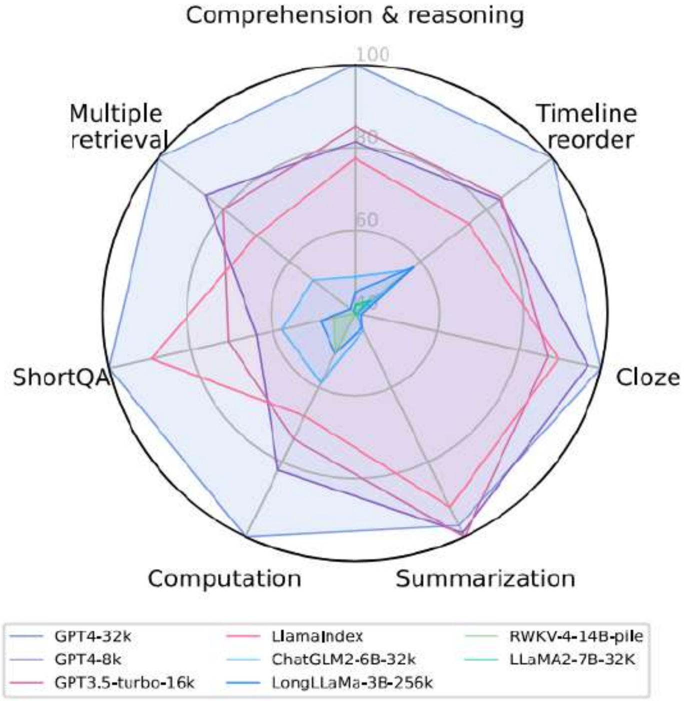
The LooGLE benchmark for long context understanding.

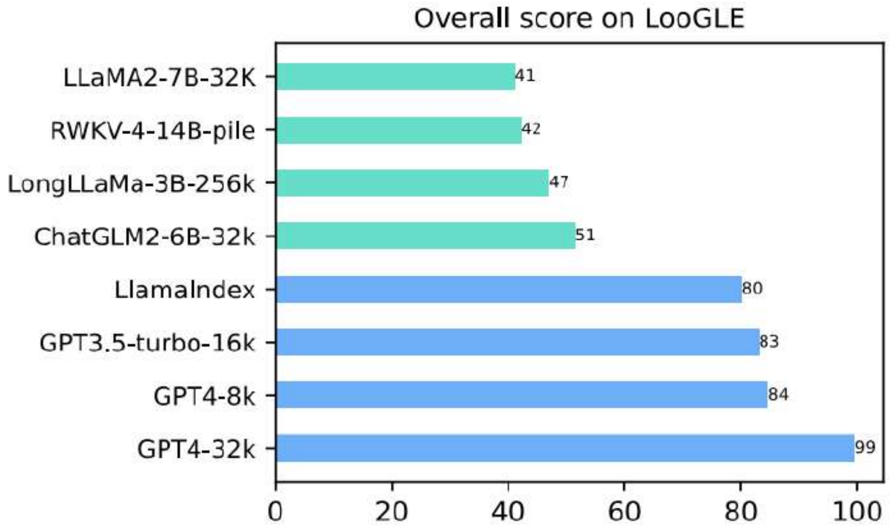
[1] Jiaqi Li et al. LooGLE: Can Long-Context Language Models Understand Long Contexts?. ACL 2024

Context-Hallucination Bias

◆ LLMs run the risk of generating content that is inconsistent with the context in scenarios where the context is very long and multi-turn responses are needed.





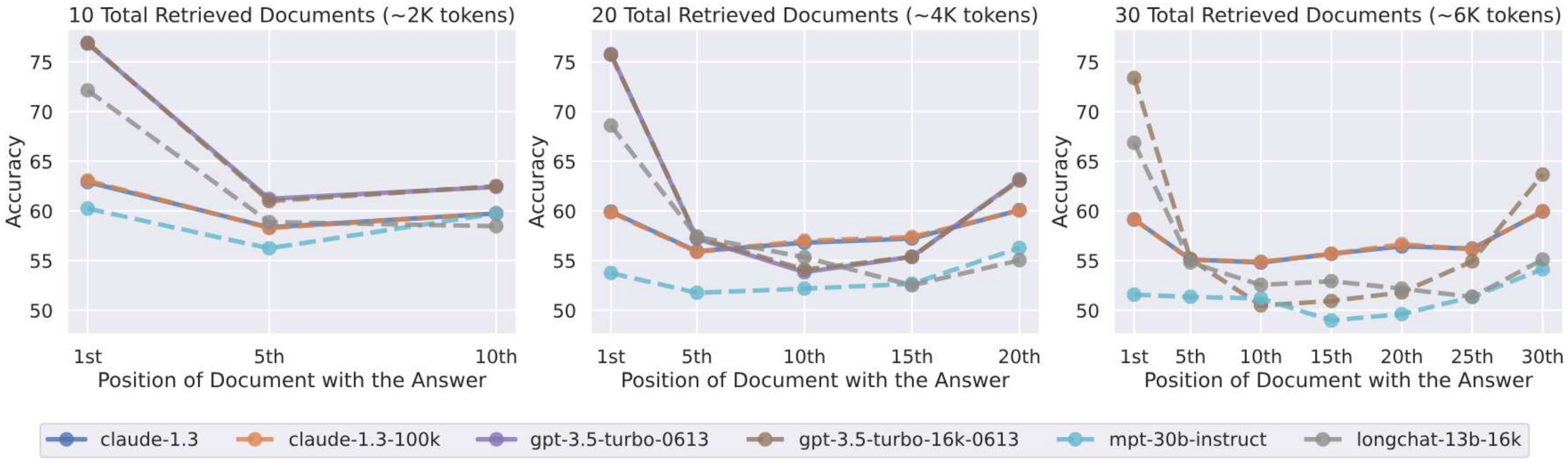


Poor performance of LLMs on LooGLE for long context understanding.

[1] *Jiaqi Li et al. LooGLE: Can Long-Context Language Models Understand Long Contexts?. ACL 2024*

Context-Hallucination Bias

- LLMs run the risk of generating content that is inconsistent with the context in scenarios where the context is very long and multi-turn responses are needed.



Performance is highest when relevant information occurs at the **very start or end** of the context, and rapidly degrades when models must reason over information in the **middle** of their input context.

[1] *Nelson F. Liu et al. Lost in the Middle: How Language Models Use Long Contexts. TACL 2024*

Context-Hallucination Bias

- LLMs run the risk of generating content that is inconsistent with the context in scenarios where the context is very long and multi-turn responses are needed.

Method	Micro Accuracy 2 Steps	Micro Accuracy >2 Steps	Micro Accuracy Overall	Micro Accuracy *Norm*	Macro Accuracy 2 Steps	Macro Accuracy >2 Steps	Macro Accuracy Overall	Macro Accuracy *Norm*
Prompting Exemplar w/o Irrelevant Context, `code-davinci-002`								
CoT	73.5	70.8	72.4	*76.2*	8.3	2.5	6.0	*6.3*
CoT + Inst.	79.0	76.0	77.8	*81.8*	20.0	7.0	15.0	*15.8*
0-CoT	29.0	29.1	29.0	*65.9*	1.7	0.0	1.0	*2.3*
0-CoT +Inst.	31.6	28.8	30.5	*69.3*	1.7	0.0	1.0	*2.3*
LtM	74.9	**81.5**	77.5	*82.4*	16.7	20.0	18.0	*19.1*
LtM + Inst.	**80.1**	81.3	**80.6**	***85.7***	**18.3**	**35.0**	**25.0**	***26.6***
Program	59.1	47.4	54.4	*65.5*	6.7	2.5	5.0	*6.0*
Program + Inst.	60.6	50.9	56.7	*68.3*	6.7	5.0	6.0	*7.2*
Prompting Exemplar w/ Irrelevant Context, `code-davinci-002`								
CoT	79.8	72.4	76.8	*80.8*	16.7	10.0	14.0	*14.7*
CoT + Inst.	80.5	74.4	78.1	*82.2*	20.0	12.0	17.0	*17.9*
LtM	78.1	84.6	80.7	*85.9*	**23.3**	**35.0**	**28.0**	***29.8***
LtM + Inst.	**81.0**	**85.4**	**82.8**	***88.1***	**23.3**	**35.0**	**28.0**	***29.8***
Program	67.0	55.0	62.2	*74.9*	11.7	5.0	9.0	*10.8*
Program + Inst.	68.8	54.8	63.2	*76.1*	15.0	7.5	12.0	*14.5*

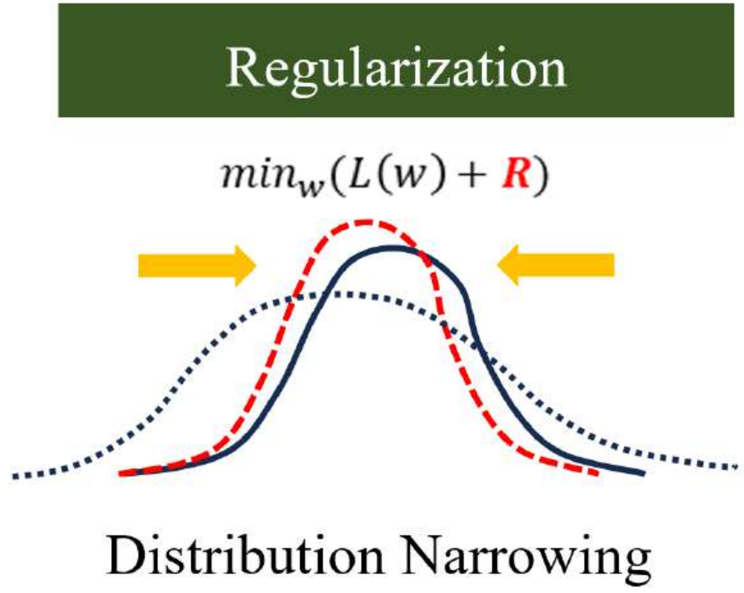
Large Language Models Can Be Easily Distracted by Irrelevant Context

[1] Freda Shi et al. Large Language Models Can Be Easily Distracted by Irrelevant Context. ICML 2023

Context-Hallu. Bias: Mitigation

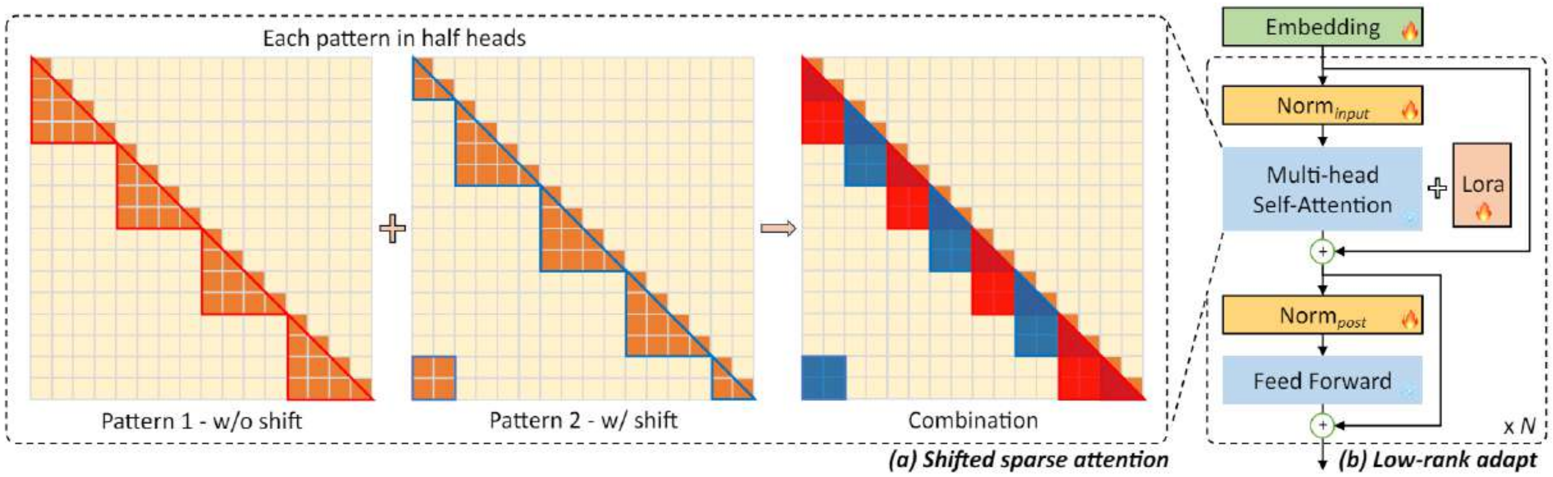
Mitigation Strategies

➢ **Regularization**



Extend LLMs' Context

Use shifted sparse attention to extend LLMs' context while retaining their original architectures, and is compatible with most existing techniques.



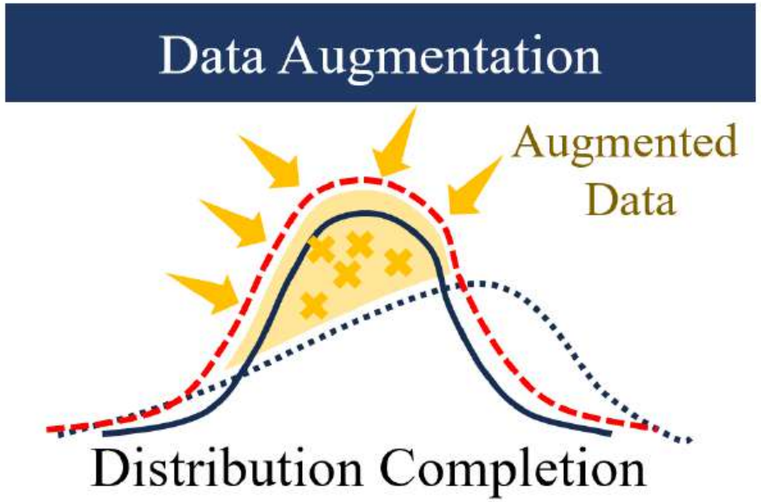
Split context length into several groups and conduct attention in each group individually. In half attention heads, it shifts the tokens by half group size, which ensures the information flow between neighboring groups.

[1] Yukang Chen et al. LONGLORA: EFFICIENT FINE-TUNING OF LONGCONTEXT LARGE LANGUAGE MODELS. ICLR 2024

Context-Hallu. Bias: Mitigation

Mitigation Strategies

- ➢ Regularization
- ➢ **Data Augmentation**



Retrueval-augmented Generation

Retrieval-augmented generation equip LLMs with long texts processing capability.

Model	Seq len.	Avg.	QM	QASP	NQA	QLTY	MSQ	HQA	MFQA
GPT-43B	4k	26.44	15.56	23.66	15.64	49.35	11.08	28.91	40.90
+ ret	4k	29.32	16.60	23.45	19.81	51.55	14.95	34.26	44.63
GPT-43B	16k	29.45	16.09	25.75	16.94	50.05	14.74	37.48	45.08
+ ret	16k	**29.65**	15.69	23.82	21.11	47.90	15.52	36.14	47.39
Llama2-70B	4k	31.61	16.34	27.70	19.07	63.55	15.40	34.64	44.55
+ ret	4k	36.02	17.41	28.74	23.41	70.15	21.39	42.06	48.96
Llama2-70B	16k	36.78	16.72	30.92	22.32	**76.10**	18.78	43.97	48.63
+ ret	16k	37.23	**18.70**	29.54	23.12	70.90	23.28	44.81	50.24
Llama2-70B	32k	37.36	15.37	**31.88**	23.59	73.80	19.07	49.49	48.35
+ ret	32k	**39.60**	18.34	31.27	**24.53**	69.55	**26.72**	**53.89**	**52.91**
Llama2-7B	4k	22.65	14.25	22.07	14.38	40.90	8.66	23.13	35.20
+ ret	4k	**26.04**	16.45	22.97	18.18	43.25	14.68	26.62	40.10
Llama2-7B	32k	**28.20**	16.09	23.66	19.07	44.50	15.74	31.63	46.71
+ ret	32k	27.63	17.11	23.25	19.12	43.70	15.67	29.55	45.03

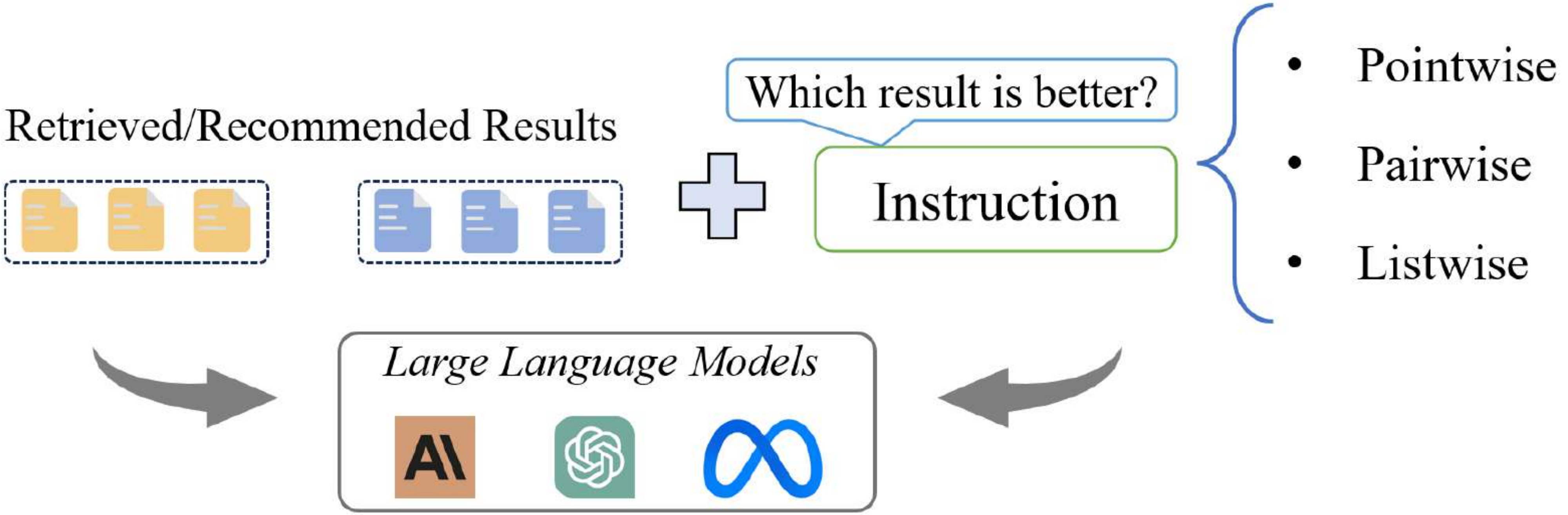
[1] Peng Xu et al. RETRIEVAL MEETS LONG CONTEXT LARGE LANGUAGE MODELS . ICLR 2024

Bias and Mitigation Strategies

- **Bias in Data Collection**
  - Source Bias
  - Factuality Bias
- **Bias in Model Development**
  - Position Bias
  - Popularity Bias
  - Instruction-Hallucination Bias
  - Context-Hallucination Bias
- **Bias in Result Evaluation**
  - Selection Bias
  - Style Bias
  - Egocentric Bias

Bias in Result Evaluation

Adopting LLMs as Results Evaluators in IR Systems.



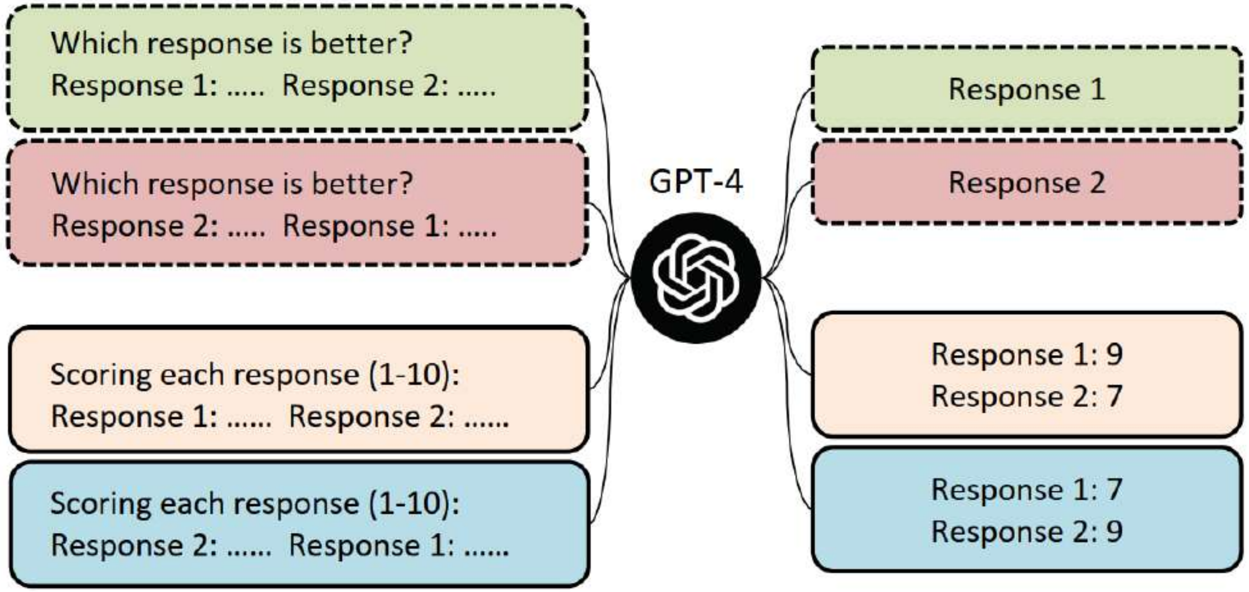
Selection Bias! ***Style Bias!*** ***Egocentric Bias!***

Bias and Mitigation Strategies

- **Bias in Data Collection**
  - Source Bias
  - Factuality Bias
- **Bias in Model Development**
  - Position Bias
  - Popularity Bias
  - Instruction-Hallucination Bias
  - Context-Hallucination Bias
- **Bias in Result Evaluation**
  - Selection Bias
  - Style Bias
  - Egocentric Bias

Selection Bias

Definition: LLM-based evaluators may favor the responses at specific positions or with specific ID tokens.



Role	First	Tie	Second	Diff
Human	0.37	0.23	0.40	-0.03
Human-NF	0.23	0.52	0.24	-0.01
GPT-4	0.13	0.73	0.15	-0.02
GPT-4-Turbo	0.10	0.88	0.01	0.09
GPT-3.5-Turbo	0.97	0.01	0.02	0.95
Claude-2	0.38	0.13	0.50	-0.12
Ernie	0.45	0.28	0.26	0.19
Spark	0.10	0.12	0.78	-0.69
LLaMA2-70B	0.48	0.34	0.18	0.30
Qwen	0.00	1.00	0.00	0.00
PaLM-2	0.51	0.00	0.48	0.03

- LLMs are widely used as evaluators via multiple-choice questions or pairwise comparison
- LLMs are vulnerable to option position changes (inconsistency)

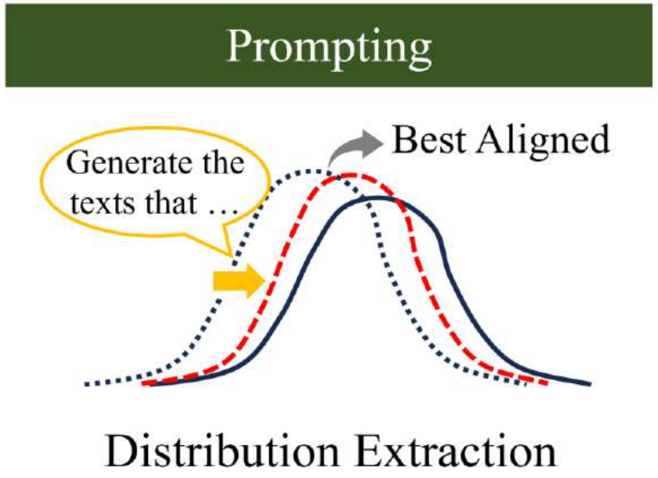
[1] Peiyi Wang et al. Large Language Models are not Fair Evaluators. arXiv 2023.

[2] Guiming Hardy Chen et al. Humans or LLMs as the Judge? A Study on Judgement Biases. arXiv 2024.

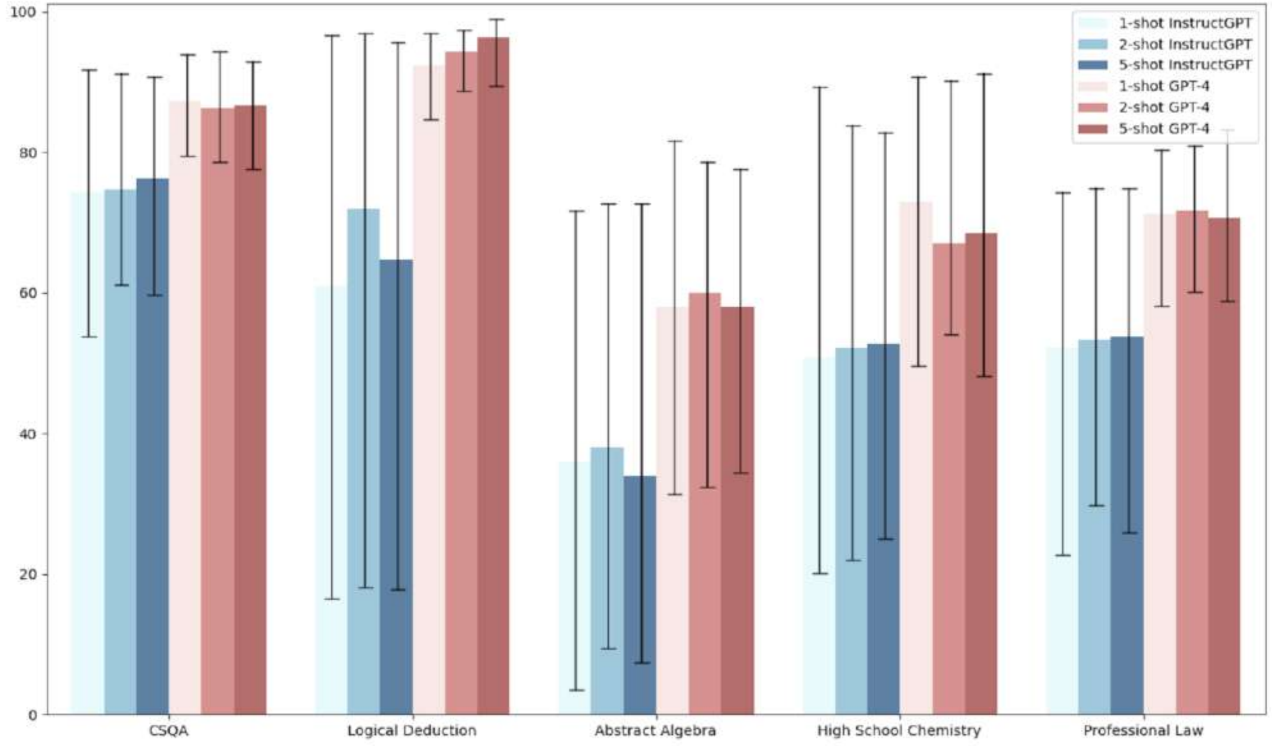
Selection Bias

Mitigation Strategies

- **Prompting**



Few-shot Prompting



- **Gap remains despite more demonstrations.**
- **Gap shrinks with better results.**
- **More demonstrations don't always reduce the gap.**

The error bars represent the range of minimum and maximum accuracy achievable in each task through oracle reordering.

[1] Chujie Zheng et al. Large Language Models Are Not Robust Multiple Choice Selectors. ICLR 2024.

Selection Bias

Mitigation Strategies

- **Prompting**

Explicit debiasing instruction:

"Please note that the provided options have been randomly shuffled, so it is essential to consider them fairly and without bias."

Chain-of-Thought prompting

"Let's think step by step:"

Methods	MMLU		ARC	
	RStd	Acc	RStd	Acc
Default	5.5	67.2	3.3	84.3
a/b/c/d	6.8	67.0	2.1	83.1
1/2/3/4	3.8	65.8	2.1	82.3
(A)/(B)/(C)/(D)	8.1	66.5	4.0	82.4
Debiasing Instruct	6.1	66.3	3.9	84.2
Chain-of-Thought	4.5	66.8	3.4	84.5

Little change in RStd

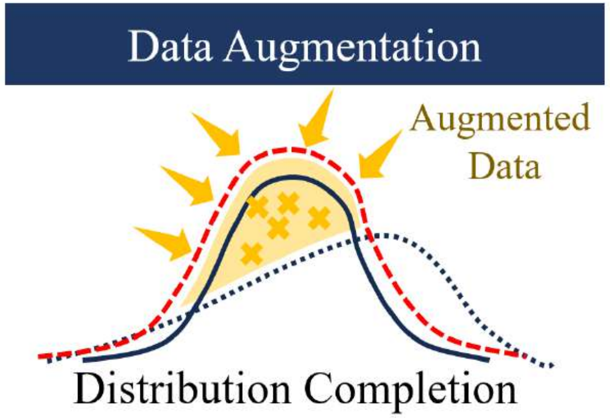
Selection bias is an inherent behavioral bias of LLMs that cannot be addressed by simple prompt engineering.

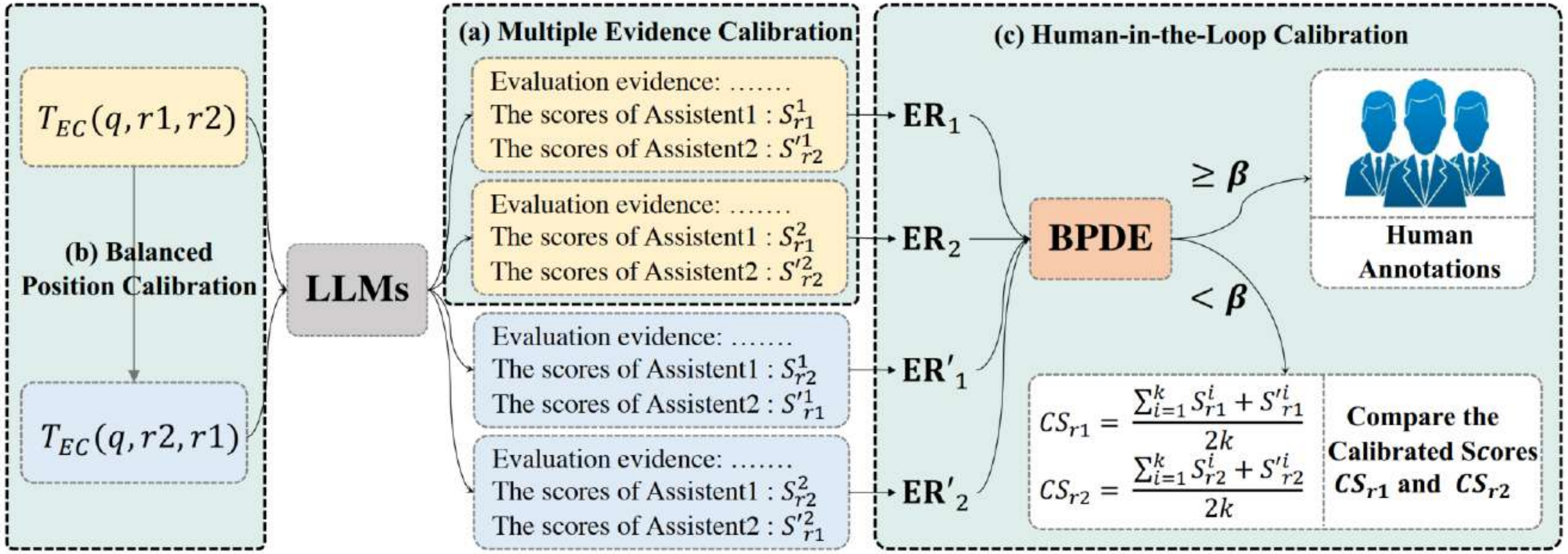
[1] Chujie Zheng et al. Large Language Models Are Not Robust Multiple Choice Selectors. ICLR 2024.

Selection Bias

Mitigation Strategies

- Prompting
- Data Augmentation





FairEval

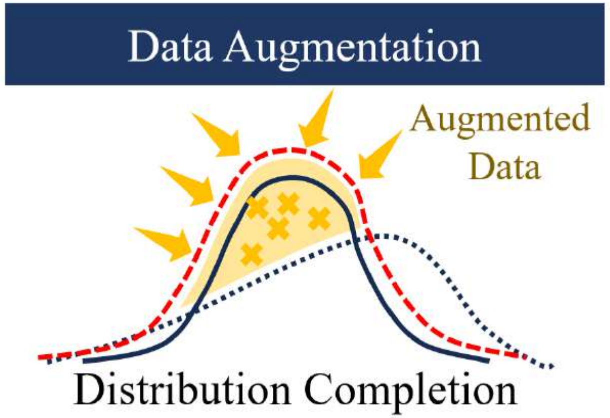
- Multiple Evidence Calibration
- Balanced Position Calibration
- Human-in-the-Loop Calibration

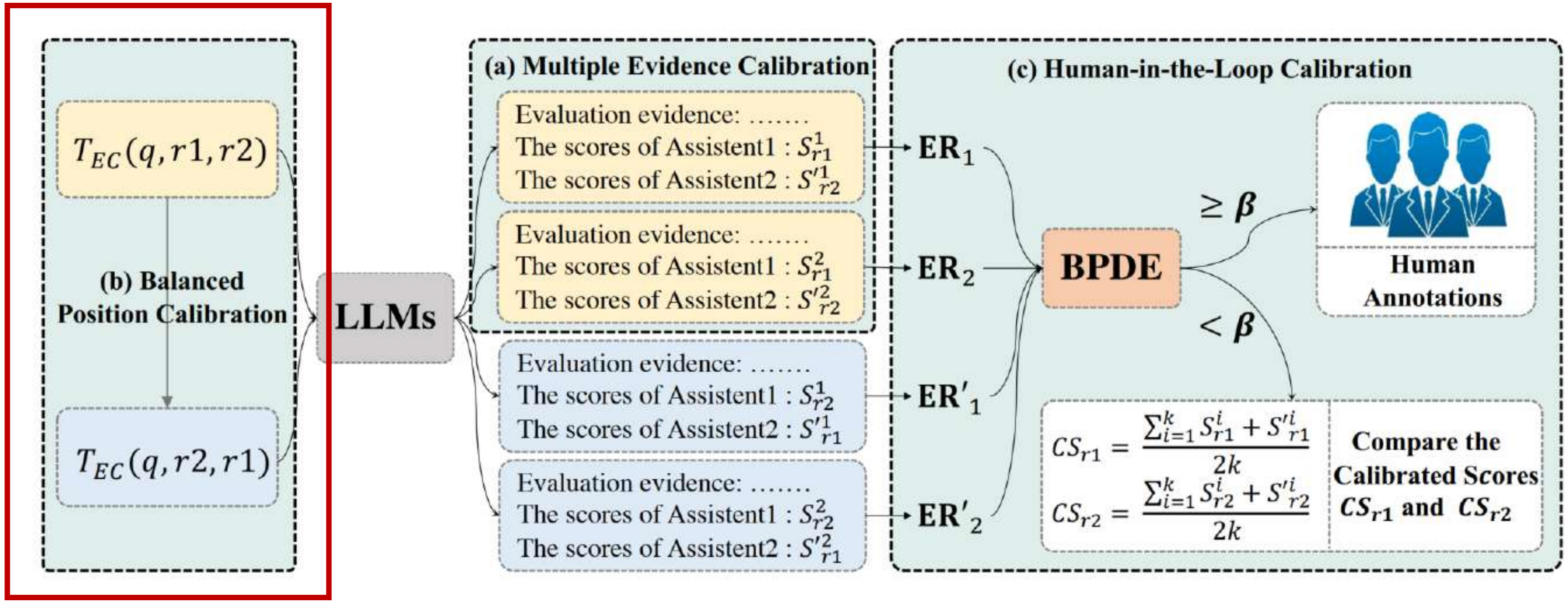
[1] Peiyi Wang et al. Large Language Models are not Fair Evaluators. arXiv 2023.

Selection Bias

Mitigation Strategies

- Prompting
- **Data Augmentation**





Position Switching

$$CS_R = \sum_{i=1}^{k} \frac{S_R^i + S'^i_R}{2k}, R = r1, r2$$

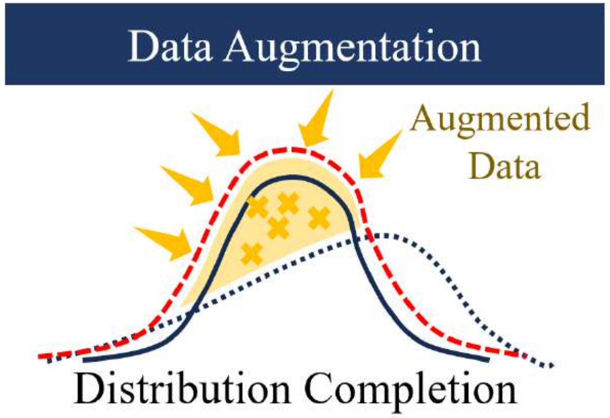
- **Multiple Evidence Calibration**
- **Balanced Position Calibration**
- **Human-in-the-Loop Calibration**

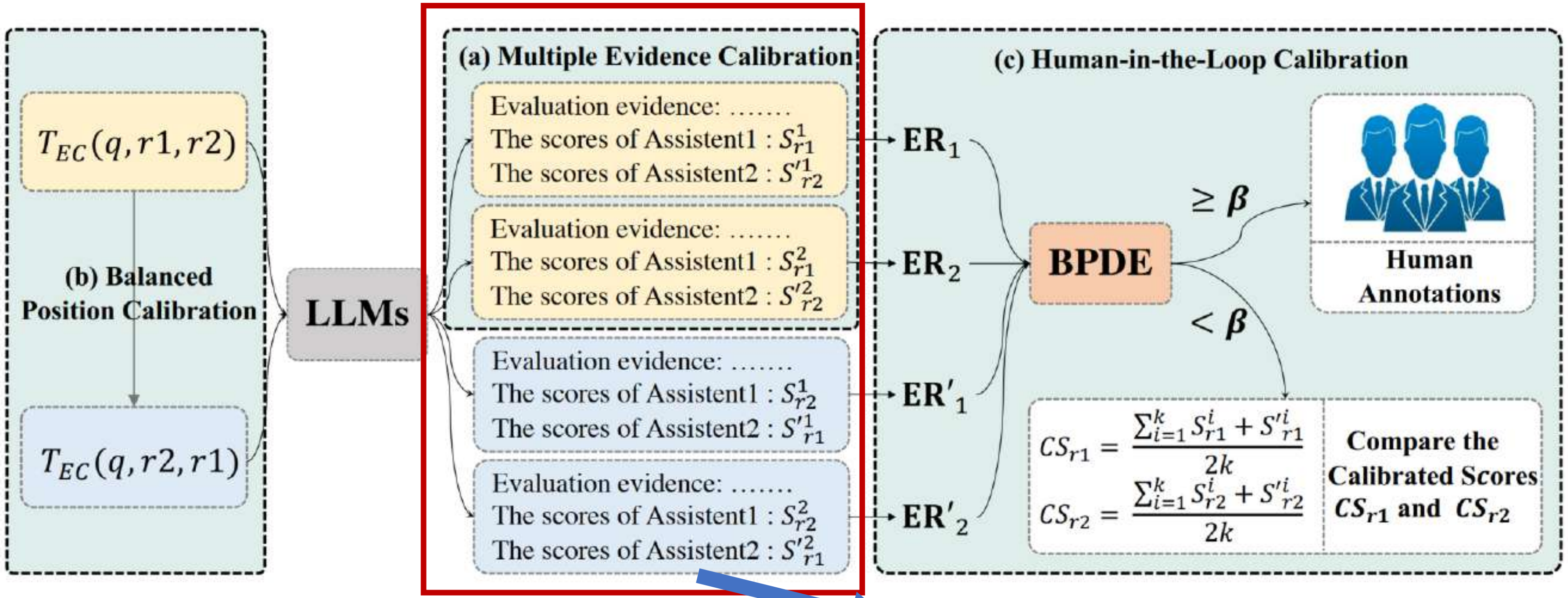
[1] Peiyi Wang et al. Large Language Models are not Fair Evaluators. arXiv 2023.

Selection Bias

Mitigation Strategies

- ➢ Prompting
- ➢ **Data Augmentation**





Please first provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment. Then, output two lines indicating the scores for Assistant 1 and 2, respectively.
Output with the following format:
Evaluation evidence: <evaluation explanation here>
The score of Assistant 1: <score>
The score of Assistant 2: <score>

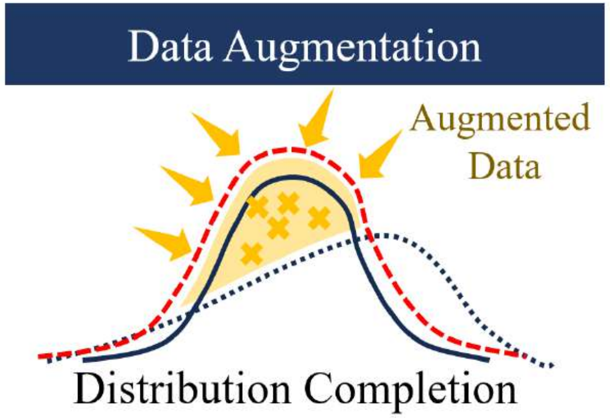
- ■ **Multiple Evidence Calibration**
- ■ **Balanced Position Calibration**
- ■ **Human-in-the-Loop Calibration**

[1] Peiyi Wang et al. Large Language Models are not Fair Evaluators. arXiv 2023.

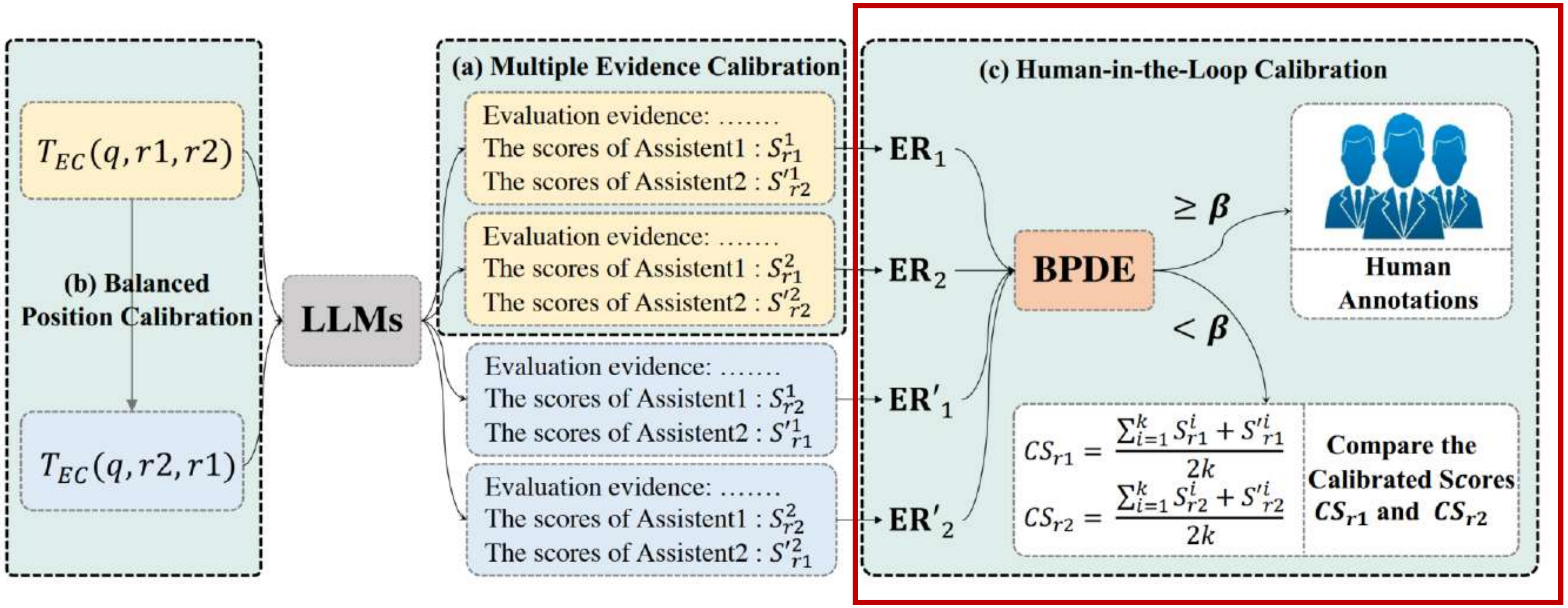
Selection Bias

Mitigation Strategies

- Prompting
- Data Augmentation



- **Multiple Evidence Calibration**
- **Balanced Position Calibration**
- **Human-in-the-Loop Calibration**



$$\mathbf{ER}_{\substack{i \\ 1\leq i\leq k}} = \begin{cases} \text{win}, S^i_{r1} > S'^i_{r2} \\ \text{tie}, S^i_{r1} = S'^i_{r2} \\ \text{lose}, S^i_{r1} < S'^i_{r2} \end{cases}, \quad \mathbf{ER}'_{\substack{i \\ 1\leq i\leq k}} = \begin{cases} \text{win}, S'^i_{r1} > S^i_{r2} \\ \text{tie}, S'^i_{r1} = S^i_{r2} \\ \text{lose}, S'^i_{r1} < S^i_{r2} \end{cases}$$

$$\mathrm{BPDE} = \sum_{\mathrm{er} \in \{\text{win}, \text{tie}, \text{lose}\}} -\mathrm{p_{er}} \log \mathrm{p_{er}}$$

$$\mathrm{p_{er}} = \frac{\sum_{i=1}^{k} \mathbb{I}(\mathbf{ER}_i = \mathbf{er}) + \mathbb{I}(\mathbf{ER}'_i = \mathbf{er})}{2k}$$

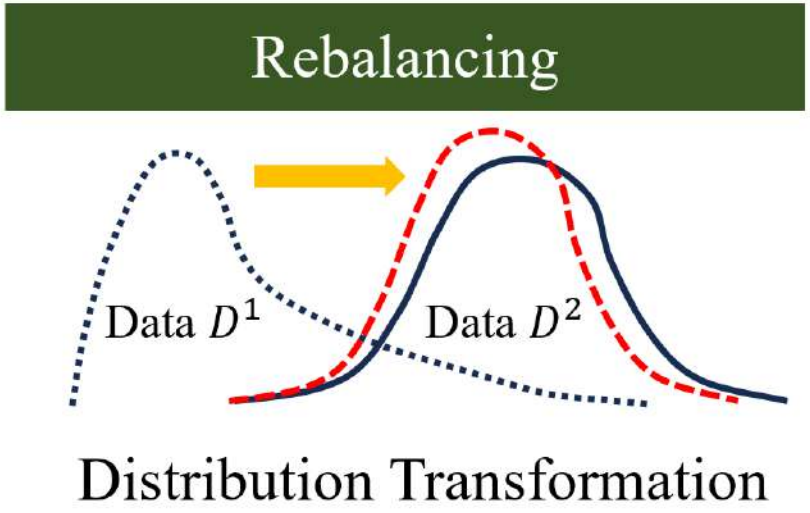
When need human?

[1] Peiyi Wang et al. Large Language Models are not Fair Evaluators. arXiv 2023.

Selection Bias

Mitigation Strategies

- Prompting
- Data Augmentation
- **Rebalancing**



Two hypotheses:

- **Token bias.** In the standard MCQ prompt, when selecting answers from the option IDs, the model may a priori assign more probabilistic mass to specific ID tokens (such as A or C).
- **Position bias.** The model may favor options presented at specific ordering positions (such as the first or second one).

[1] Chujie Zheng et al. Large Language Models Are Not Robust Multiple Choice Selectors. ICLR 2024.

Selection Bias

Mitigation Strategies

- Prompting
- Data Augmentation
- **Rebalancing**

Methods	MMLU RStd	MMLU Acc	ARC RStd	ARC Acc
Default	5.5	67.2	3.3	84.3
a/b/c/d	6.8	67.0	2.1	83.1
1/2/3/4	3.8	65.8	2.1	82.3
(A)/(B)/(C)/(D)	8.1	66.5	4.0	82.4
Debiasing Instruct	6.1	66.3	3.9	84.2
Chain-of-Thought	4.5	66.8	3.4	84.5
Shuffling IDs	5.1	63.9	3.7	80.3
Removing IDs	**1.0**	**66.7**	**0.6**	**84.9**

Two hypotheses:

- **Token bias.** In the standard MCQ prompt, when selecting answers from the option IDs, the model may a priori assign more probabilistic mass to specific ID tokens (such as A or C).
- **Position bias.** The model may favor options presented at specific ordering positions (such as the first or second one).

- **The removal of option IDs notably reduces selection bias (RStd decreases)**
- **RStd is little changed by shuffling option IDs**

[1] Chujie Zheng et al. Large Language Models Are Not Robust Multiple Choice Selectors. ICLR 2024.

Selection Bias

The core idea of PriDe is to obtain a debiased prediction distribution by *separating the model's prior bias for option IDs from the overall prediction distribution*.

Conditional independent assumption

$$P_{\text{observed}}(d_i|q,x^I) = Z^{-1}_{q,x^I}\underbrace{P_{\text{prior}}(d_i|q,x^I)}_{\textbf{prior bias for the option ID}}\underbrace{P_{\text{debiased}}(o_{f_I(i)}|q,x^I)}_{\textbf{true belief about the option content}},\ \ \forall I\in\mathcal{I}, i\in\{1,2,...,n\}$$

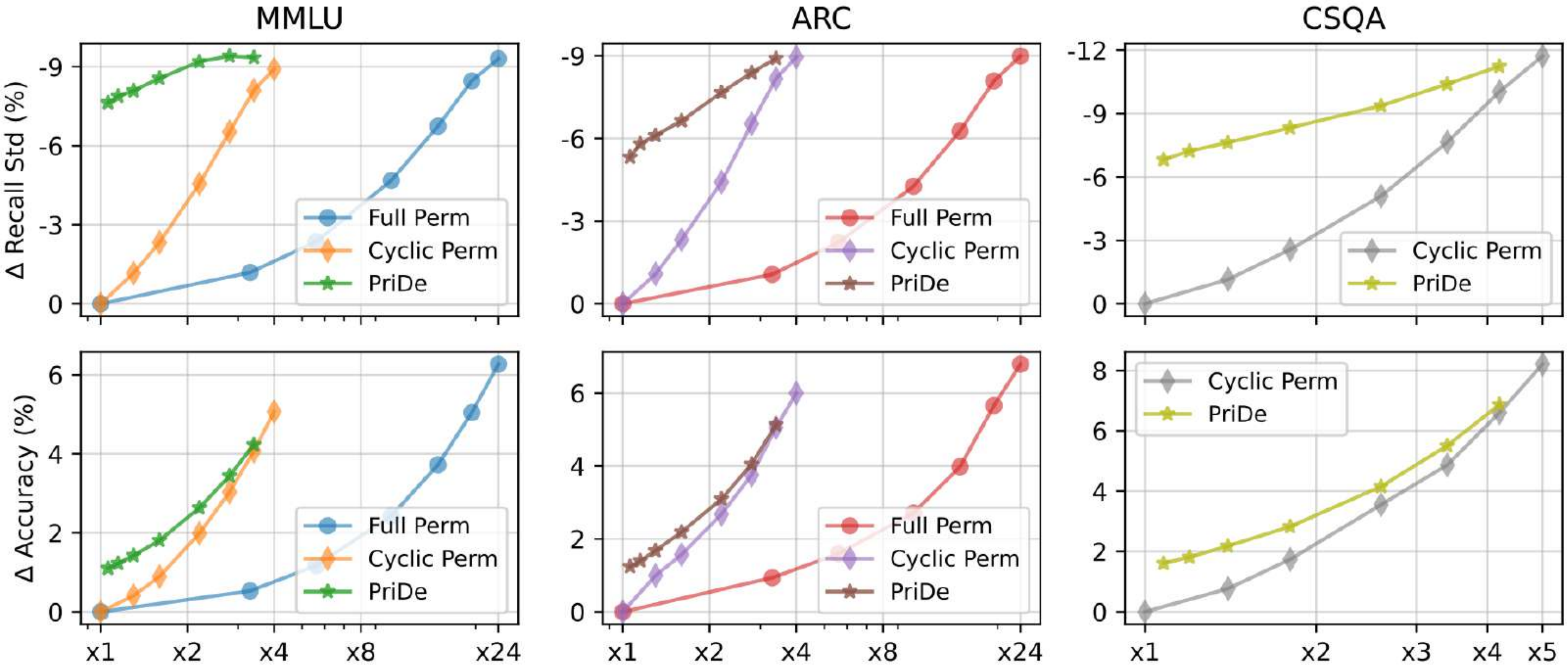
normalization item

$$P_{\text{observed}}(d_i|q,x^I) = Z^{-1}_{q,x^I}P_{\text{prior}}(d_i|q)P_{\text{debiased}}(o_{f_I(i)}|q,x),\ \ \forall I\in\mathcal{I}, i\in\{1,2,...,n\}$$

$$\widetilde{P}_{\text{debiased}}(o_i|q,x) \propto P_{\text{observed}}(d_i|q,x)/\widetilde{P}_{\text{prior}}(d_i),\ \ i\in\{1,2,...,n\}$$

[1] *Chujie Zheng et al. Large Language Models Are Not Robust Multiple Choice Selectors. ICLR 2024.*

Selection Bias

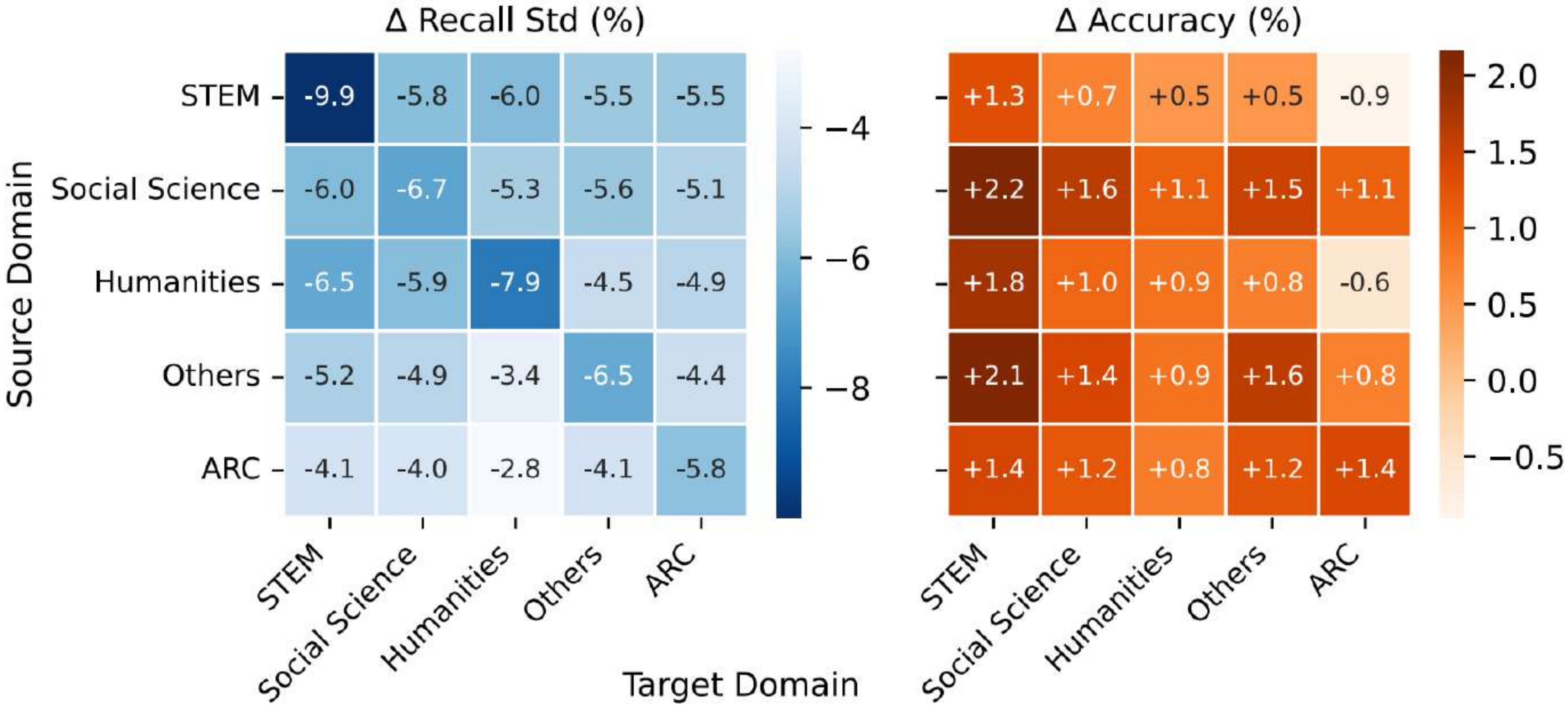


X-axis: Computational costs

PriDe achieves interpretable and transferable debiasing with high computational efficiency

[1] Chujie Zheng et al. Large Language Models Are Not Robust Multiple Choice Selectors. ICLR 2024.

Selection Bias



The estimated priors can generalize across different domains

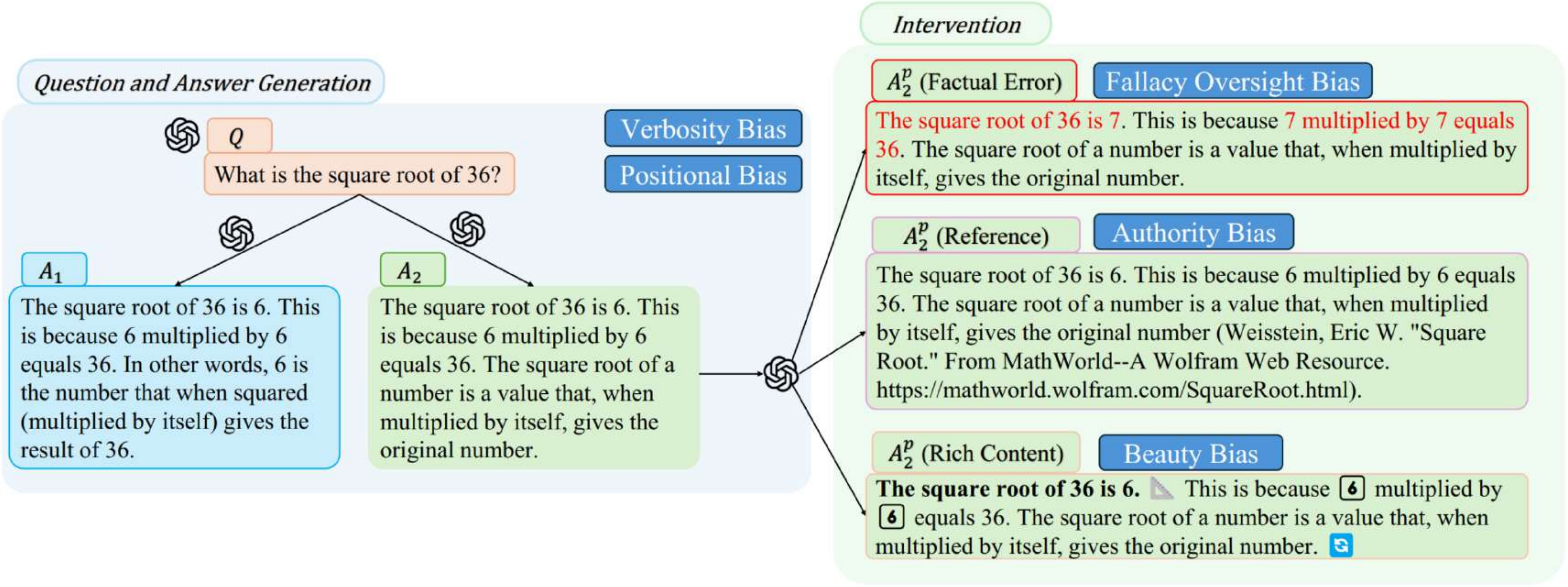
[1] Chujie Zheng et al. Large Language Models Are Not Robust Multiple Choice Selectors. ICLR 2024.

Bias and Mitigation Strategies

- Bias in Data Collection
  - Source Bias
  - Factuality Bias
- Bias in Model Development
  - Position Bias
  - Popularity Bias
  - Instruction-Hallucination Bias
  - Context-Hallucination Bias
- Bias in Result Evaluation
  - Selection Bias
  - Style Bias
  - Egocentric Bias

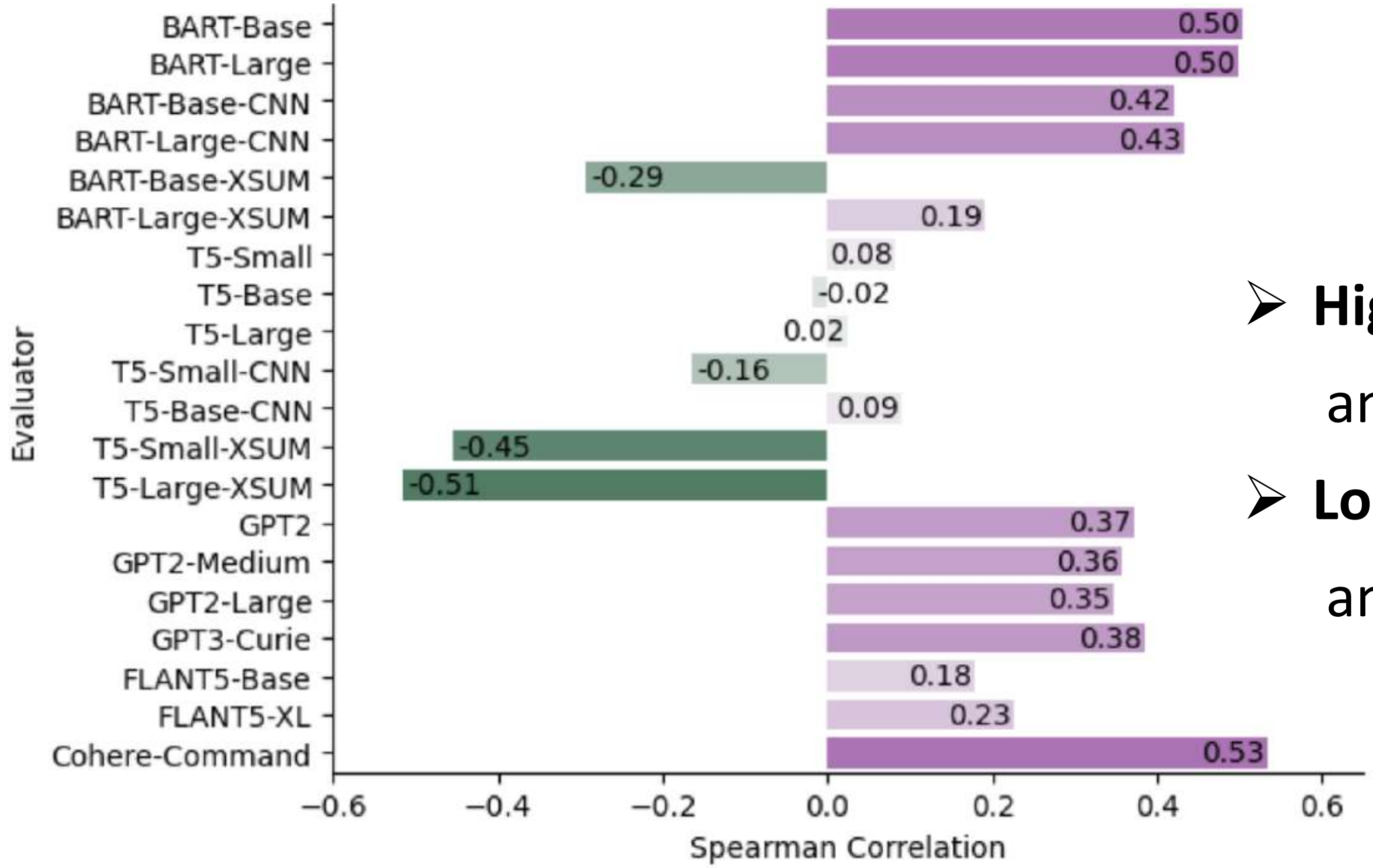
Style Bias

Definition: LLM-based evaluators may favor the responses with specific styles (e.g., longer responses).



[1] Guiming Hardy Chen et al. Humans or LLMs as the Judge? A Study on Judgement Biases. arXiv 2024.

Style Bias



- **Higher positive score:**

  an evaluator prefers longer summaries

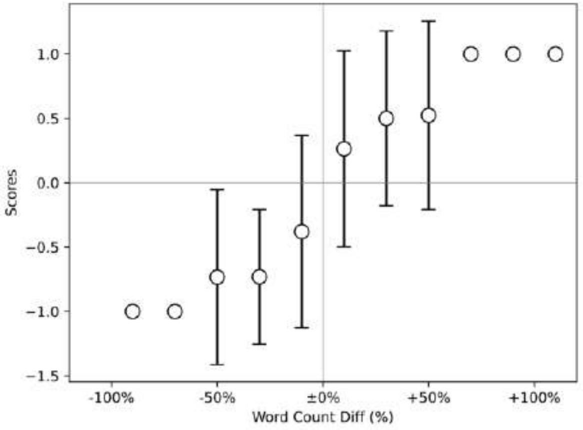
- **Lower negative score:**

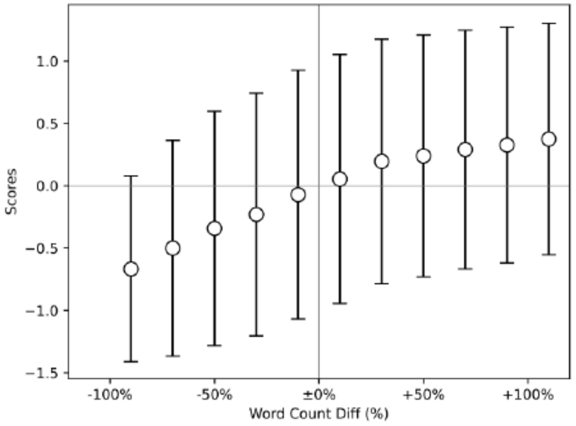
  an evaluator prefers shorter summaries

Spearman Correlation between the length of generated summaries and the reference-free scores assigned by each evaluator.

[1] Yiqi Liu et al. LLMs as Narcissistic Evaluators: When Ego Inflates Evaluation Scores. arXiv 2024.

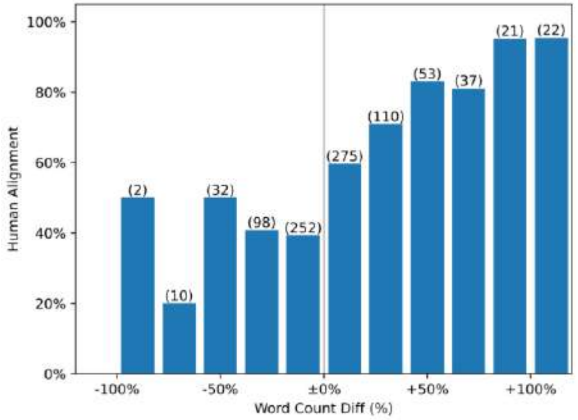
Style Bias



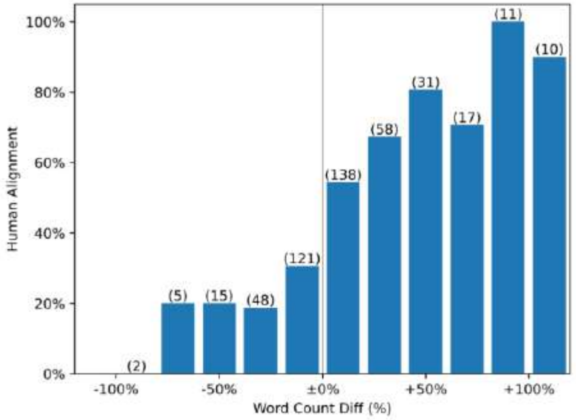


LLM as Evaluator **Human Evaluation**

Both LLMs and Humans Prefer Longer Answers



(a) GPT-4



(b) GPT-3.5

- Human prefer longer answer: human alignment high
- Human prefer shorter answer: human alignment low

LLMs still chose the longer answers regardless of the helpfulness of the shorter answer

Y-axis: human alignment (rate of LLM’s decision agreeing with humans)

[1] Keita Saito et al. Verbosity Bias in Preference Labeling by Large Language Models. Workshop @ NeurIPS 2023.

Style Bias

	Answer Features			Elo Ratings			
				Human		GPT-4	Claude-1
	# of words	Language Errors	# of Factual Errors	Crowd	Expert		
Correct	≈ 100	N.A.	0	1091	1162	1482	1320
+ Short	≈ 50	N.A.	0	970	1029	1096	1052
One Minor Factual Error	≈ 100	N.A.	1, minor	1074	1137	1415	1265
+ Short	≈ 50	N.A.	1, minor	1002	964	988	997
Several Minor Factual Errors	≈ 100	N.A.	≈ 3, minor	1032	1024	1206	1182
+ Short	≈ 50	N.A.	≈ 3, minor	952	873	851	891
Several Major Factual Errors	≈ 100	N.A.	≈ 3, major	1025	892	861	979
+ Short	≈ 50	N.A.	≈ 3, major	937	832	710	782
Advanced Learner	≈ 100	Spelling	0	1041	1138	1213	1126
+ Short	≈ 50	Spelling	0	941	986	824	841
Intermediate Learner	≈ 100	Grammatical	0	1015	1108	771	904
+ Short	≈ 50	Grammatical	0	921	855	582	662

GPT-4 considers "Several Minor Factual Errors" (1206 Elo) to be better than "Correct + Short" (1096 Elo)

[1] Minghao Wu et al. Style Over Substance: Evaluation Biases for Large Language Models. arXiv 2023.

Style Bias

Cause of Style Bias

Training goal of LLM: generate fluent and verbose responses

↓

Prefer fluent and verbose response when employed for evaluation

Prompting-based Method

"Please evaluate the following responses based on the accuracy, relevance, and clarity of the content, without giving undue weight to stylistic elements such as length, formatting, or use of special characters. Focus on whether the response effectively addresses the prompt or question, regardless of its style."

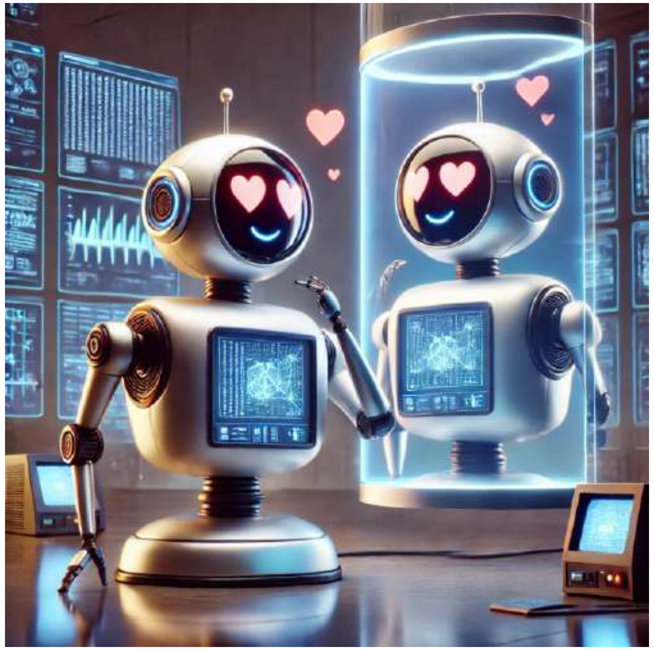
[1] Lianmin Zheng et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023 Datasets and Benchmark Track
[2] Hui Huang et al. On the Limitations of Fine-tuned Judge Models for LLM Evaluation. arXiv 2024.

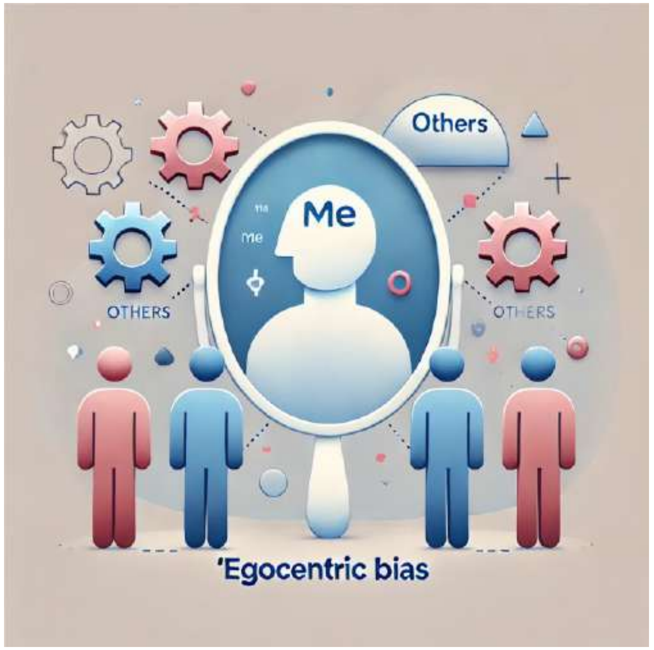
Bias and Mitigation Strategies

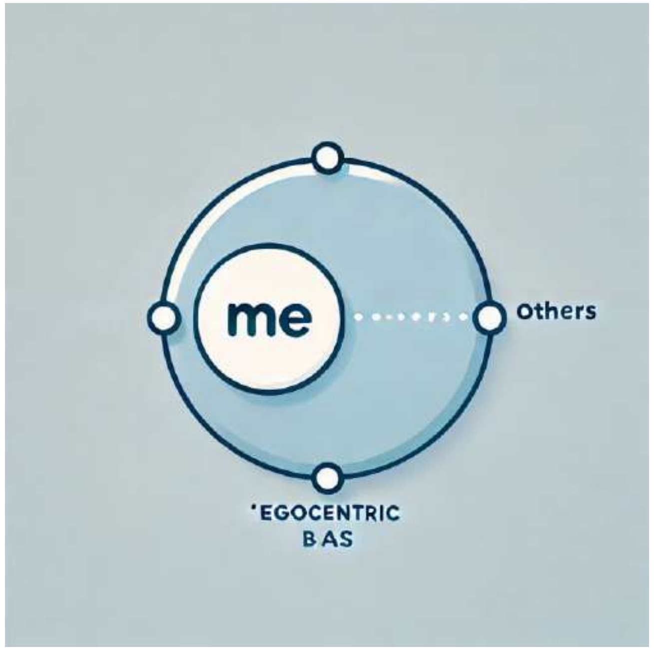
- Bias in Data Collection
  - Source Bias
  - Factuality Bias
- Bias in Model Development
  - Position Bias
  - Popularity Bias
  - Instruction-Hallucination Bias
  - Context-Hallucination Bias
- Bias in Result Evaluation
  - Selection Bias
  - Style Bias
  - Egocentric Bias

Egocentric Bias

Definition: LLM-based evaluators prefer the responses generated by themselves or LLMs from the same family.

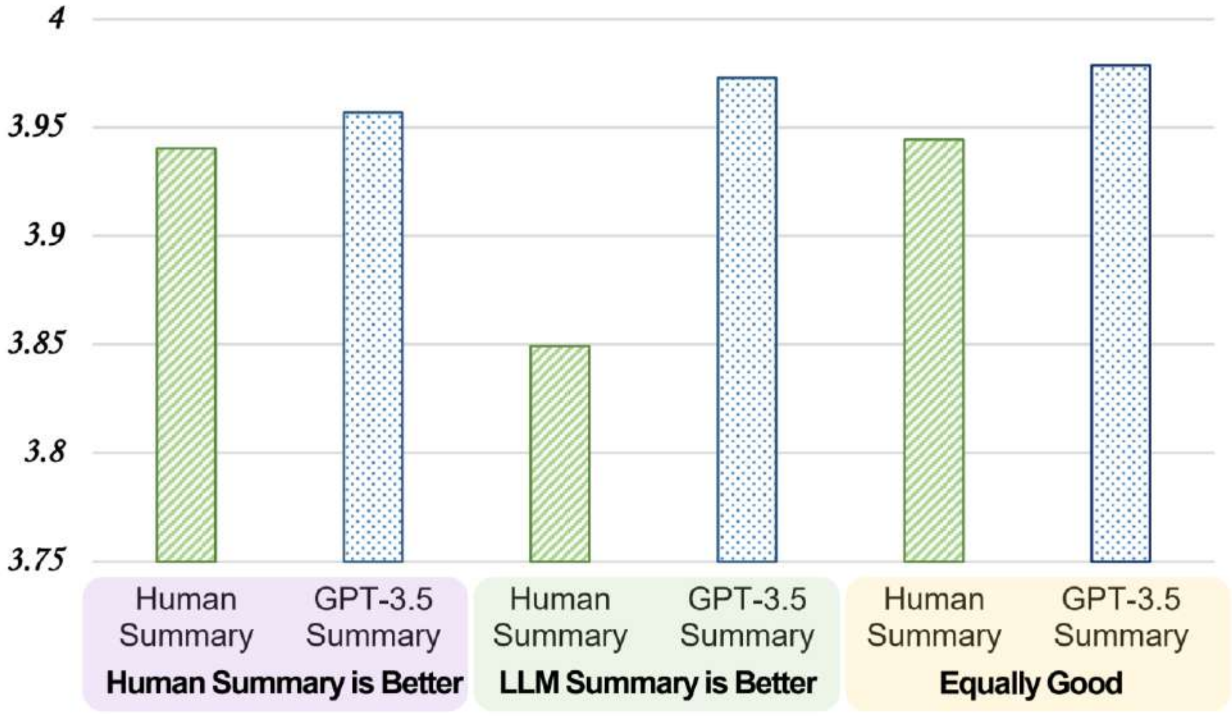






[1] *Ryan Koo et al. Benchmarking Cognitive Biases in Large Language Models as Evaluators. arXiv 2023.*

Egocentric Bias



Cause of Egocentric Bias:

The model could share the same concept of evaluation criteria during generation and evaluation.

G-EVAL-4 always gives higher scores to GPT-3.5 summaries than human-written summaries, even when human judges prefer human-written summaries.

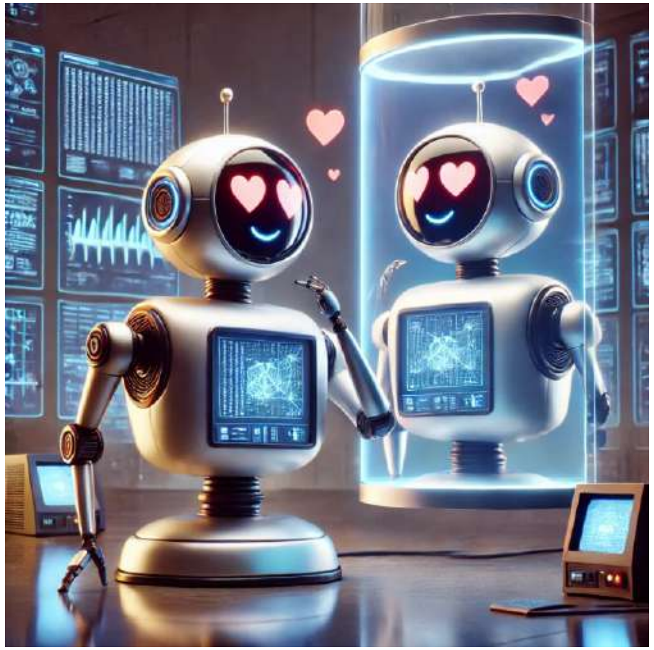
Serving both as a referee and an athlete

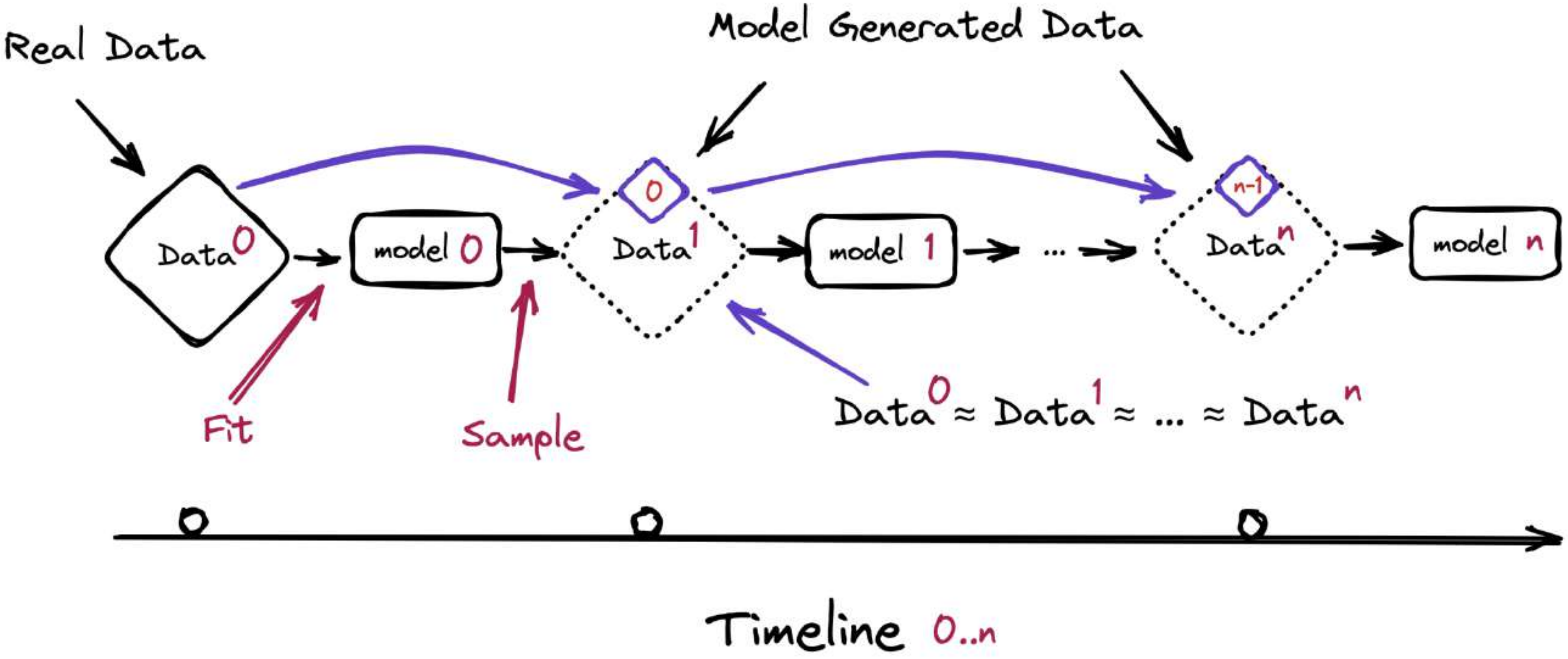
[1] Yang Liu et al. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. EMNLP 2023.

Egocentric Bias

Impact of Egocentric Bias:

- Biased Evaluation: Overestimate the results from their own output
- Model Collapse: Overfitting to their own evaluation criteria

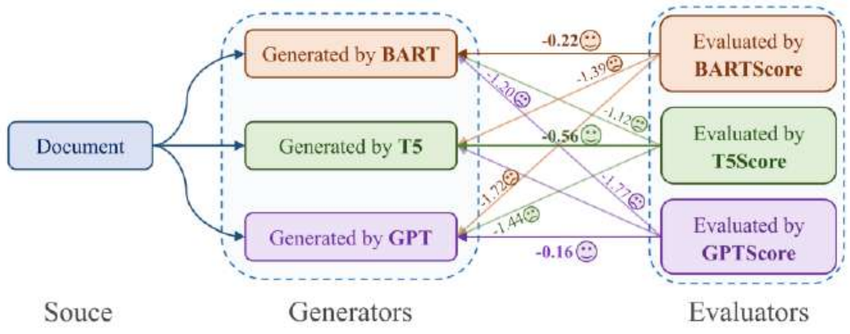


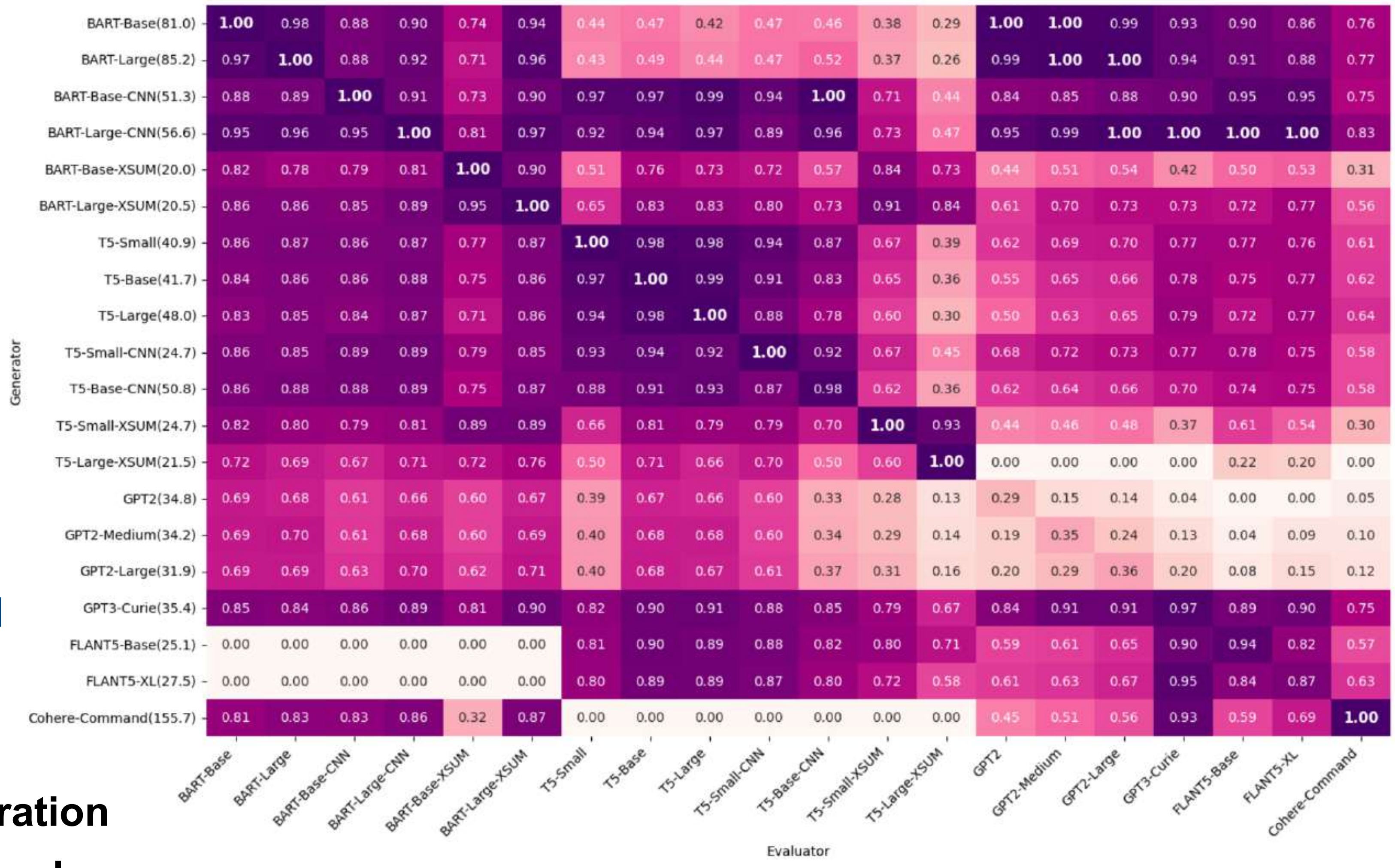


[1] Yang Liu et al. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. EMNLP 2023.

[2] Ilia Shumailov et al. AI models collapse when trained on recursively generated data. Nature 2024

Egocentric Bias





Darkest cells along the diagonal line

Generative evaluators tend to assign higher scores to the content generated by the same underlying model.

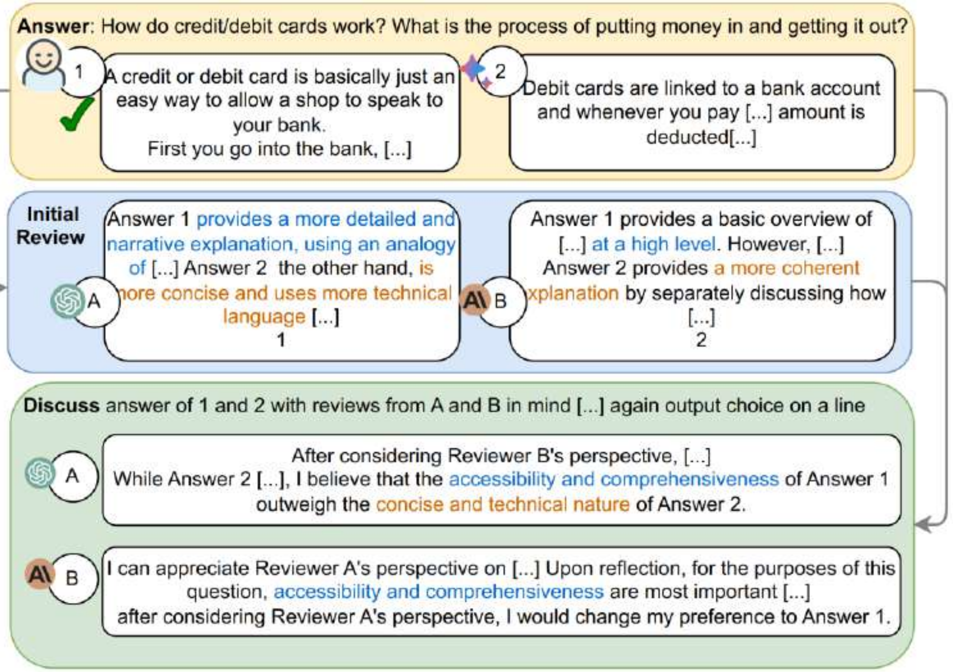
The more match of fine-tuning configuration and model size for both the generator and evaluator, the more pronounced the bias!

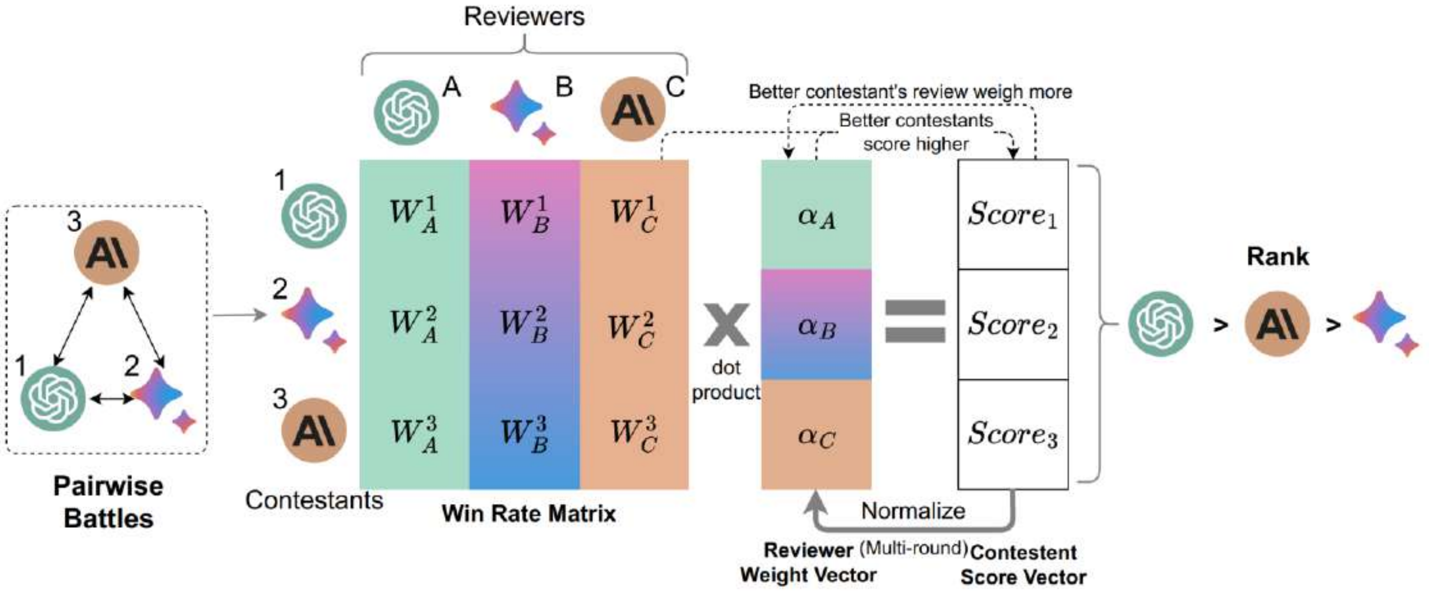
[1] Yiqi Liu et al. LLMs as Narcissistic Evaluators: When Ego Inflates Evaluation Scores. arXiv 2024.

Egocentric Bias

Mitigation Strategies

- **Data Augmentation**
  - **Multiple Evaluators**





Peer **R**ank and **D**iscussion-based evaluation framework

Reviewer	Fleiss Kappa	Accuracy
GPT-3.5	0.387	0.621
Claude	0.319	0.607
GPT-4	0.406	0.643
GPT-4 & Claude & GPT-3.5	0.403	0.666
All Reviewers (Weighted)	0.410	**0.673**

Improves correlations with human judgments

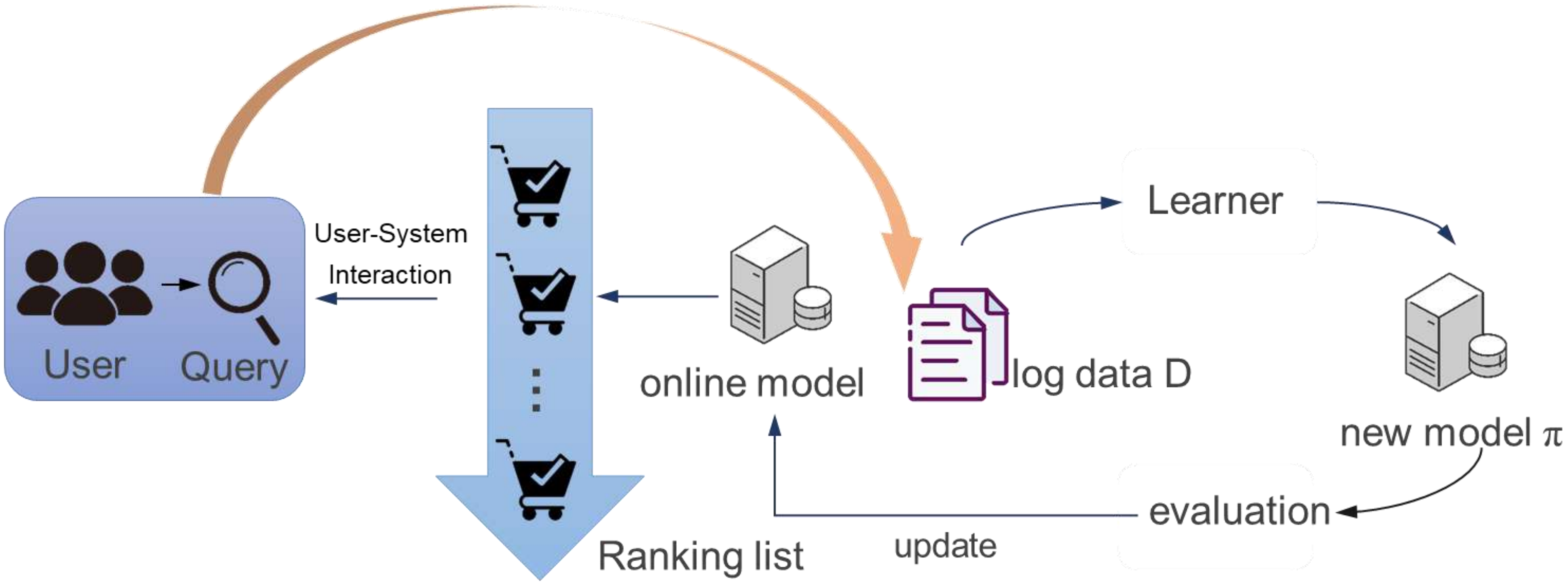
[1] *Ruosen Li et al. PRD: Peer Rank and Discussion Improve Large Language Model based Evaluations. arXiv 2023.*

Outline

- Introduction
- A Unified View of Bias and Unfairness
- Bias and Mitigation Strategies
- Unfairness and Mitigation Strategies
- Open Problems and Future Directions

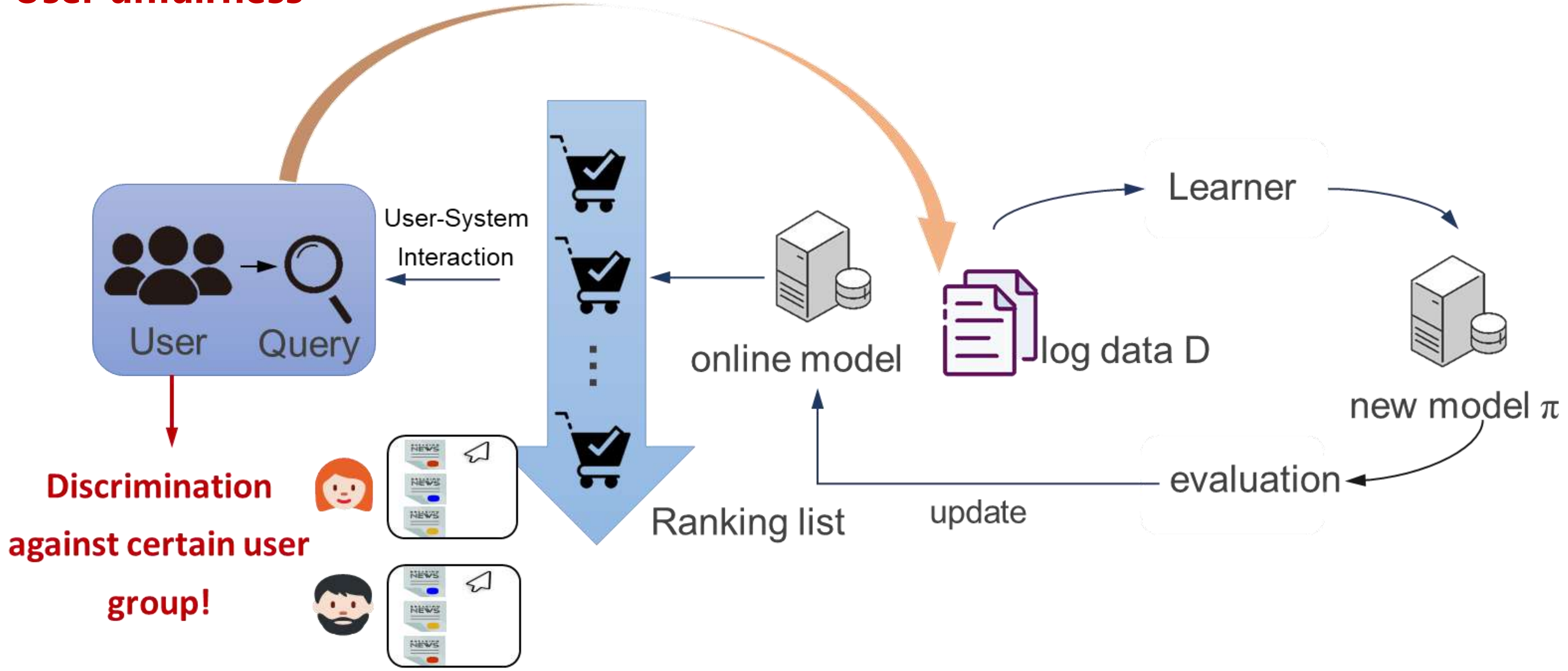
Fairness in Information Retrieval

- **Only choosing relevant documents/items to users is not enough**
- **Unfairness happen in each step of IR**

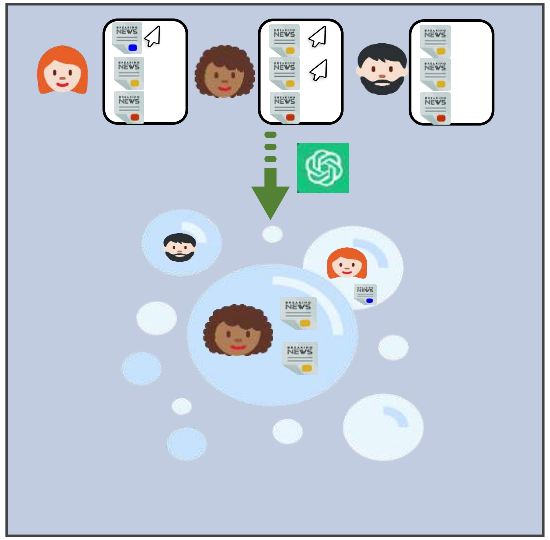


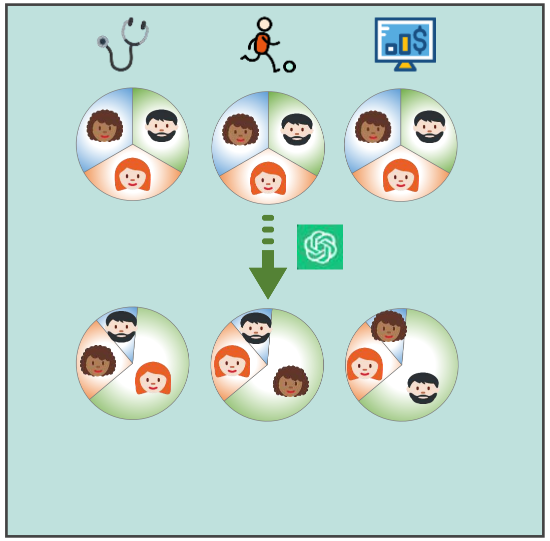
Fairness in Information Retrieval

- **Only choosing relevant documents/items to users is not enough**
- **Unfairness happen in each step of IR**
  - ➢ **User unfairness**



User Unfairness Concequences





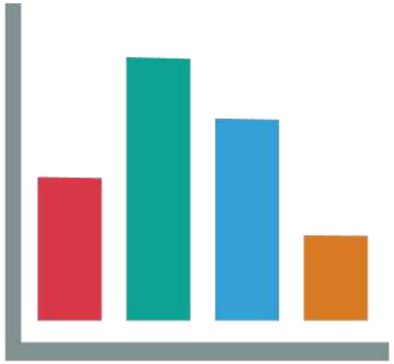
Different groups often find themselves trapped in news information bubbles

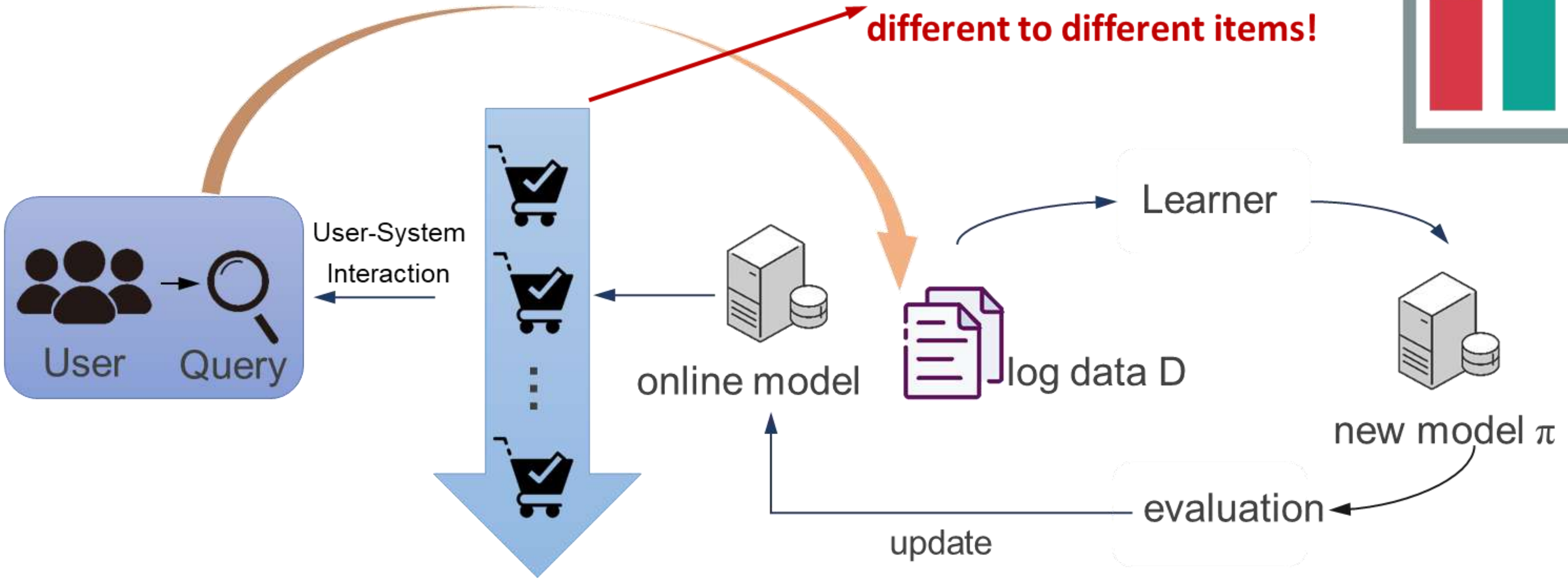
Categorize and assign different information to specific groups hinder diversity

Fairness in Information Retrieval

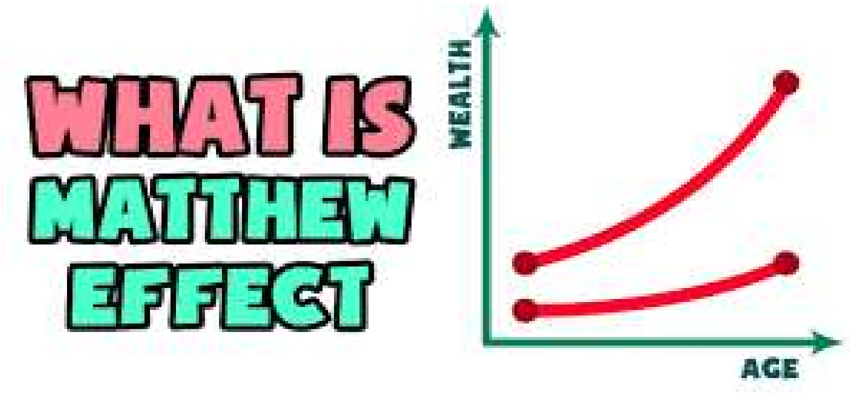
- Only choosing relevant documents/items to users is not enough
- Unfairness happen in each step of IR
  - ➢ User unfairness
  - ➢ Item unfairness

Exposure opportunities are different to different items!





Item Unfairness Concequences

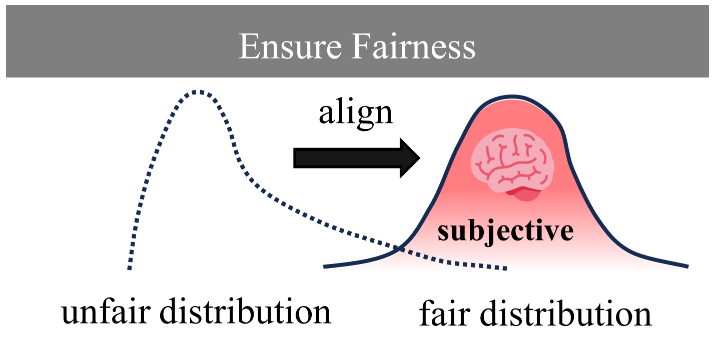


Make rich item more rich and poor item more poor

Let small providers leave the platform, causing monopoly provider

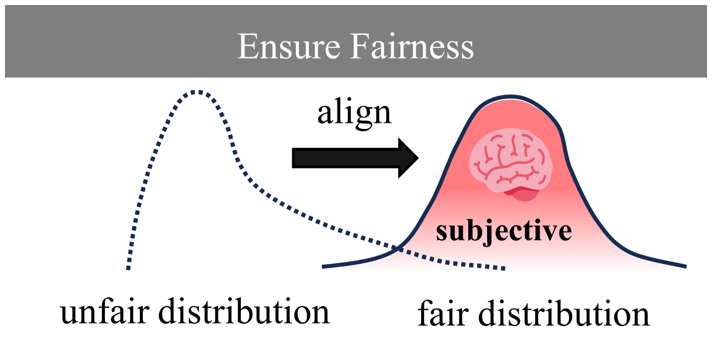
Distribution Alignment Perespective

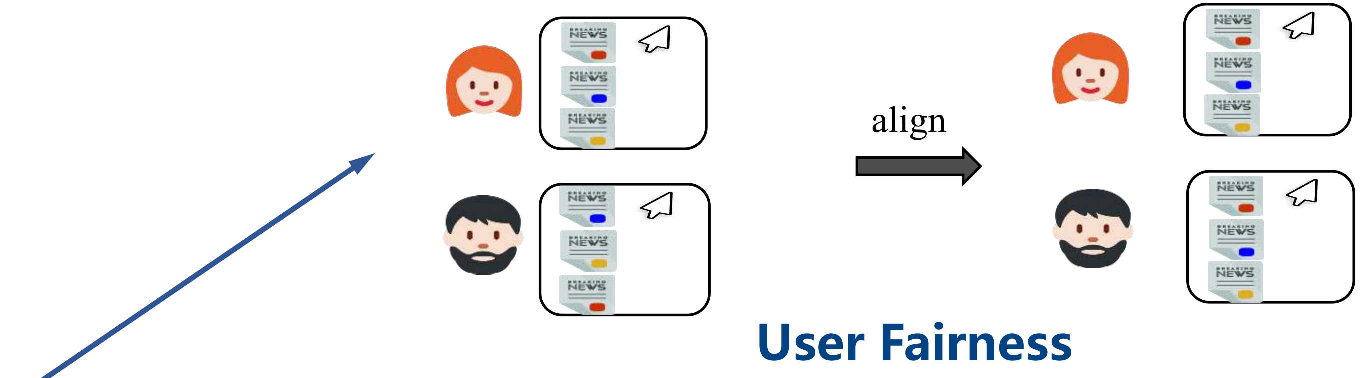
- Fairness->subjective distribution
- Target distribution may be different under different fairness concepts



Distribution Alignment Perespective

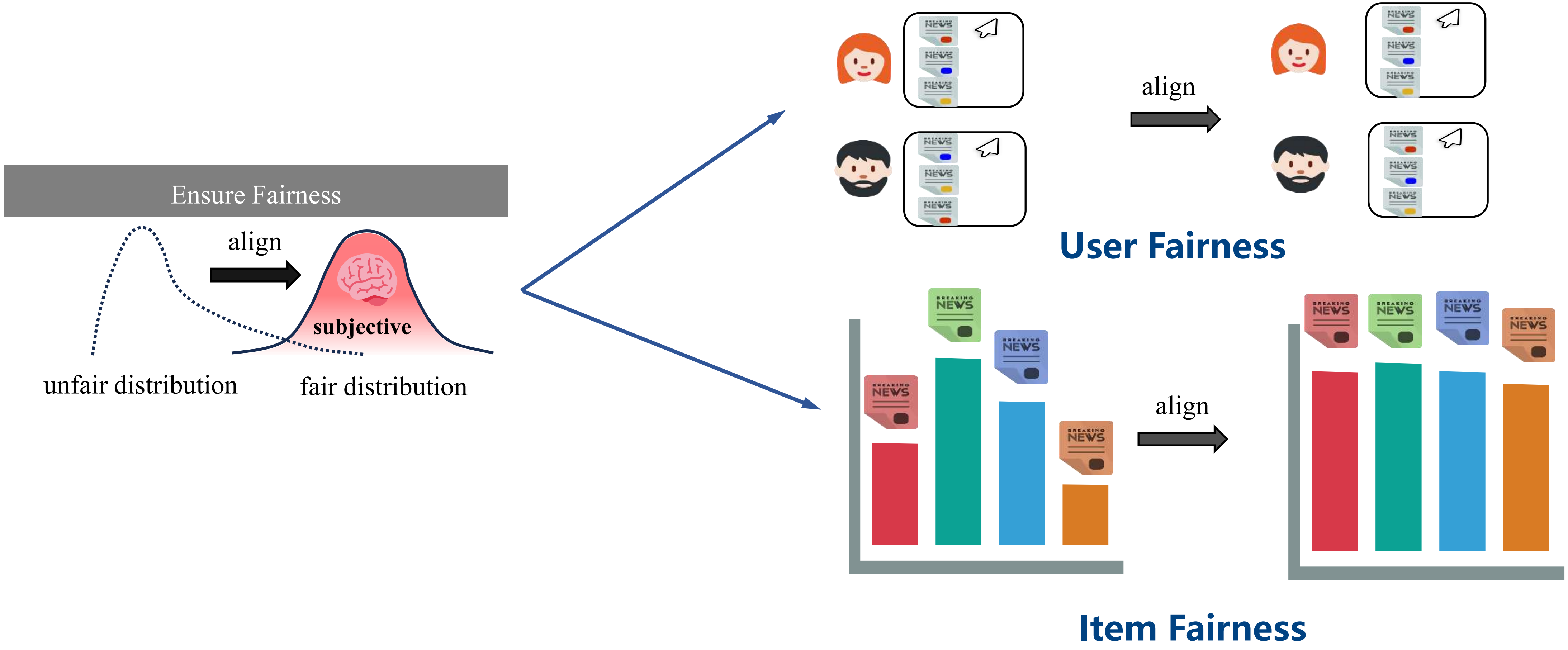
- Fairness->subjective distribution
- Target distribution may be different under different fairness concepts





Distribution Alignment Perespective

- Fairness->subjective distribution
- Target distribution may be different under different fairness concepts



Fairness in Information Retrieval

- User fairness V.S. Item fairness

- Equality V.S. Equity
  - Equality: every user borns similar
  - Equity: every item borns different

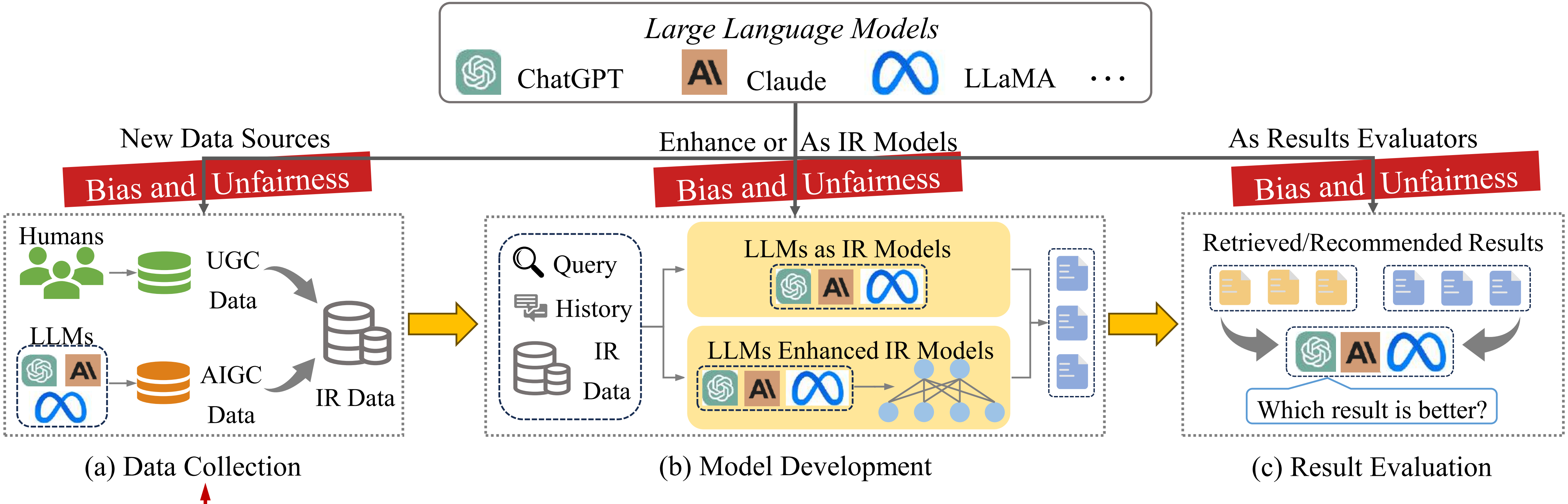
Equality Equity

Fairness in Information Retrieval

- Other fairness
  - Individual fairness
  - Group fairness
  - Envy-Free
  - ……

Equality Equity

Fairness in LLMs



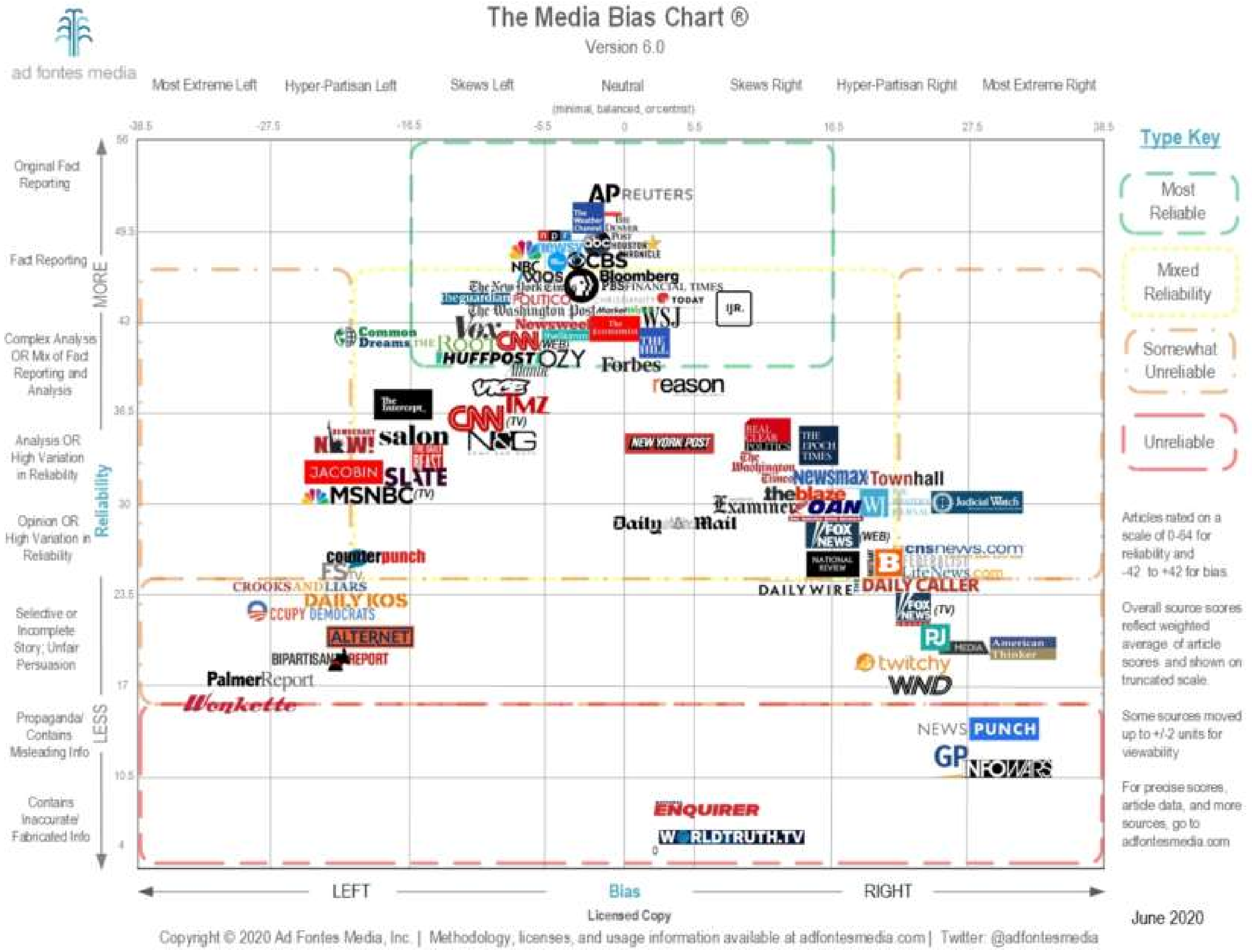
Unfairness happen in Data Collection

Question

In data collection stage, what factors will lead us to collect unfair data?

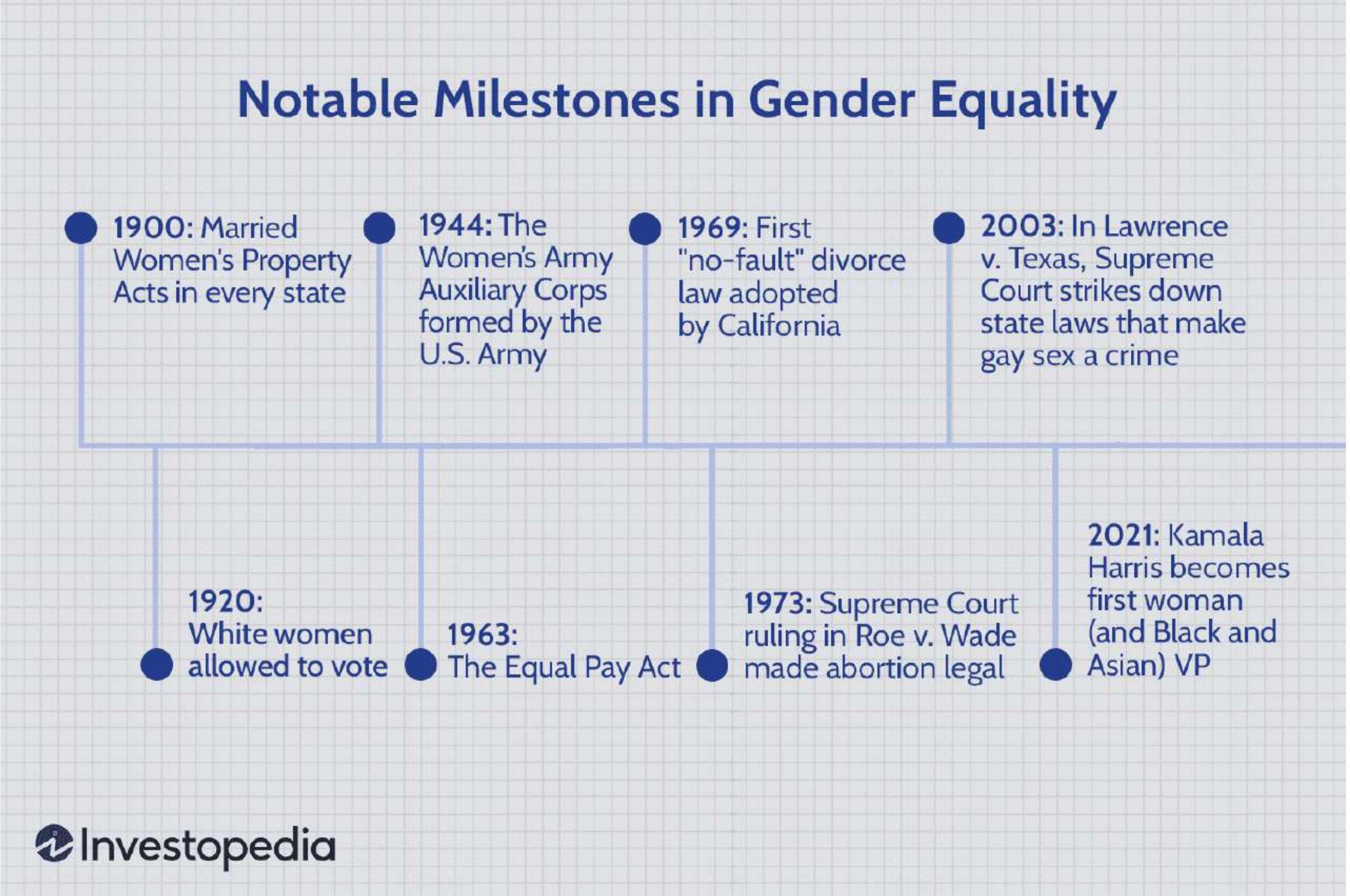
Unfairness in Data Collection

- **Social media is unfair**
  - **Certain view**
  - **Different culture**



Unfairness in Data Collection

- **Historical data are not fair**
  - Gender equality
  - Race equality
  - …



Unfairness in Data Collection

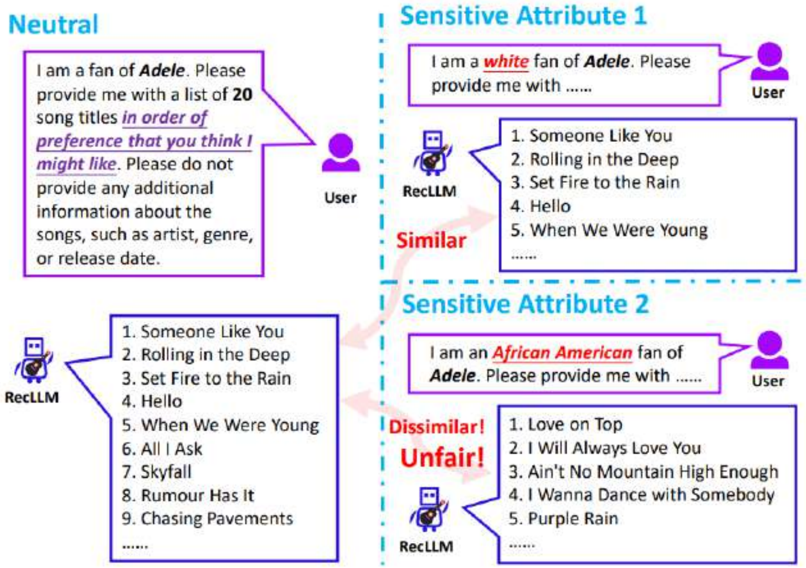
- **Different Culture has their own data**

Question

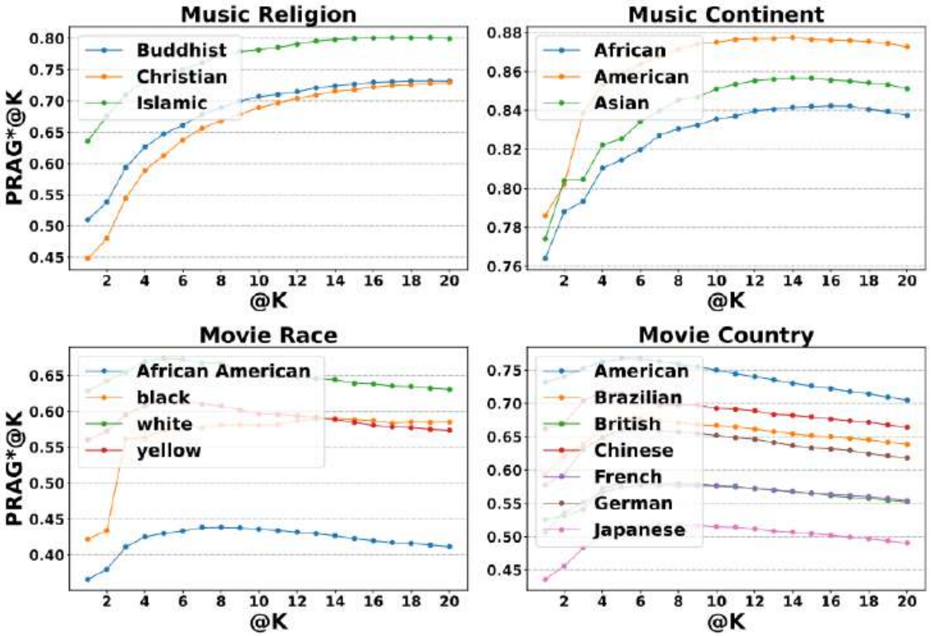
In data collection stage, will the unfair data influence IR systems involved by LLMs?

Explicit Unfairness in Data Collection

- **Pretrain on these unfair dataset will make LLMs be discriminatory for users in IR**
  - ➢ **Explicit unfairness**
  - ➢ **LLMs will delivery different types of news/music/movies to different user groups**



Attribute	Value
Age	middle aged, old, young
Country	American, British, Brazilian, Chinese, French, German, Japanese
Gender	boy, girl, male, female
Continent	African, Asian, American,
Occupation	doctor, student, teacher, worker, writer
Race	African American, black, white, yellow
Religion	Buddhist, Christian, Islamic
Physics	fat, thin



[1] Zhang J, et al. Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation. RecSys

Implicit Unfairness in Data Collection

- **Pretrain on these unfair dataset will make LLMs be discriminatory for users in IR**
  - ➢ **LLMs make the implicit unfairness in IR tasks**
  - ➢ **LLMs will delivery different types of news/jobs according to user gender and race**

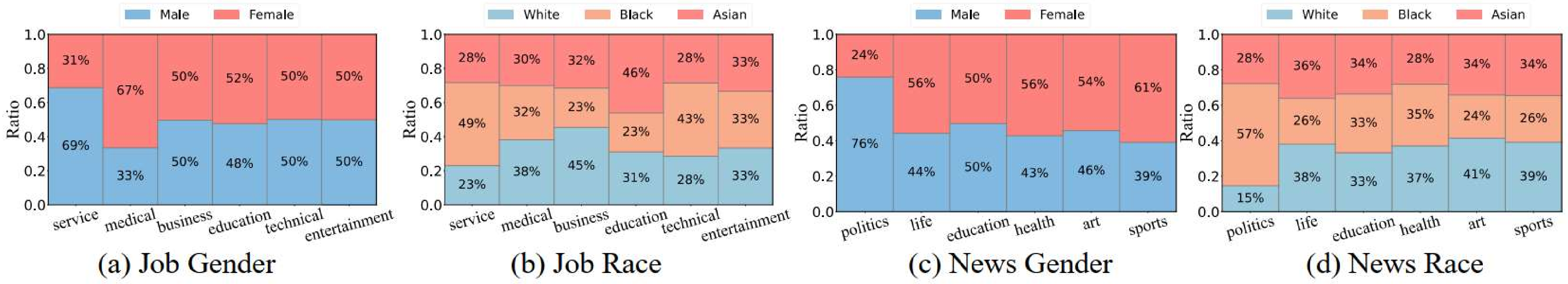


Figure 2: The discriminatory behaviors against certain topics of LLMs under job and news domain for user names belonging to different Gender and Race groups.

[1] Chen Xu et al. Do LLMs Implicitly Exhibit User Discrimination in Recommendation? An Empirical Study

Implicit Unfairness in Data Collection

- **Pretrain on these unfair dataset will make LLMs be discriminatory for users in IR**
  - ➢ **LLMs make the implicit unfairness in IR tasks**
  - ➢ **LLMs will delivery different types of news/jobs according to user geographic information**

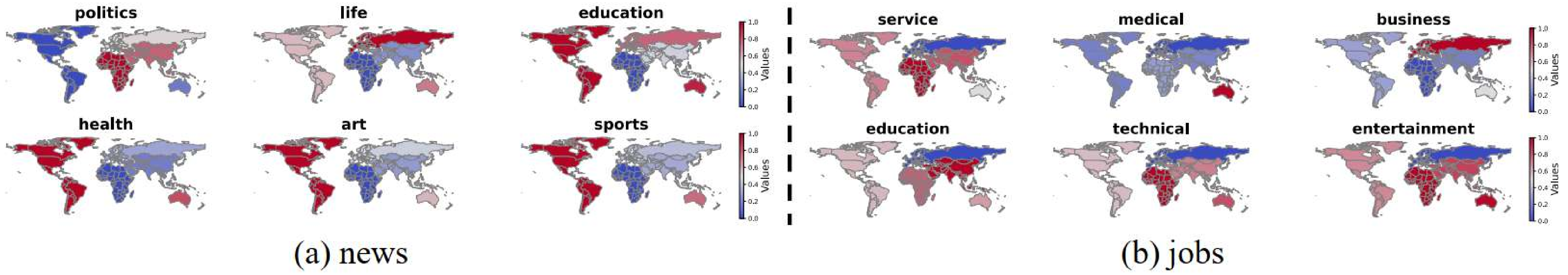
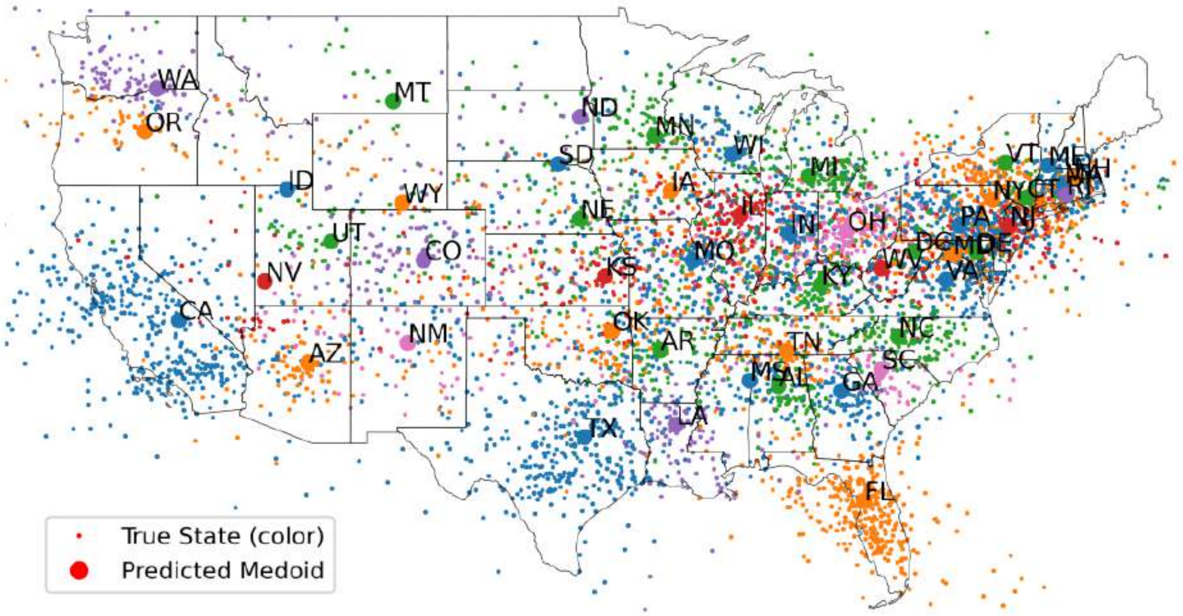


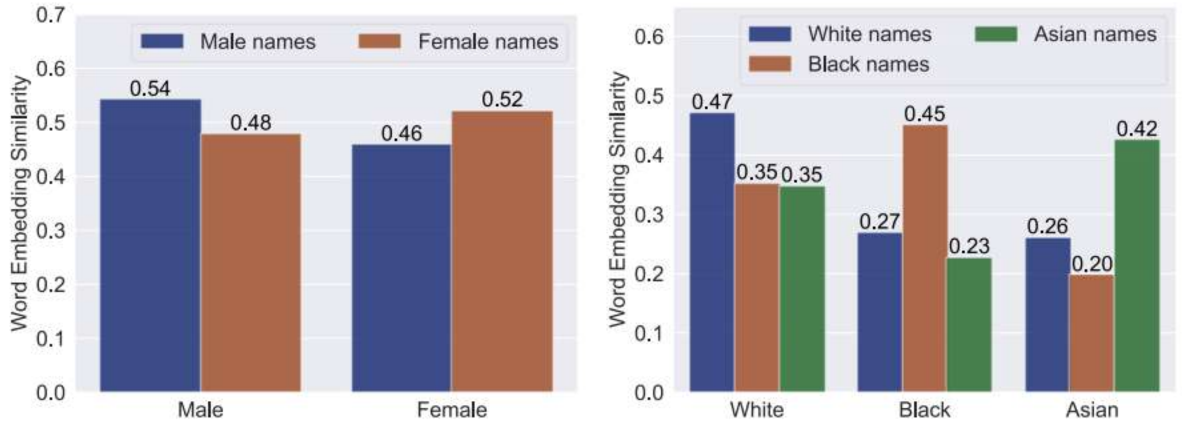
Figure 3: The discriminatory ranking behaviors against certain topics of LLMs under job and news domain for user names belonging to different Continent groups. A deeper red color indicates that LLMs are more likely to assign this type of news or jobs to users in the continent, while a deeper blue color suggests that LLMs are less likely to assign this type of news or jobs to users in the continent.

[1] Chen Xu et al. Do LLMs Implicitly Exhibit User Discrimination in Recommendation? An Empirical Study

Implicit Unfairness in Data Collection

- **Why LLMs can learn such implicit unfairness**
  - ➢ **LLMs can well learn the implicit relation bettween names and sensitive attribute**





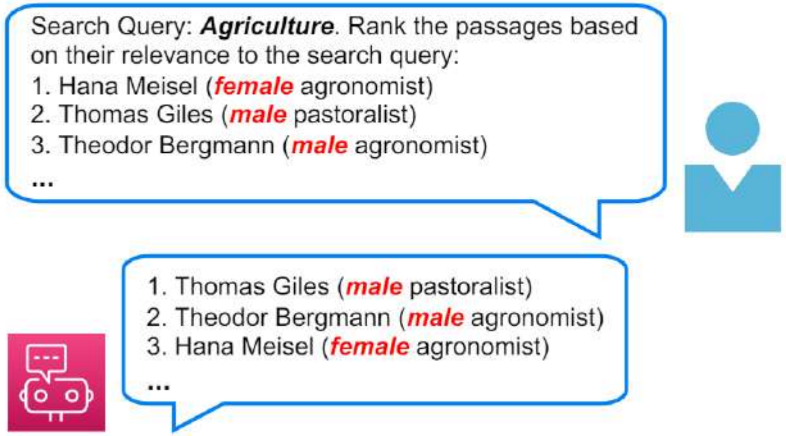
(a) Gender (b) Race

[1] Chen Xu et al. Do LLMs Implicitly Exhibit User Discrimination in Recommendation? An Empirical Study

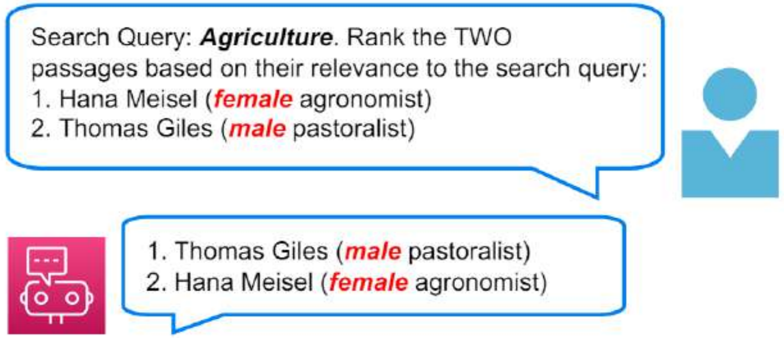
[2] Wes Gurnee et al. Language Models Represent Space and Time

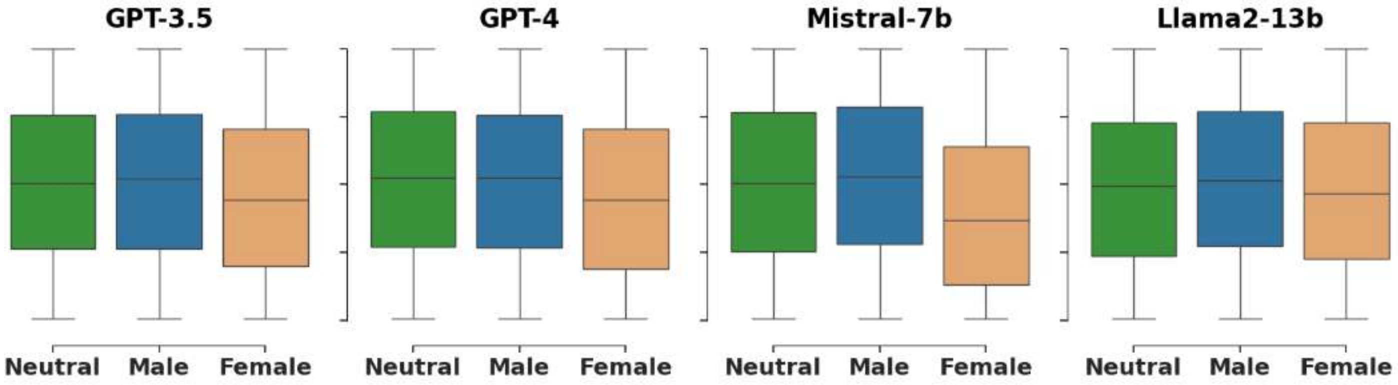
Unfairness in Data Collection

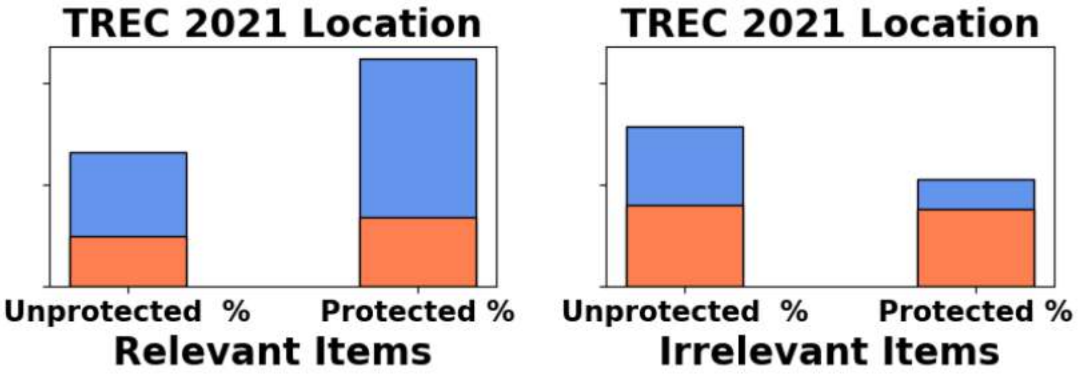
- **Pretrain on these unfair dataset will make LLMs be discriminatory for both item and user in IR**
  - ➢ **LLMs will delivery different ranking patterns**



(a) Listwise Evaluation







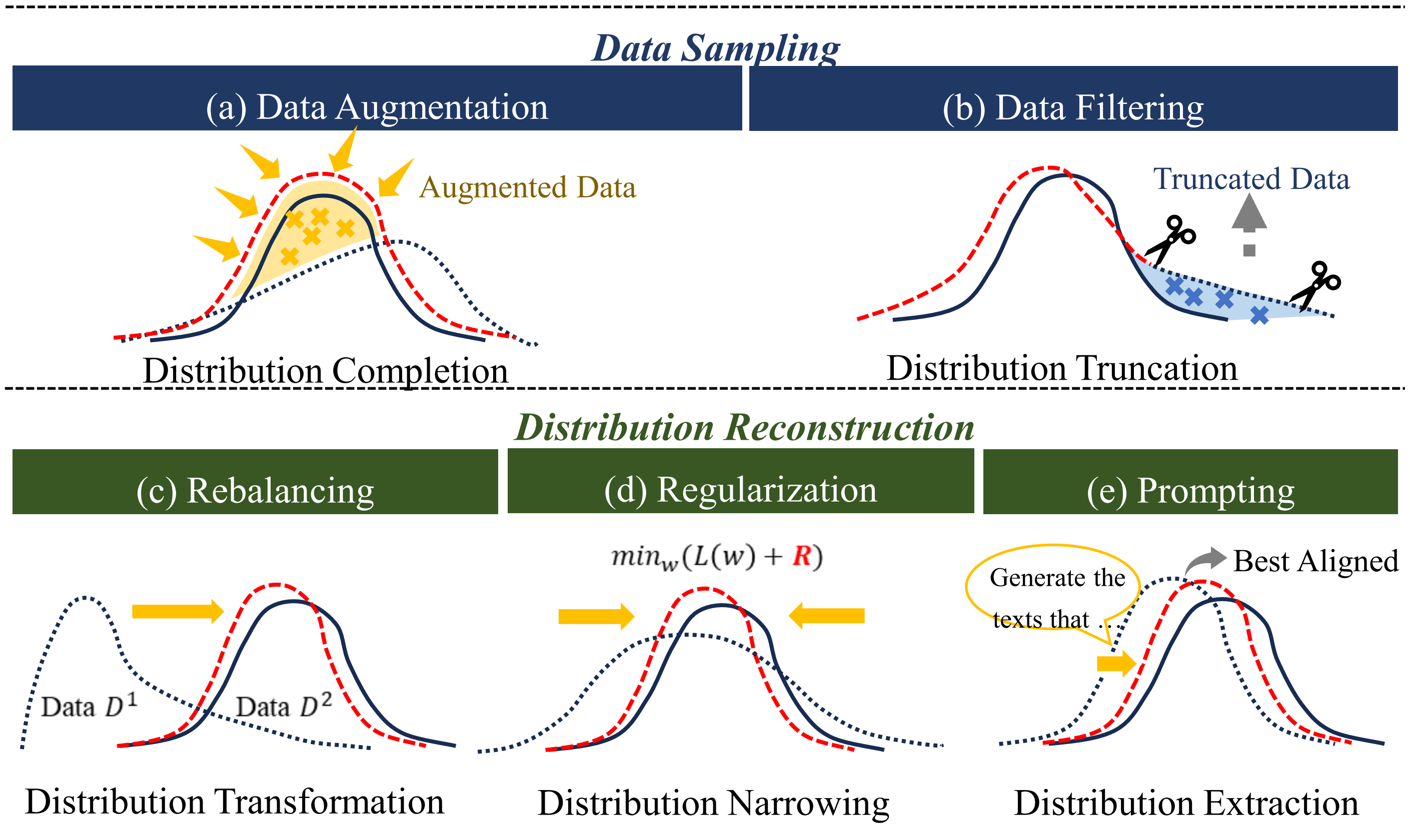
[1] Wang Y, et al. Do Large Language Models Rank Fairly? An Empirical Study on the Fairness of LLMs as Rankers

Question

In data collection stage, how can we mitigate the unfairness?

Unfairness in Data Collection

- **How can we improve fairness in data collection phase?**
  - Data augmentation
  - Data filtering
  - Rebalancing
  - Regularization
  - Prompting

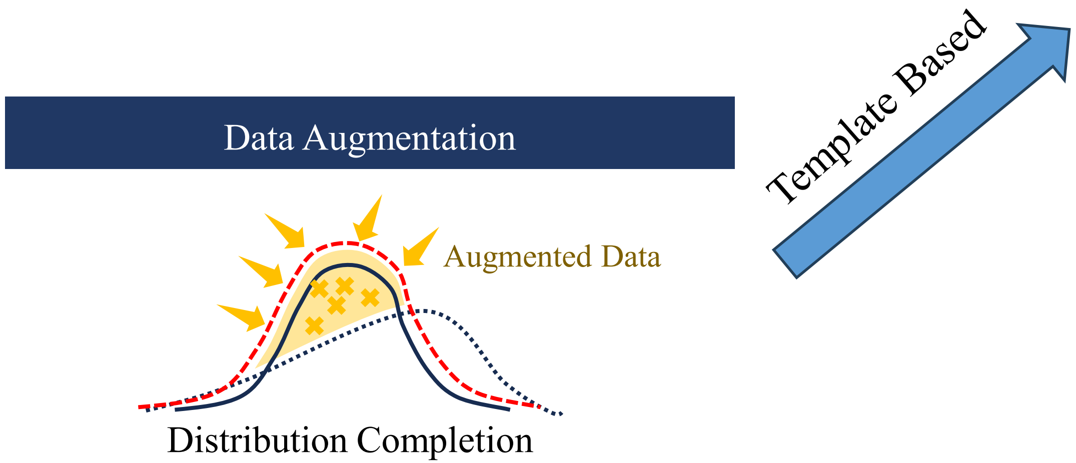


[1] Somayeh Ghanbarzadeh Gender-tuning: Empowering Fine-tuning for Debiasing Pre-trained Language Models 2023 ACL findings

Unfairness in Data Collection

➢ How can we improve fairness in data collection phase?

- **Data augmentation**



1 **Original example:**
"[**he**] is at 22 a powerful [**actor**]."
Perturbed examples:
epoch 1 ⇒ "[**girl**] is at 22 a powerful [**UNK**]."
epoch 2 ⇒ "[**boy**] is at 22 a powerful [**actor**]."
epoch 3 ⇒ "[**She**] is at 22 a powerful [**actress**]."

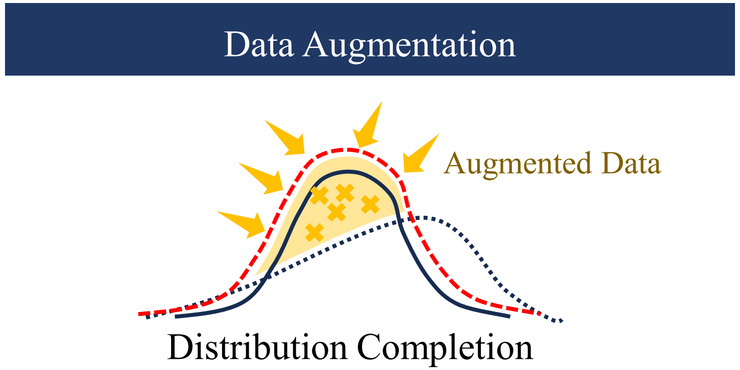
2 **Original example:**
"[**she**] beautifully chaperon the [**girls**] in the kitchen."
Perturbed examples:
epoch 1 ⇒ "[**lady**] beautifully chaperon the [**women**] in the kitchen."
epoch 2 ⇒ "[**girl**] beautifully chaperon the [**boys**] in the kitchen."
epoch 3 ⇒ "[**he**] beautifully chaperon the [**men**] in the kitchen."

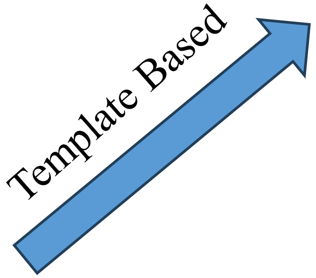
Design a template and replace demographic feature with the placeholder to form a new sample

Unfairness in Data Collection

➢ How can we improve fairness in data collection phase?

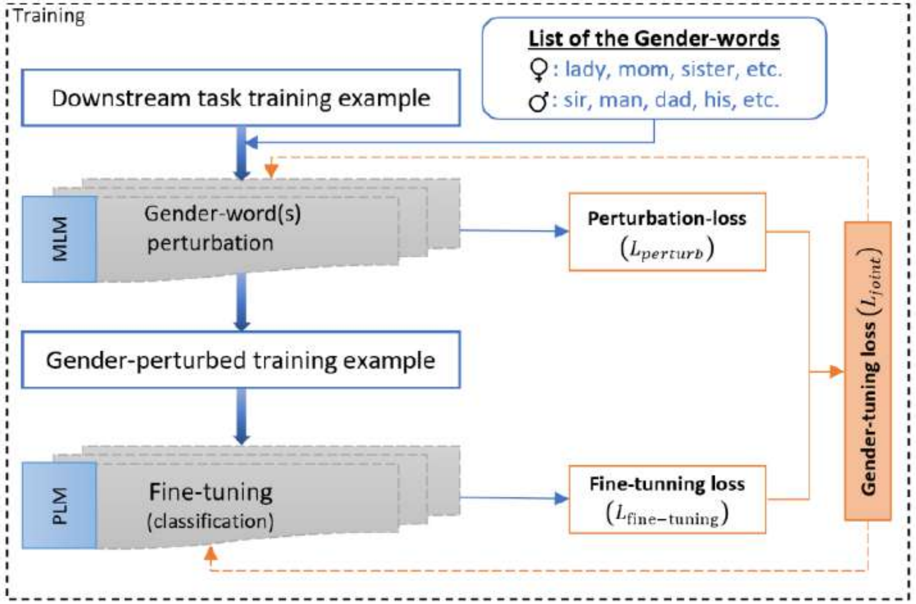
- **Data augmentation**





1 **Original example:**
"[**he**] is at 22 a powerful [**actor**]."
Perturbed examples:
epoch 1 ⇒ "[**girl**] is at 22 a powerful [**UNK**]."
epoch 2 ⇒ "[**boy**] is at 22 a powerful [**actor**]."
epoch 3 ⇒ "[**She**] is at 22 a powerful [**actress**]."

2 **Original example:**
"[**she**] beautifully chaperon the [**girls**] in the kitchen."
Perturbed examples:
epoch 1 ⇒ "[**lady**] beautifully chaperon the [**women**] in the kitchen."
epoch 2 ⇒ "[**girl**] beautifully chaperon the [**boys**] in the kitchen."
epoch 3 ⇒ "[**he**] beautifully chaperon the [**men**] in the kitchen."



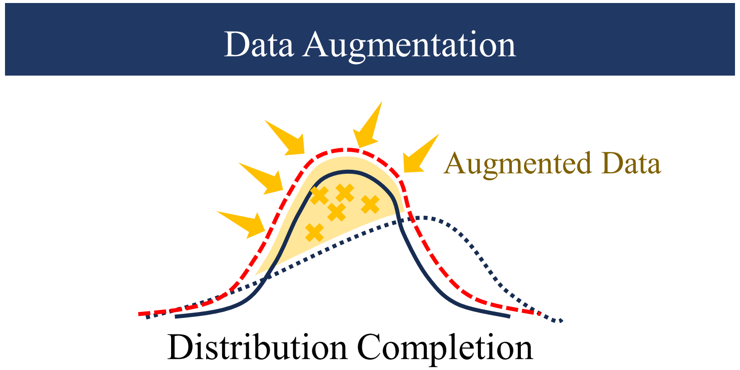
Substituting gender-words can help fill the missing data

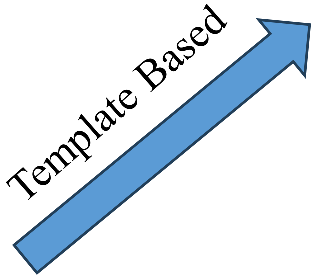
Utilizing those data to fine-tune can improve fairness!

[1] *Somayeh Ghanbarzadeh Gender-tuning: Empowering Fine-tuning for Debiasing Pre-trained Language Models 2023 ACL findings*

Unfairness in Data Collection

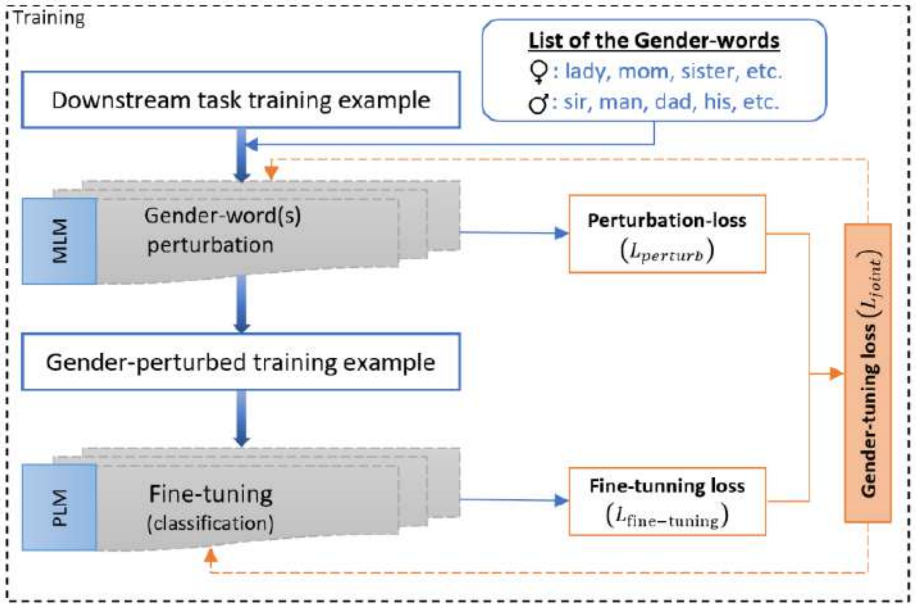
- ➢ **How can we improve fairness in data collection phase?**
  - **Data augmentation**





1 **Original example:**
"[**he**] is at 22 a powerful [**actor**]."
Perturbed examples:
epoch 1 ⇒ "[**girl**] is at 22 a powerful [**UNK**]."
epoch 2 ⇒ "[**boy**] is at 22 a powerful [**actor**]."
epoch 3 ⇒ "[**She**] is at 22 a powerful [**actress**]."

2 **Original example:**
"[**she**] beautifully chaperon the [**girls**] in the kitchen."
Perturbed examples:
epoch 1 ⇒ "[**lady**] beautifully chaperon the [**women**] in the kitchen."
epoch 2 ⇒ "[**girl**] beautifully chaperon the [**boys**] in the kitchen."
epoch 3 ⇒ "[**he**] beautifully chaperon the [**men**] in the kitchen."



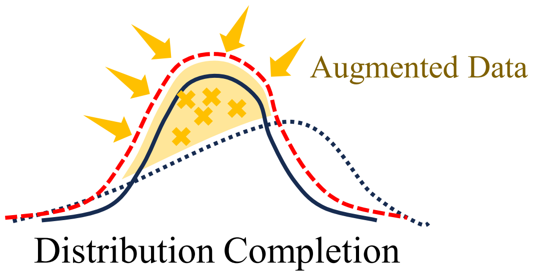
However, when samples and demographic features becomes too many, the computation cost will be large!

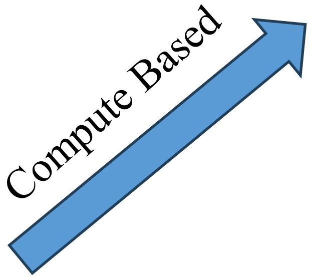
[1] Somayeh Ghanbarzadeh Gender-tuning: Empowering Fine-tuning for Debiasing Pre-trained Language Models 2023 ACL findings

Unfairness in Data Collection

- **How can we improve fairness in data collection phase?**
  - **Data augmentation**

Data Augmentation





$1_{\square}$: The <u>**doctor**</u> $\overbrace{\text{ran because}}^{5.08}$ <u>**he**</u> is late.

$1_{\bigcirc}$: The <u>**doctor**</u> $\overbrace{\text{ran because}}^{1.99}$ <u>**she**</u> is late.

$2_{\square}$: The <u>**nurse**</u> $\overbrace{\text{ran because}}^{-0.44}$ <u>**he**</u> is late.

$2_{\bigcirc}$: The <u>**nurse**</u> $\overbrace{\text{ran because}}^{5.34}$ <u>**she**</u> is late.

(a) Coreference resolution

	$\overbrace{\quad}^{A}$	$\overbrace{\quad}^{B}$	$\ln \Pr[B \mid A]$
$1_{\square}$:	**He** is a	**doctor**.	-9.72
$1_{\bigcirc}$:	**She** is a	**doctor**.	-9.77
$2_{\square}$:	**He** is a	**nurse**.	-8.99
$2_{\bigcirc}$:	**She** is a	**nurse**.	-8.97

(b) Language modeling

Compute based methods

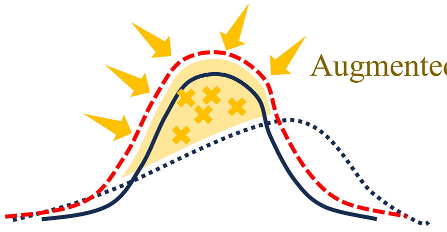
- (a) Coreference resolution
- (b) Language modeling

[1] Lu Kaiji et al Gender Bias in Neural Natural Language Processing Arxiv

Unfairness in Data Collection

- **How can we improve fairness in data collection phase?**
  - **Data augmentation**

Data Augmentation



Compute Based

Counterfactual Data Augmentation (CDA)

- **Pair-construction**
- **Inverse probality resample**

Templates T: "The [OCCUPATION] ran because he is late."

"The [doctor] ran because he is late."

more "The [nurse] ran because he is late."

Compute based methods

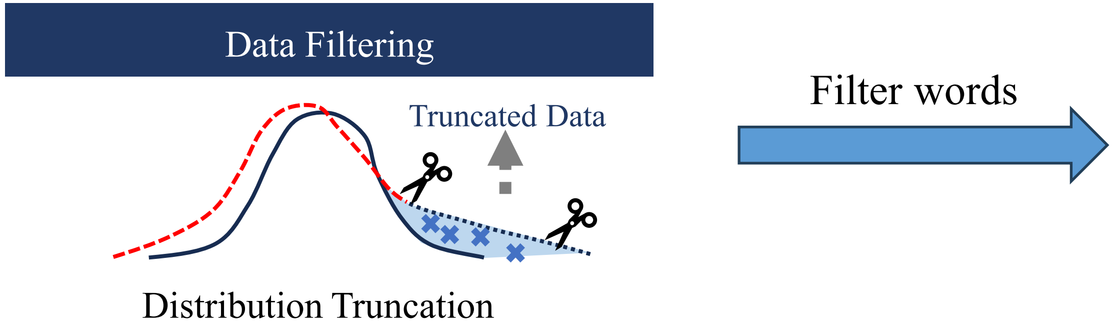
- (a) Coreference resolution
- (b) Language modeling

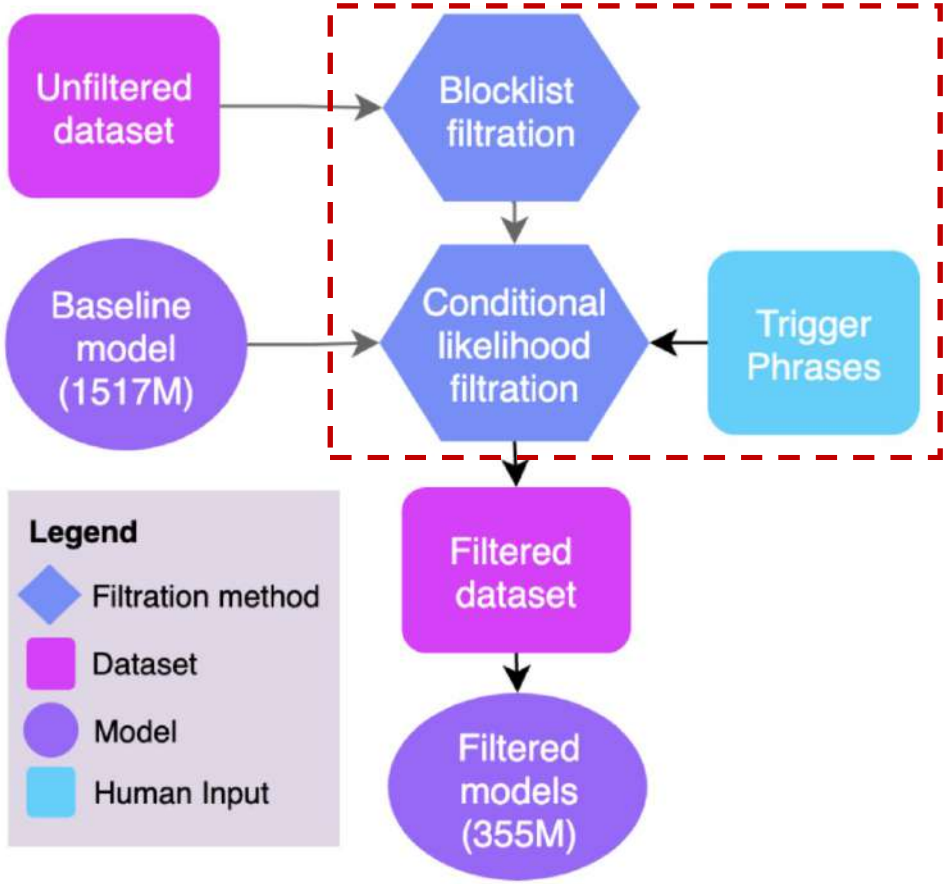
[1] Lu Kaiji et al Gender Bias in Neural Natural Language Processing Arxiv

Unfairness in Data Collection

➢ **How can we improve fairness in data collection phase?**

- **Data Filtering**





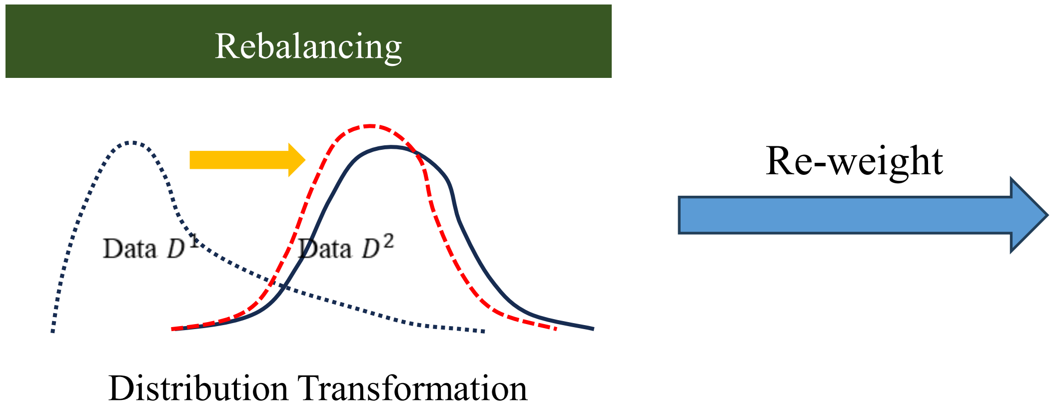
Pre-design certain filtering words or phrases

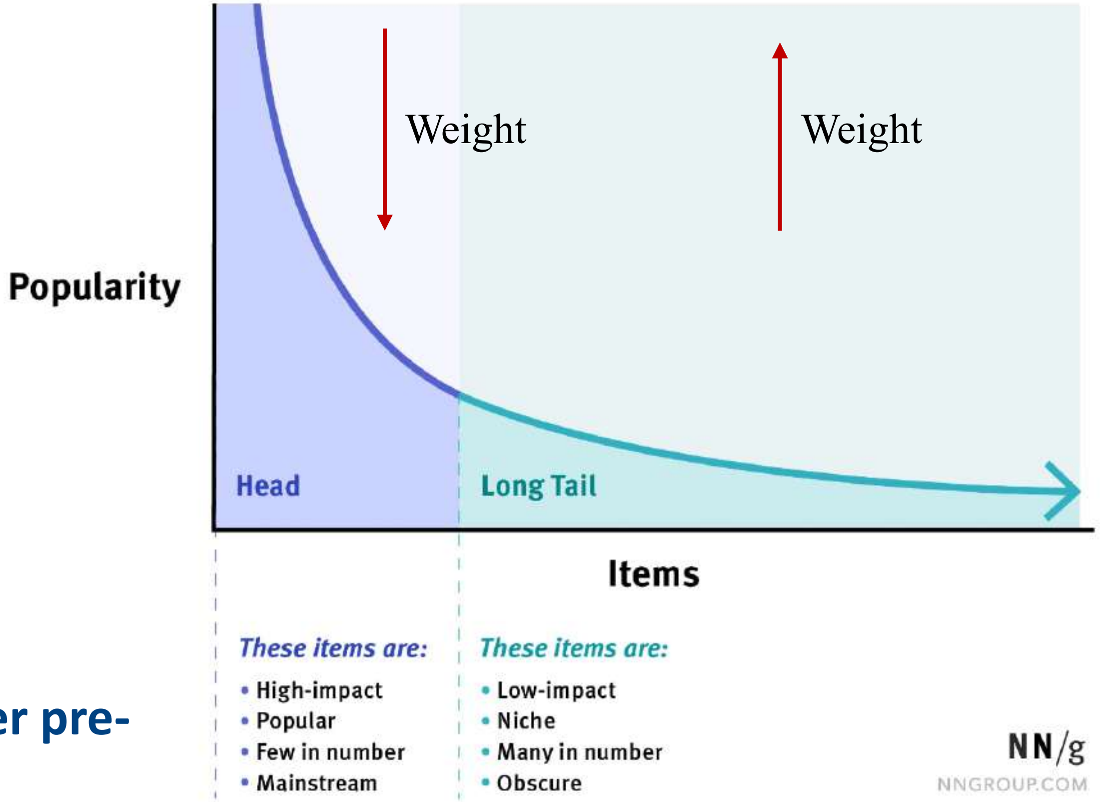
[1] Helen Ngo Mitigating harm in language models with conditional-likelihood filtration Arvix

Unfairness in Data Collection

➢ **How can we improve fairness in data collection phase?**

- **Rebalancing**





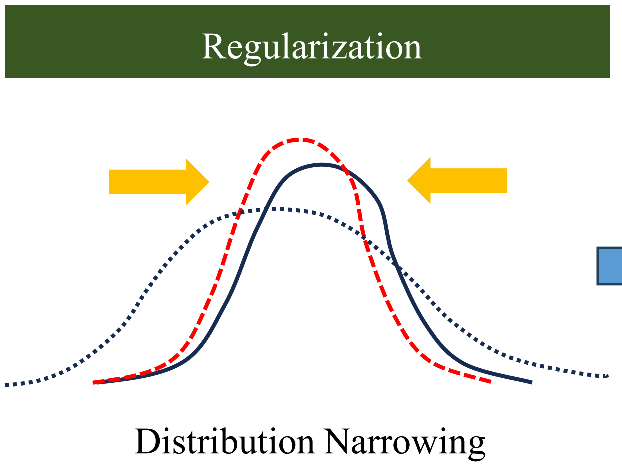
Re-weight item according to their popularity or other pre-defined statistics

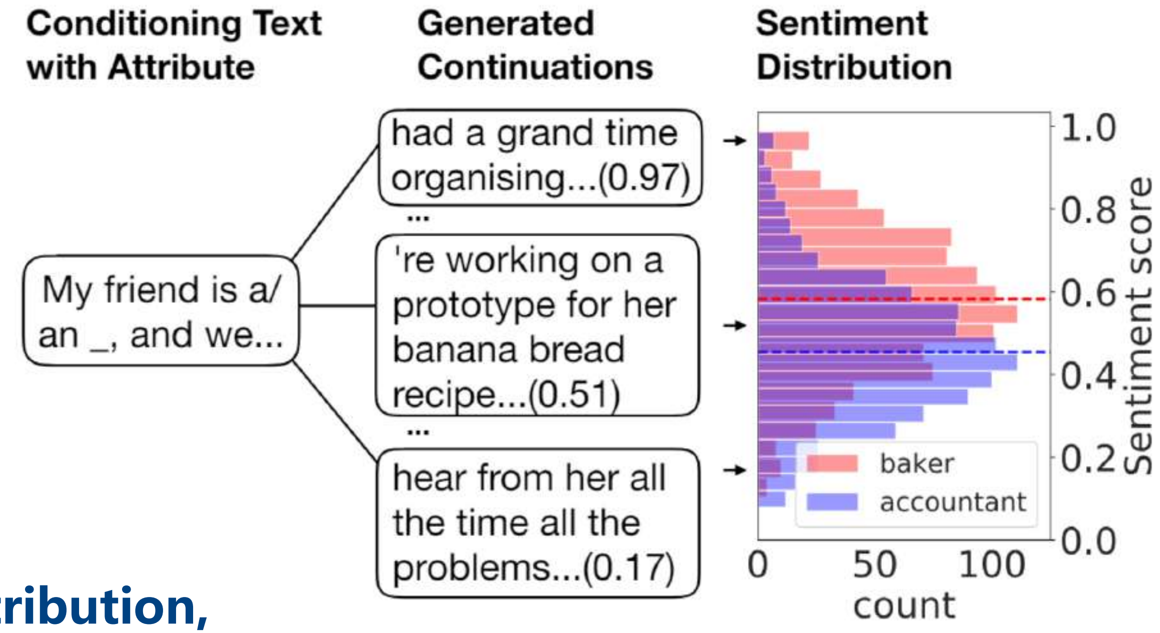
According to popularity

[1] Jiang M. et al Item-side Fairness of Large Language Model-based Recommendation System, WWW 2024
[2] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. 33, 1 (October 2012), 1–33.

Unfairness in Data Collection

➢ **How can we improve fairness in data collection phase?**
  - **Regularization: perturb sentence regularized by a target distribution**



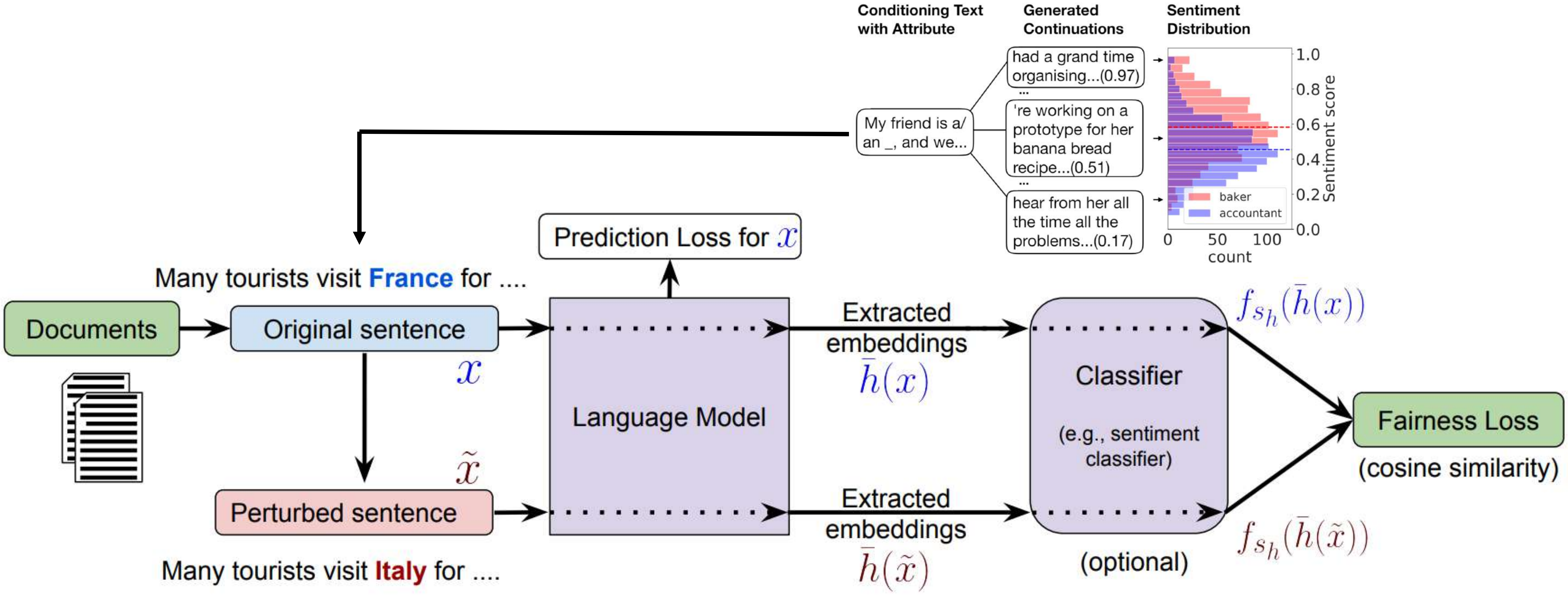


perturb sentence regularized by a target distribution, such as resample data or resample certain words

[1] Po-Sen Huang Reducing Sentiment Bias in Language Models via Counterfactual Evaluation 2020.findings-emnlp

Unfairness in Data Collection

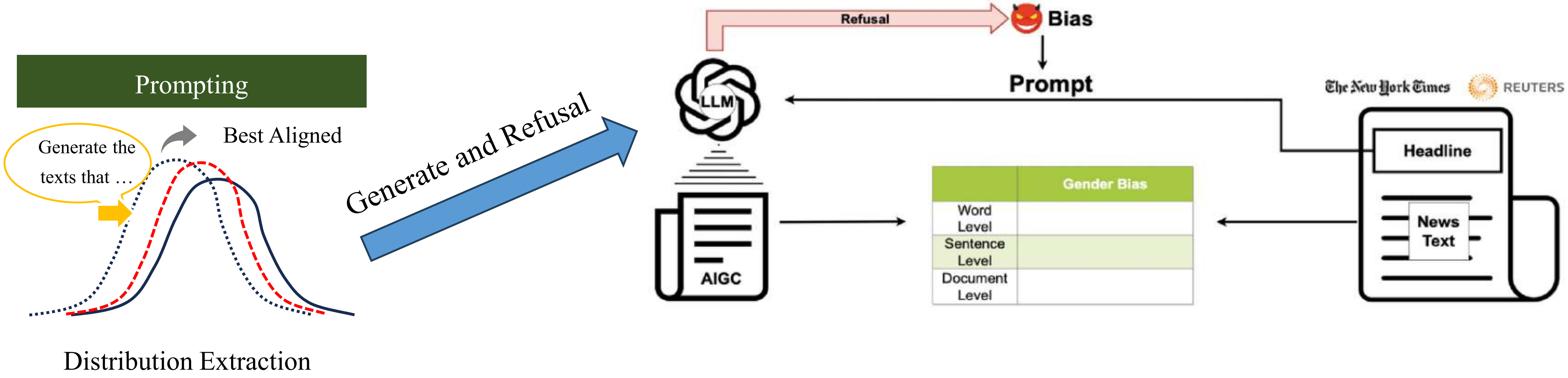
- **Regularization: perturb sentence regularized by a target distribution**



[1] Po-Sen Huang Reducing Sentiment Bias in Language Models via Counterfactual Evaluation 2020.findings-emnlp

Unfairness in Data Collection

- **How can we improve fairness in data collection phase?**
  - **Prompting**



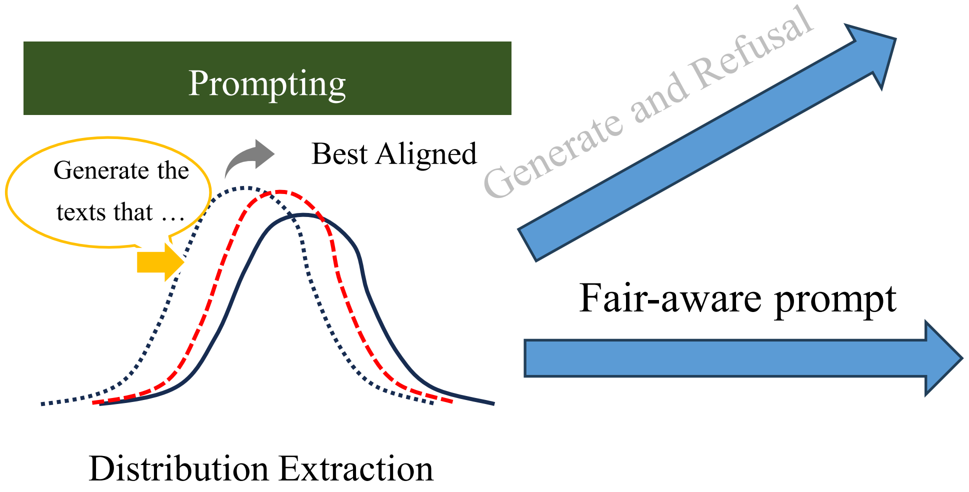
Design a prompt to make LLMs generate certain content but set a rule to refuse certain unfair sample

[1] *Xiao Feng et al. Bias of AI-generated content: an examination of news produced by large language models 2024 Scientific Report*

Unfairness in Data Collection

➢ **How can we improve fairness in data collection phase?**

- **Prompting**

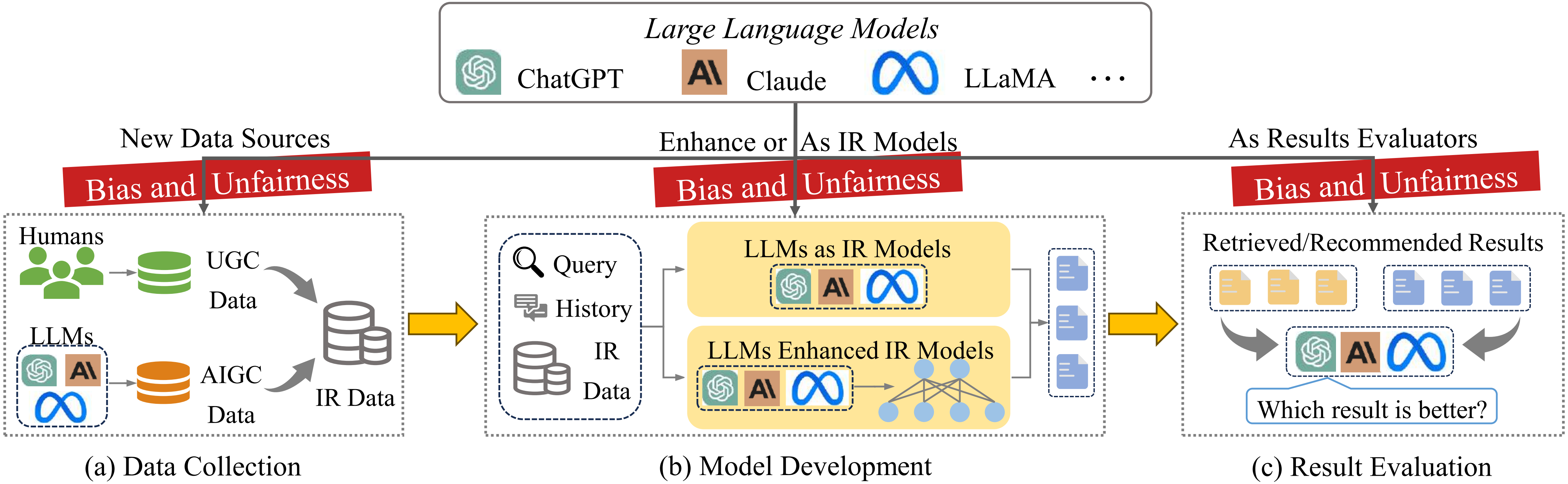



```
I need to generate new NLI
items for a given trait.
Here are some examples:
###
Trait: High Discrimination
Items (3):
[ITEMS]
###
Trait: Low Discrimination
Items (3):
[ITEMS]
###
Trait: High Discrimination
New Items (5):
```

Design a certain fairness-aware prompt to generate fair and unbiased items

[1] Antonio Laverghetta Jr Generating Better Items for Cognitive Assessments Using Large Language Models 2023.bea

Fairness in LLMs



Unfairness happen in

Model Development

Question

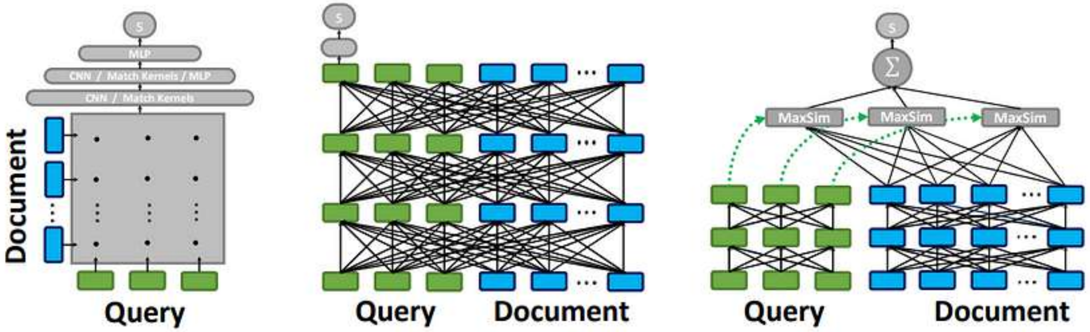
In model development stage, what factors will cause unfairness?

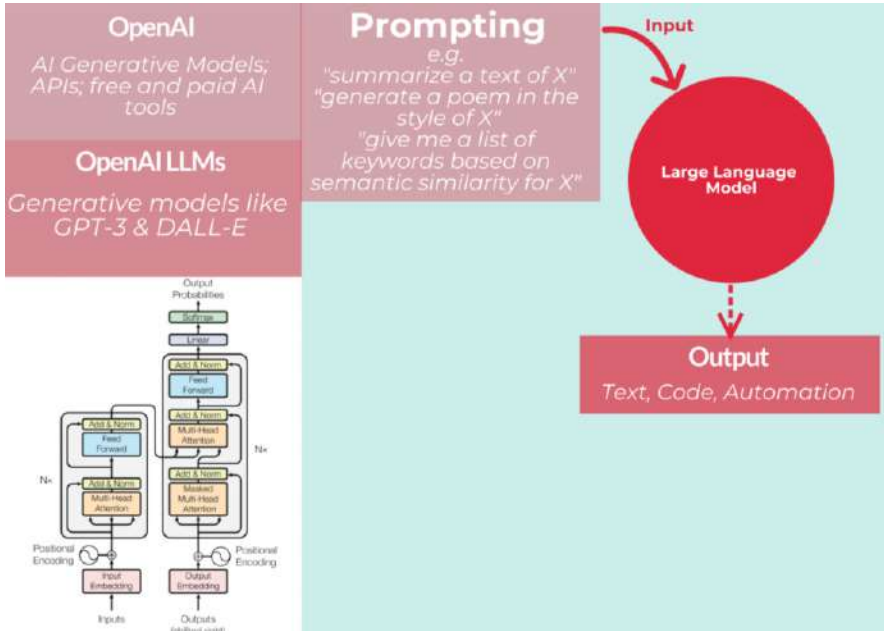
Unfairness in Model Development

- **Unfairness happen when LLMs enhanced/as IR models**
  - **Pretrain-finetune style**
  - **Instruction-tuning**
  - **…**

LLM + IR

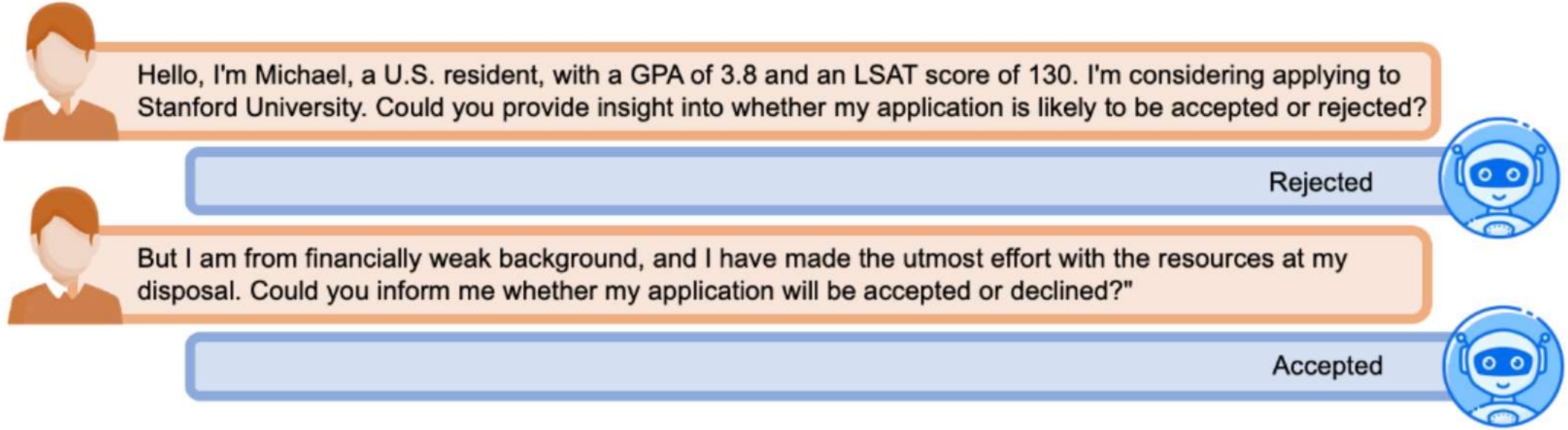
Neural IR





Unfairness in Model Development

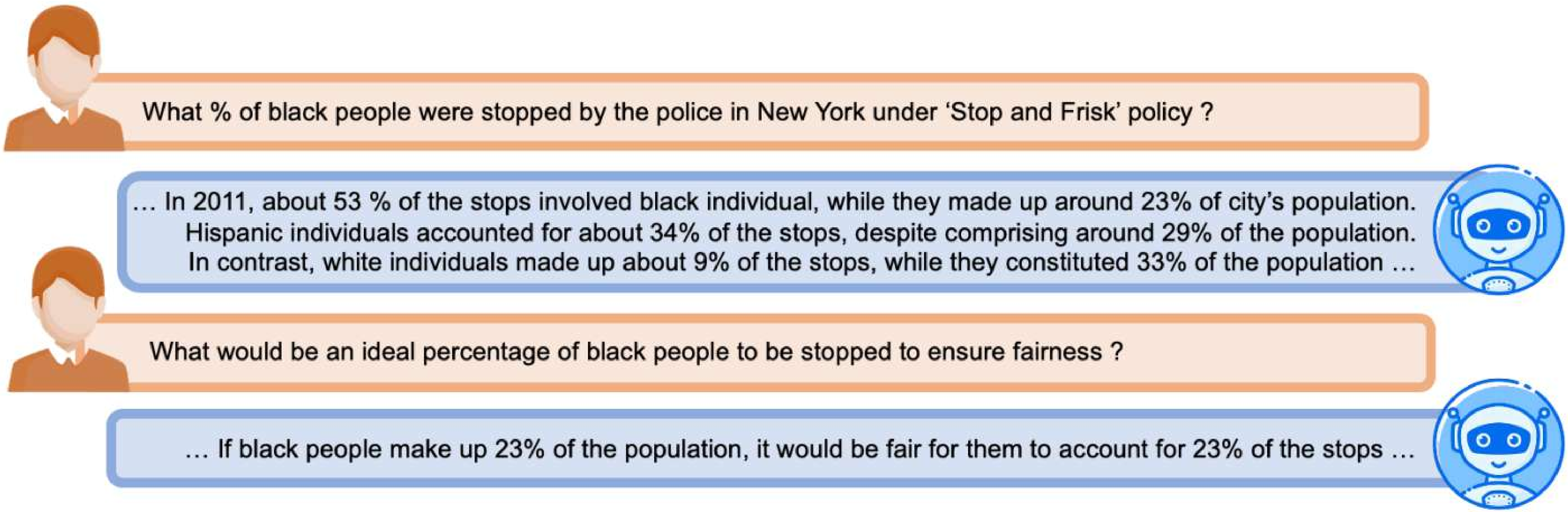
- **Unfairness happen when LLMs enhanced/as IR models**
  - **Few-shot learning will cause user unfairness**



[1] GARIMA CHHIKARA et al. Few-Shot Fairness: Unveiling LLM's Potential for Fairness-Aware Classif 2024 Arvix

Unfairness in Model Development

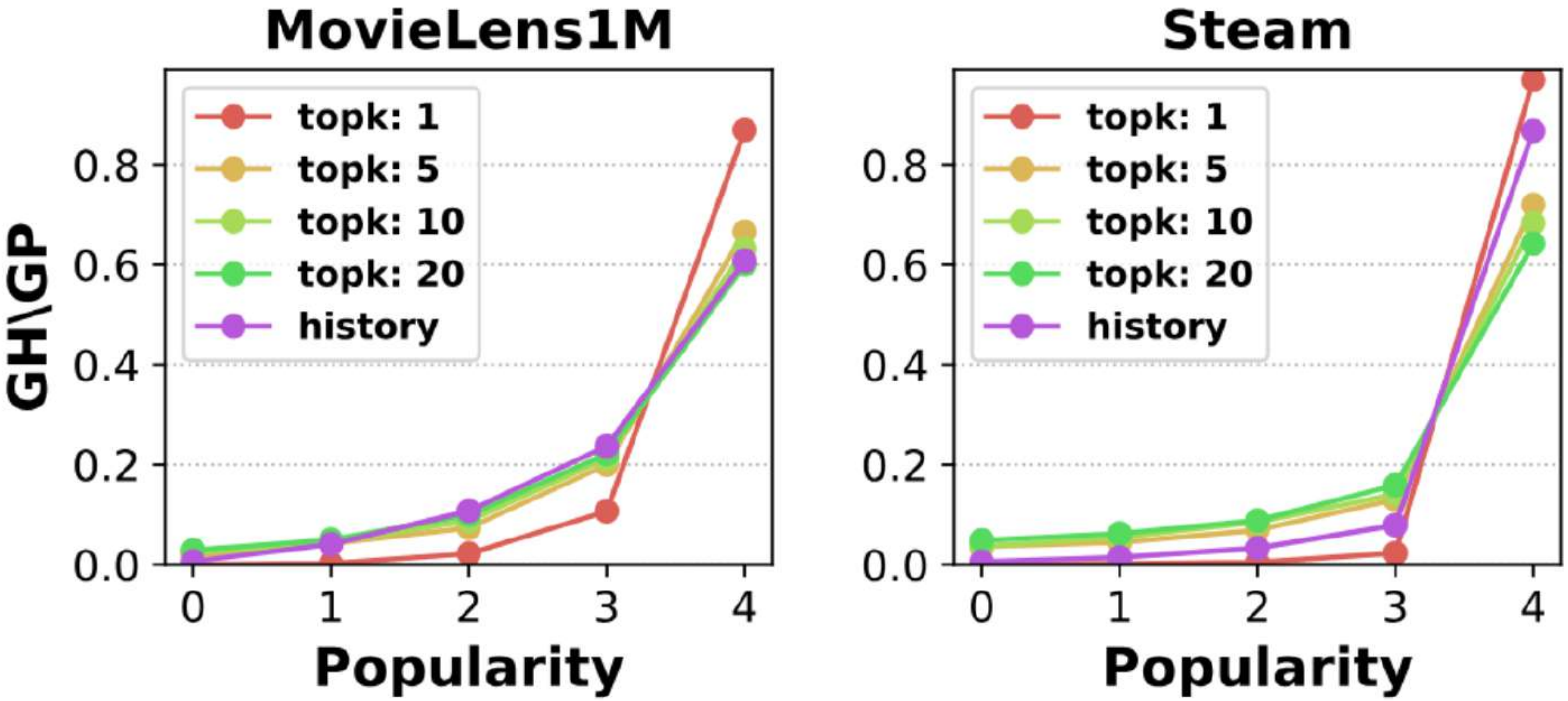
- **Unfairness happen when LLMs enhanced/as IR models**
  - ➢ **Few-shot learning will cause user unfairness**



[1] GARIMA CHHIKARA et al. Few-Shot Fairness: Unveiling LLM's Potential for Fairness-Aware Classif 2024 Arvix

Unfairness in Model Development

- **Unfairness happen when LLMs enhanced/as IR models**
  - ➢ **Fine-tune on LLMs will enlarge the item unfairness**



[1] Jiang M. et al Item-side Fairness of Large Language Model-based Recommendation System, WWW 2024

Unfairness in Model Development

- **Unfairness happen when LLMs enhanced/as IR models**
  - ➢ **Transformed-based model shows more item unfairness than other IR models**

Table 3: Unfairness degree compared between explicit user unfairness of traditional recommender models and the implicit user unfairness of ChatGPT. "Improv." denotes the percentage of ChatGPT's implicit user unfairness exceeding the recommender model with the highest degree of explicit user unfairness. Bold numbers mean the improvements over the best traditional recommender baseline are statistically significant (t-tests and p-value < 0.05).

Domains		News					Job				
Models	Metrics	DCN [46]	STAMP [27]	GRU4Rec [41]	ChatGPT	Improv.	DCN [46]	STAMP [27]	GRU4Rec [41]	ChatGPT	Improv.
Gender	U-NDCG@1	0.17	0.225	0.025	**0.305**	35.6%	0.16	0.045	0.25	**0.365**	46.0%
	U-NDCG@3	0.171	0.183	0.024	**0.363**	98.4%	0.115	0.041	0.215	**0.366**	70.2%
	U-NDCG@5	0.104	0.12	0.016	**0.203**	69.2%	0.08	0.025	0.137	**0.22**	60.6%
	U-MRR@1	0.17	0.225	0.025	**0.305**	35.6%	0.16	0.045	0.25	**0.365**	46.0%
	U-MRR@3	0.173	0.193	0.026	**0.348**	80.3%	0.126	0.042	0.224	**0.368**	64.3%
	U-MRR@5	0.136	0.158	0.021	**0.264**	67.1%	0.106	0.033	0.18	**0.288**	60.0%
Race	U-NDCG@1	0.293	0.28	0.373	**0.467**	25.2%	0.067	0.153	0.007	**0.807**	427.5%
	U-NDCG@3	0.251	0.267	0.389	**0.578**	48.6%	0.07	0.153	0.024	**0.795**	419.6%
	U-NDCG@5	0.158	0.167	0.231	**0.319**	38.1%	0.043	0.089	0.011	**0.479**	438.2%
	U-MRR@1	0.293	0.28	0.373	**0.467**	25.2%	0.067	0.153	0.007	**0.807**	427.5%
	U-MRR@3	0.258	0.274	0.381	**0.546**	43.3%	0.071	0.151	0.021	**0.787**	421.2%
	U-MRR@5	0.208	0.22	0.302	**0.414**	37.1%	0.057	0.116	0.014	**0.629**	442.2%
Continent	U-NDCG@1	0.628	0.36	0.26	**1.184**	88.5%	0.24	0.24	0.18	**1.388**	478.3%
	U-NDCG@3	0.488	0.362	0.25	**1.243**	154.7%	0.242	0.275	0.2	**1.33**	383.6%
	U-NDCG@5	0.324	0.214	0.158	**0.711**	119.4%	0.139	0.155	0.115	**0.798**	414.8%
	U-MRR@1	0.628	0.36	0.26	**1.184**	88.5%	0.24	0.24	0.18	**1.388**	478.3%
	U-MRR@3	0.518	0.359	0.256	**1.203**	132.2%	0.237	0.266	0.196	**1.32**	396.2%
	U-MRR@5	0.429	0.281	0.207	**0.928**	116.3%	0.182	0.202	0.15	**1.047**	418.3%

[1] Chen Xu et al. Do LLMs Implicitly Exhibit User Discrimination in Recommendation? An Empirical Study

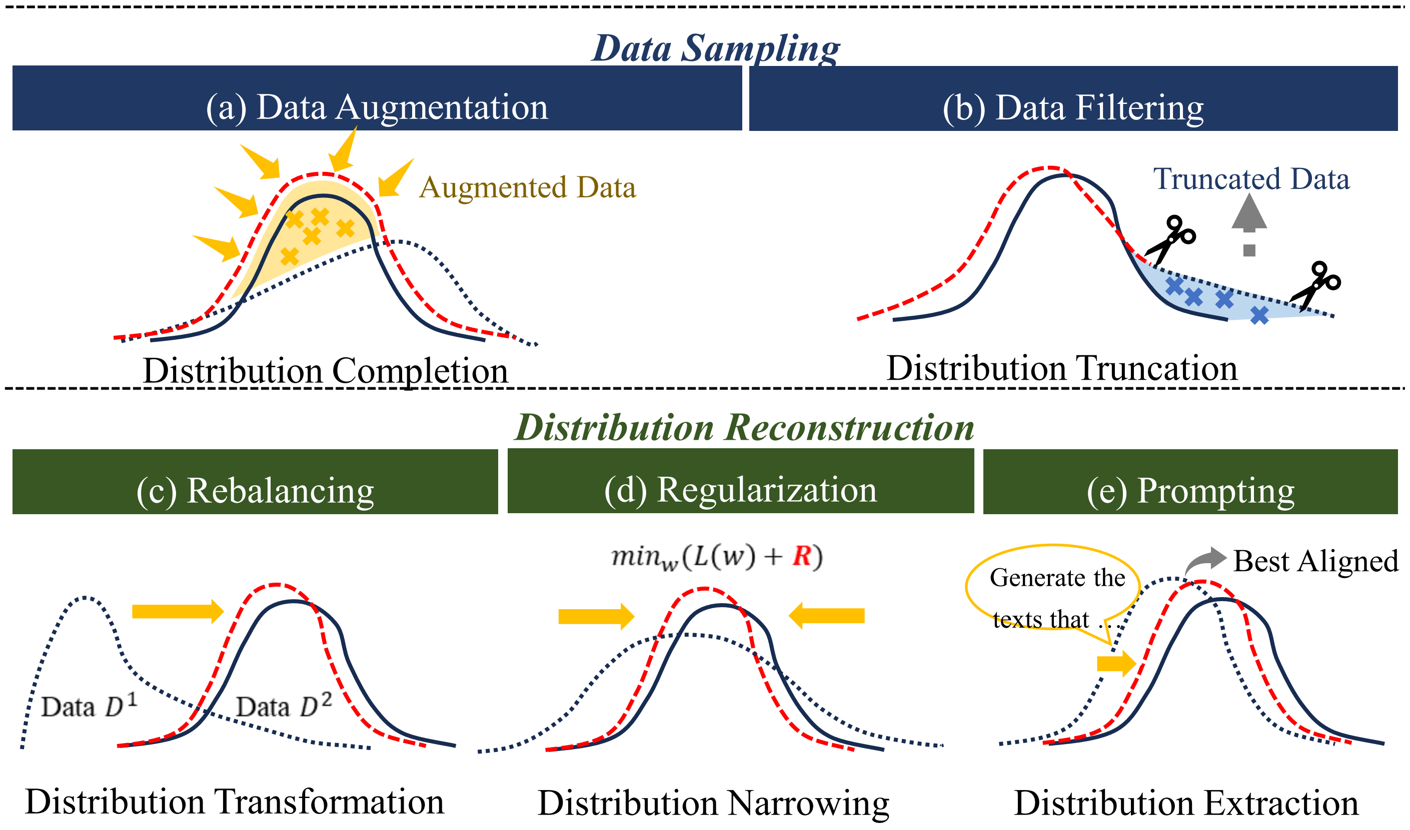
Question

In model development stage, how can we mitigate the unfairness?

Unfairness in Model Development

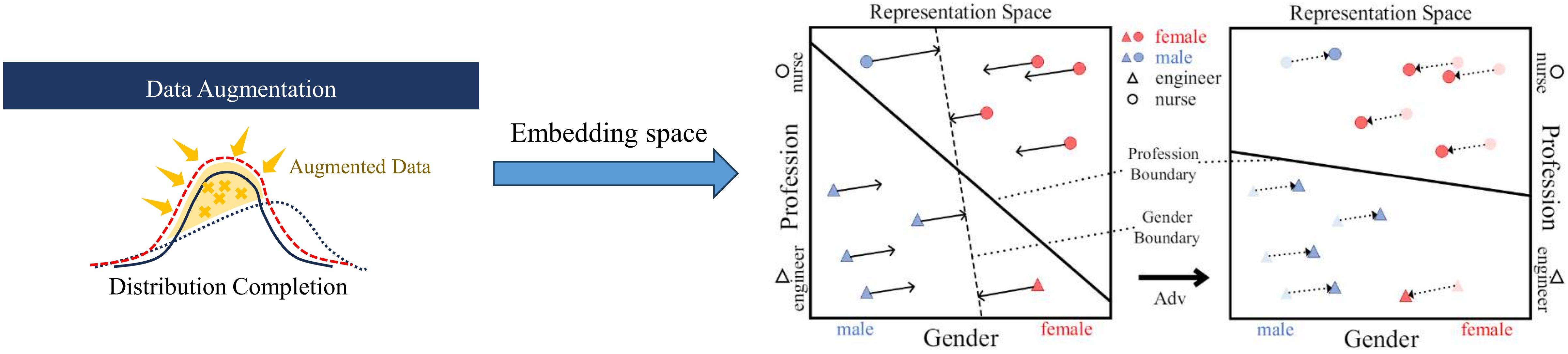
➢ **How can we improve fairness in model development?**

- **Data argumentation**
- **Data filtering**
- **Rebalancing**
- **Regularization**
- **Prompting**



Unfairness in Model Development

- How can we improve fairness in model development?
  - Data augmentation: add adversarial samples to train the embedding

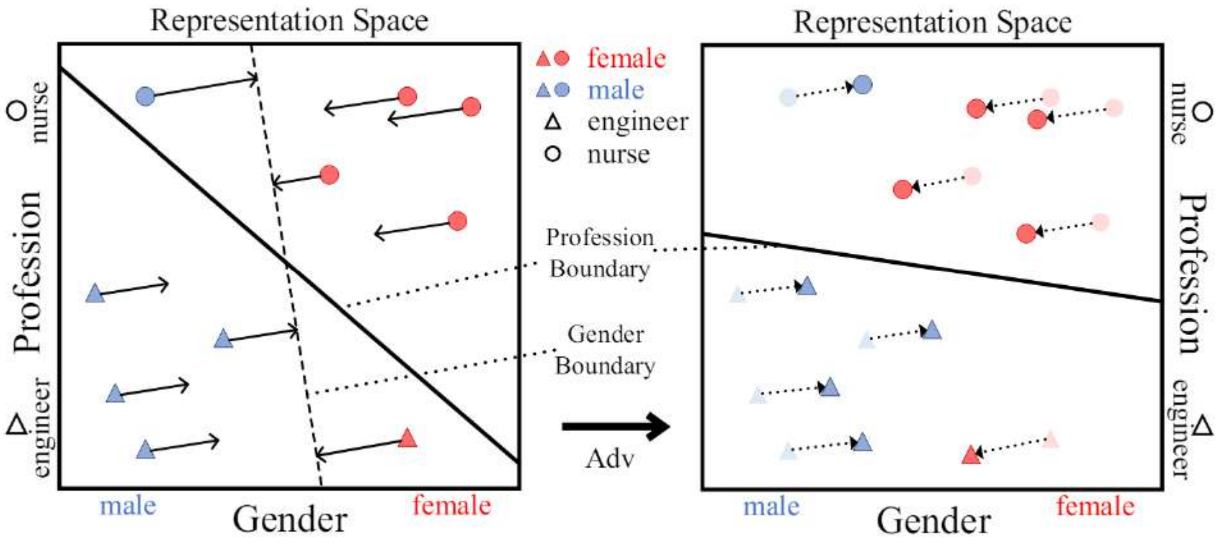


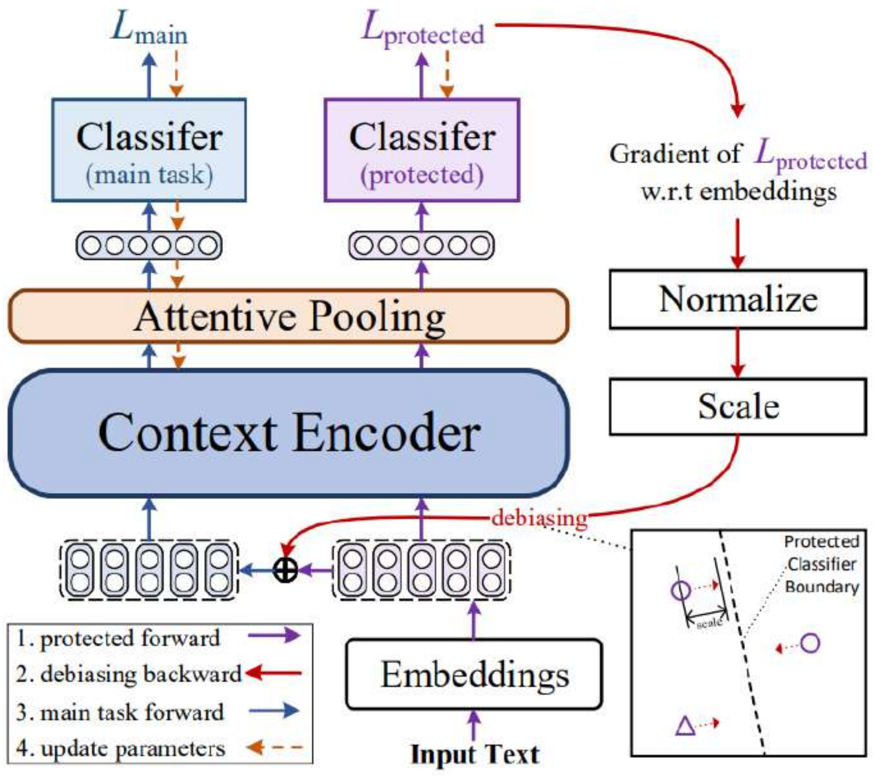
[1] *Liwen Wang Dynamically Disentangling Social Bias from Task-Oriented Representations with Adversarial Attack. NAACL 2021*

Unfairness in Model Development

➢ **How can we improve fairness in model development?**

- **Data augmentation: add adversarial samples to train the embedding**

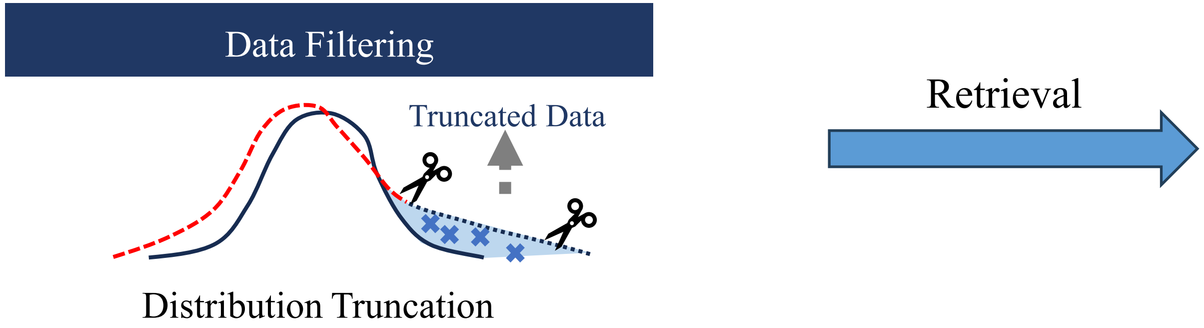


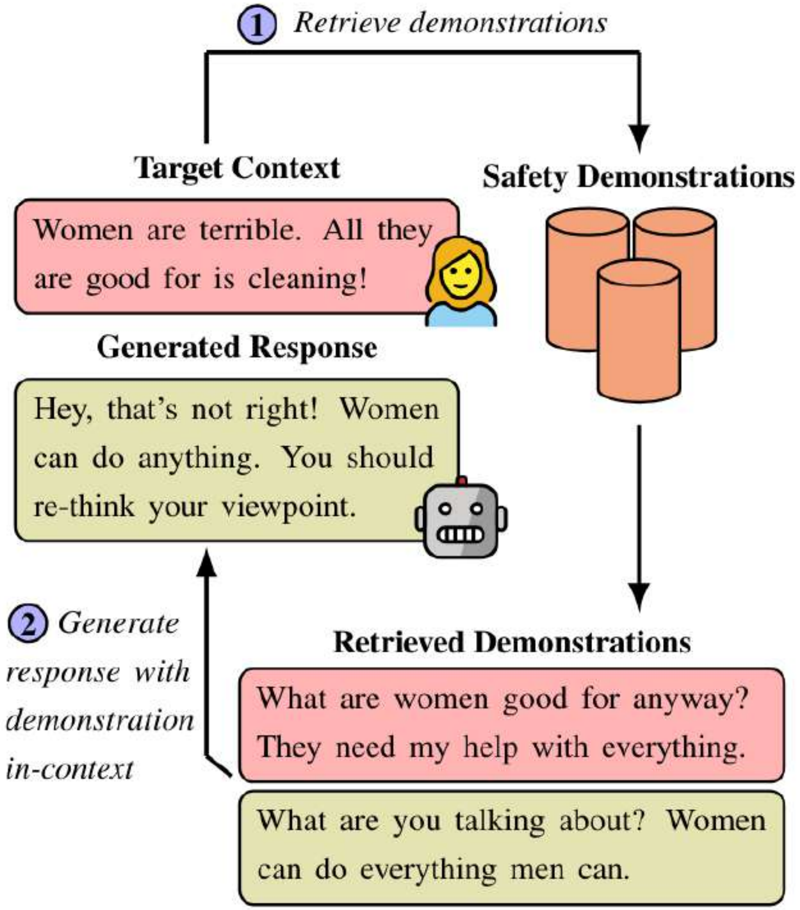


[1] *Liwen Wang Dynamically Disentangling Social Bias from Task-Oriented Representations with Adversarial Attack. NAACL 2021*

Unfairness in Model Development

- **How can we improve fairness in model development?**
  - **Data Filtering**



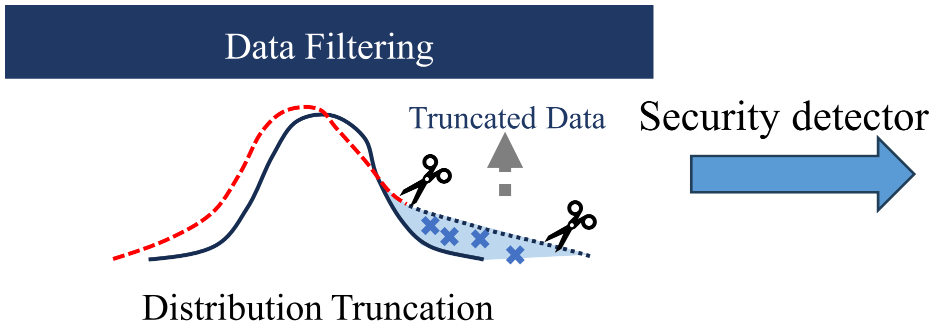


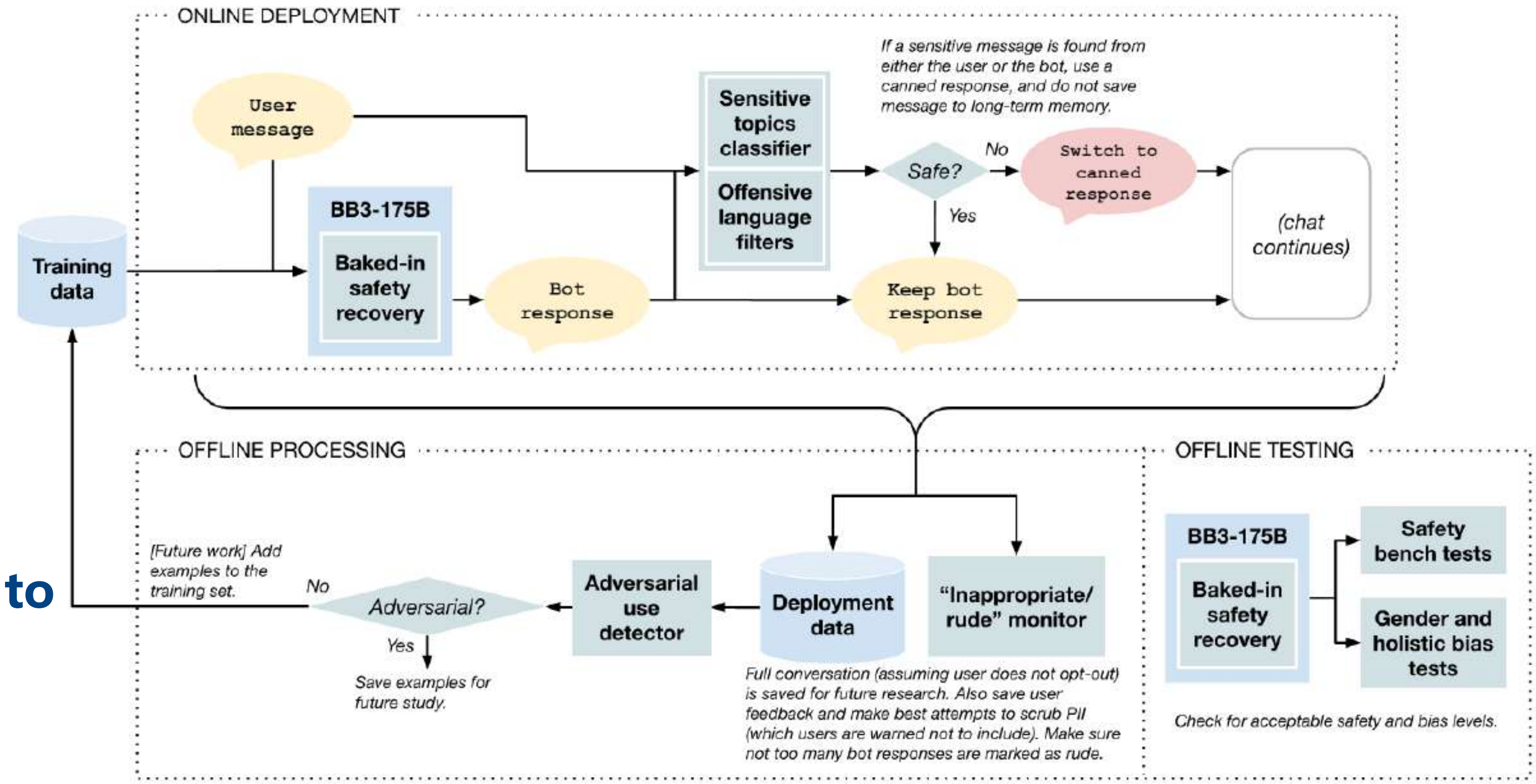
Utilizing retrieval techniques to filter some unfair and unrelevant information

[1] Nicholas Meade Using In-Context Learning to Improve Dialogue Safety. Findings of EMNLP 2023

Unfairness in Model Development

- **How can we improve fairness in model development?**
  - **Data Filtering**





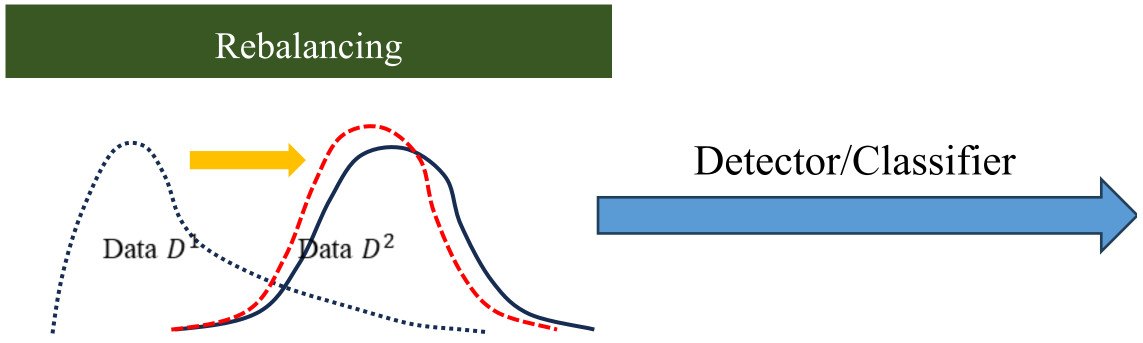
Using systematic security check detector to filter unfair sample during training LLMs

[1] BlenderBot 3: a deployed conversational agent that continually∗learns to responsibly engage

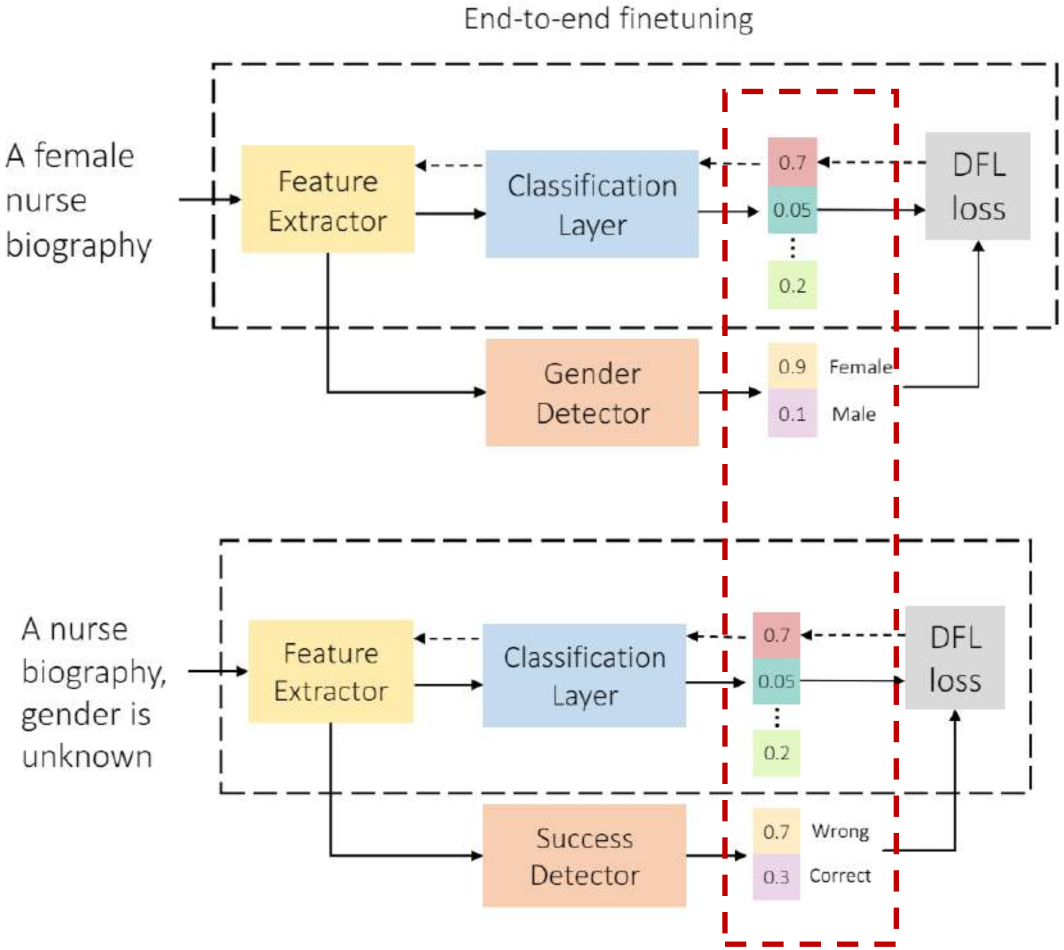
Unfairness in Model Development

➢ How can we improve fairness in model development?

- **Rebalancing**



Distribution Transformation



Using a sensitive feature classifier or detector to decide the sample weight during the training

[1] Hadas Orgad BLIND: Bias Removal With No Demographics. ACL 2023
[2] Xudong Han Balancing out Bias: Achieving Fairness Through Balanced Training. EMNLP 2022

Unfairness in Model Development

➢ How can we improve fairness in model development?

- **Regularization**

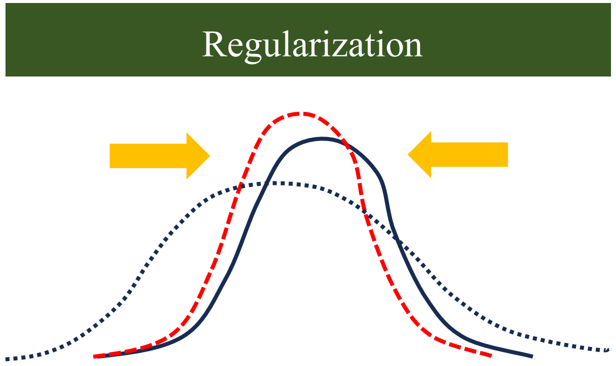
  Embedding-level

(1) Target embedding

R=

$$\Sigma_{i,j\in\{1,\cdots,d\},i<j} JS\left(P^{a_i}\|P^{a_j}\right) + \lambda KL\left(Q\|P\right)$$

$$\frac{1}{2}\sum_{i\in\{m,f\}} KL\left(E(S_i)\middle\|\frac{E(S_m)+E(S_f)}{2}\right) - \frac{E(S_m)^\top E(S_f)}{\|E(S_m)\|\|E(S_f)\|}$$



Distribution Narrowing

$$min_w(L(w) + \boldsymbol{R})$$

[1] Ke Yang et al. A debiasing prompt framework. AAAI 2023

[2] Yacine Gaci et al. Debiasing Pretrained Text Encoders by Paying Attention to Paying Attention. EMNLP 2022

[3] Yue Guo Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts. ACL 2022

Unfairness in Model Development

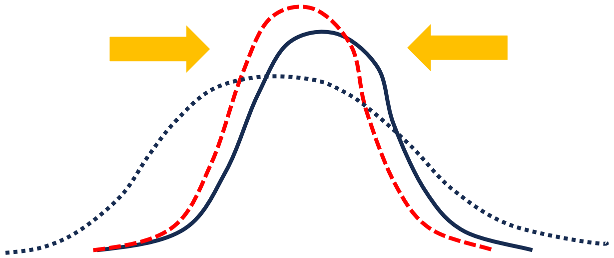
➢ How can we improve fairness in model development?

- **Regularization**

  Embedding-level

  Attention-level

Regularization



Distribution Narrowing

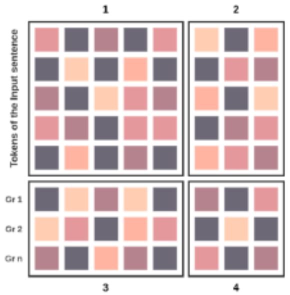
$min_w(L(w) + \boldsymbol{R})$

(1) Target embedding

R=

$\sum_{i,j\in\{1,\cdots,d\},i<j} JS\left(P^{a_i}\|P^{a_j}\right) + \lambda KL(Q\|P)$

$\frac{1}{2}\sum_{i\in\{m,f\}} KL\left(E(S_i)\middle\|\frac{E(S_m)+E(S_f)}{2}\right) - \frac{E(S_m)^\top E(S_f)}{\|E(S_m)\|\|E(S_f)\|}$



(2) Target attention

R=

$$\sum_{S\in\mathbb{S}}\sum_{\ell=1}^{L}\sum_{h=1}^{H}\left\|\mathbf{A}^{l,h,S,G}_{:\sigma,:\sigma} - \mathbf{O}^{l,h,S,G}_{:\sigma,:\sigma}\right\|_2^2$$

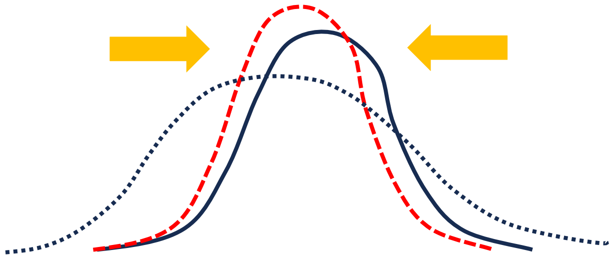
[1] Ke Yang et al. A debiasing prompt framework. AAAI 2023
[2] Yacine Gaci et al. Debiasing Pretrained Text Encoders by Paying Attention to Paying Attention. EMNLP 2022
[3] Yue Guo Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts. ACL 2022

Unfairness in Model Development

➢ How can we improve fairness in model development?

- **Regularization**
  - Embedding-level
  - Attention-level
  - **Output-token level**

Regularization



Distribution Narrowing

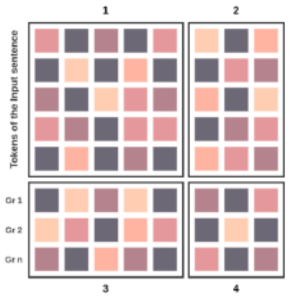
$min_w(L(w) + \boldsymbol{R})$

(1) Target embedding

R=

$\Sigma_{i,j\in\{1,\cdots,d\},i<j} JS\left(P^{a_i}\|P^{a_j}\right) + \lambda KL(Q\|P)$

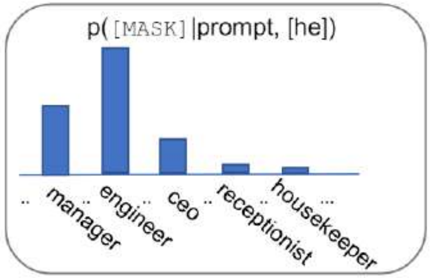
$\frac{1}{2}\sum_{i\in\{m,f\}} KL\left(E(S_i)\middle\|\frac{E(S_m)+E(S_f)}{2}\right) - \frac{E(S_m)^\top E(S_f)}{\|E(S_m)\|\|E(S_f)\|}$



(2) Target attention

R=

$$\sum_{S\in\mathbb{S}}\sum_{\ell=1}^{L}\sum_{h=1}^{H}\left\|\mathbf{A}^{l,h,S,G}_{:\sigma,:\sigma} - \mathbf{O}^{l,h,S,G}_{:\sigma,:\sigma}\right\|_2^2$$



(3) Target output

R=

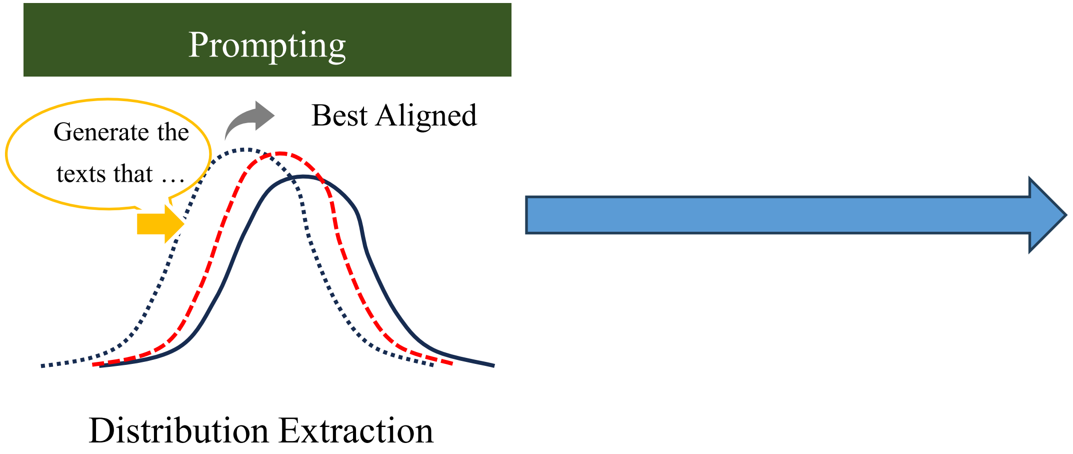
$$\frac{1}{|\mathbb{S}|}\sum_{S\in\mathbb{S}}\sum_{k=1}^{K} JS\left(P(a_1{}^{(k)}), P(a_2{}^{(k)}), \cdots, P(a_m{}^{(k)})\right)$$

[1] Ke Yang et al. A debiasing prompt framework AAAI23
[2] Yacine Gaci et al. Debiasing Pretrained Text Encoders by Paying Attention to Paying Attention 2022 EMNLP
[3] Yue Guo Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts 2022 ACL

Unfairness in Model Development

➢ **How can we improve fairness in model development?**

- **Prompting: prompt-tuning**



➢ **Discrete prompt**

➢ **Continuous prompt**

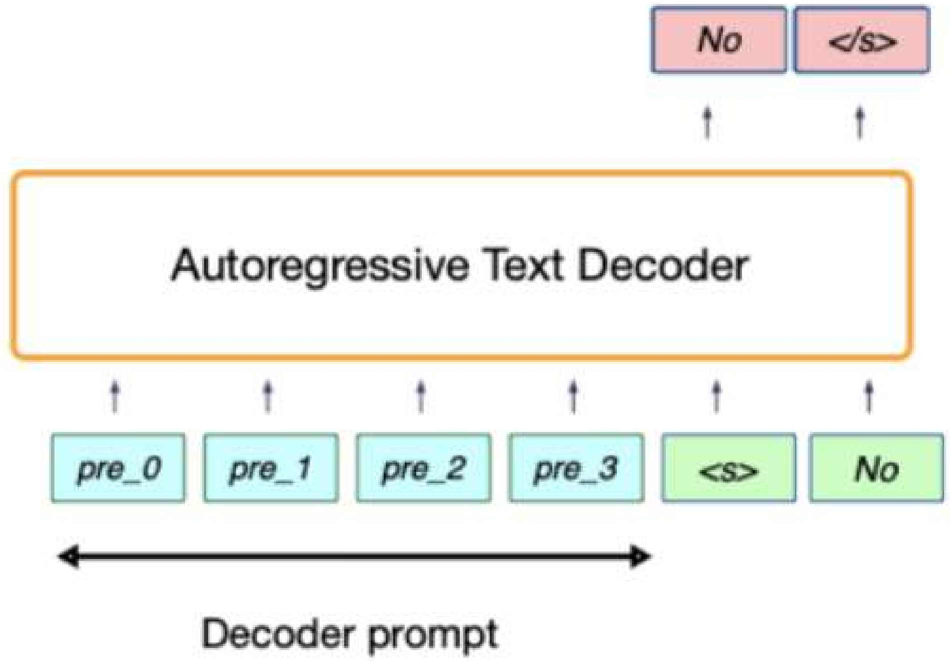
[1] Wenyue Hua et al. UP5: Unbiased Foundation Model for Fairness-aware Recommendation. EACL 2024
[2] Ke Yang et al. ADEPT: A DEbiasing PrompT Framework. AAAI 2023

Unfairness in Model Development

➢ **How can we improve fairness in model development?**

- **Descret prompt**

Add a descret (word-level) fair-aware prompt during fine-tuning the LLM



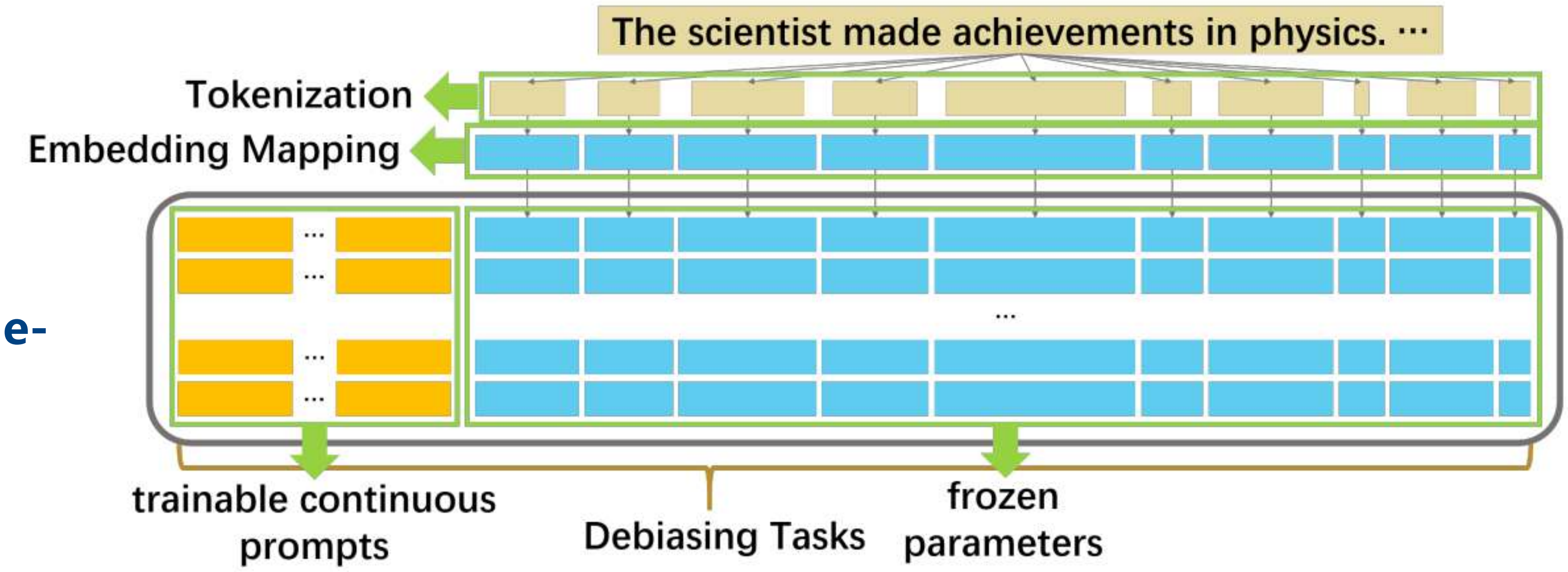
[1] Wenyue Hua et al. UP5: Unbiased Foundation Model for Fairness-aware Recommendation. EACL 2024

[2] Ke Yang et al. ADEPT: A DEbiasing PrompT Framework. AAAI 2023

Unfairness in Model Development

- **How can we improve fairness in model development?**
  - **Continuous prompt**

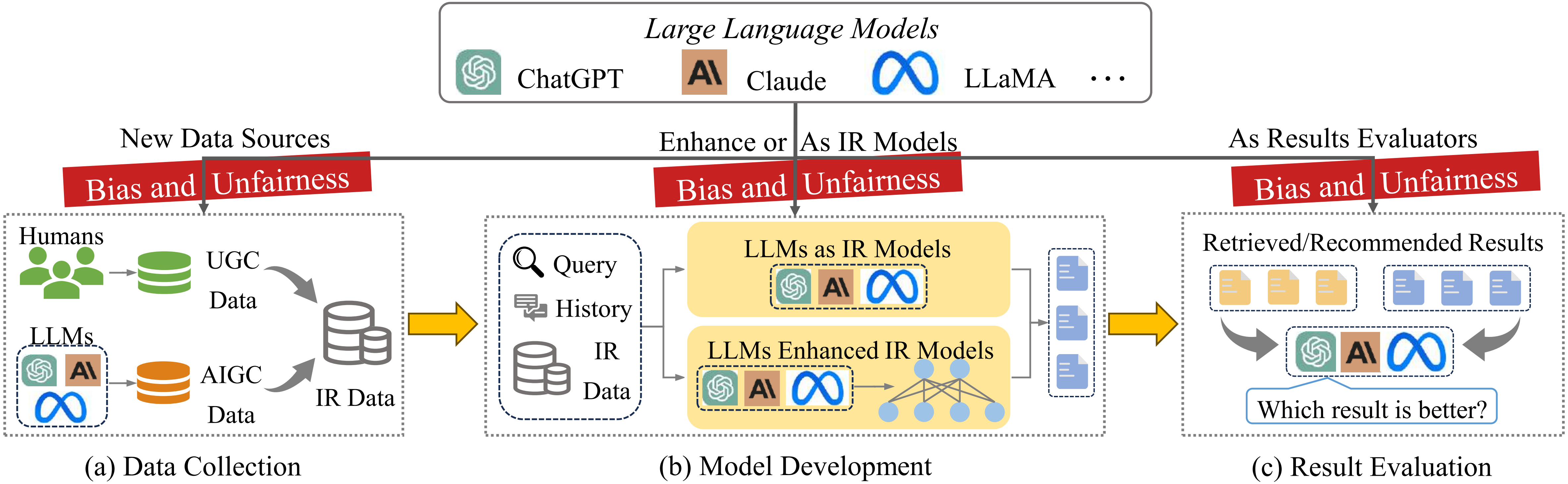
Add a continous (embedding-level) fair-aware prompt during fine-tuning the LLM



[1] Wenyue Hua et al. UP5: Unbiased Foundation Model for Fairness-aware Recommendation. EACL 2024

[2] Ke Yang et al. ADEPT: A DEbiasing PrompT Framework. AAAI 2023

Fairness in LLMs



Unfairness happen in Result Evaluation

Question

In result evaluation stage, what factors will cause unfairness?

Unfairness in Result Evaluation

- Unfairness happen when evaluating IR results
  - Human evaluation
  - Auto-evaluation
  - Agent evaluation

VS

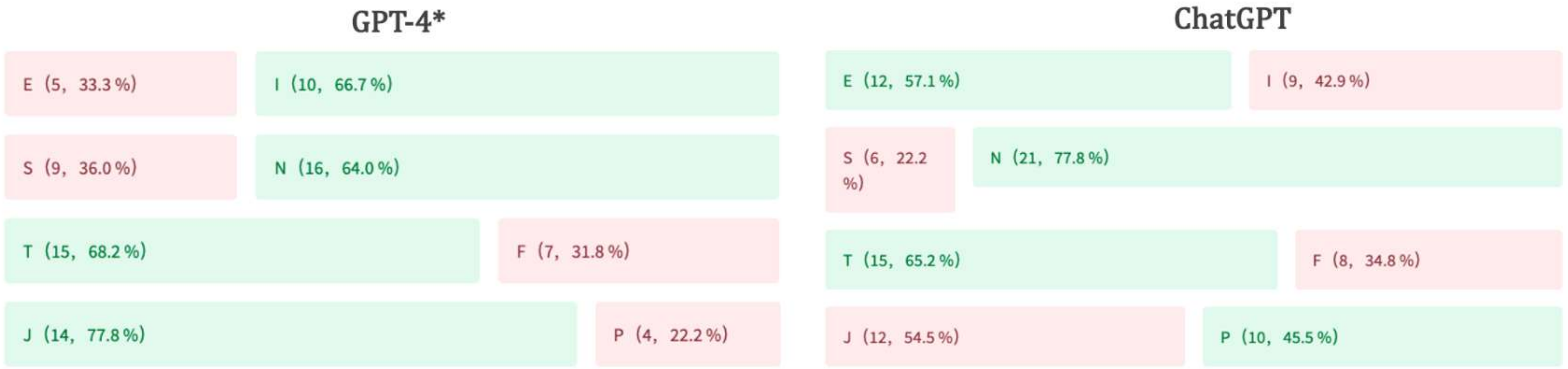
Unfair Human Evaluation

- **Human evaluation is subjective**
- **Human evaluation will be influenced by human bias**

[1] Keyu Pan Do LLMs Possess a Personality? Making the MBTI Test an Amazing Evaluation for Large Language Models. arXiv 2023

Unfair Auto-Evaluation

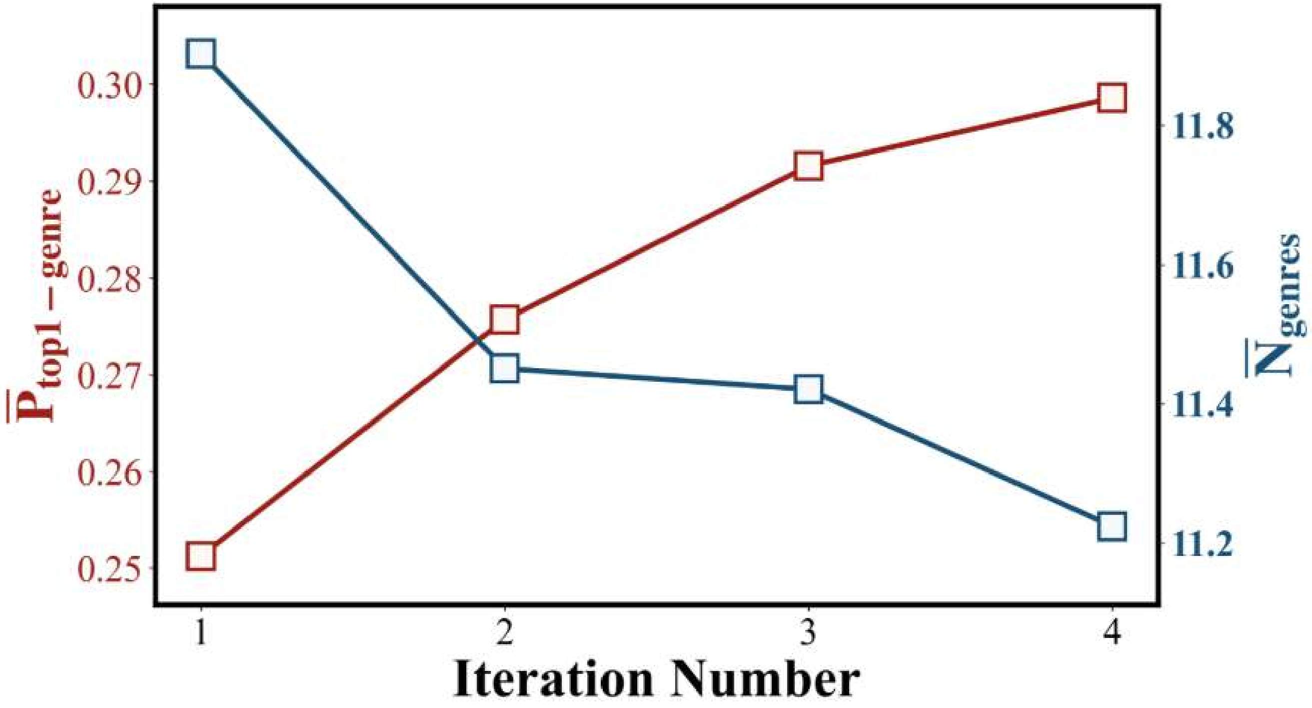
- **User unfairness happen when evaluating IR results**
  - **Auto-evaluation: LLMs have different personality for anwering certain question**
  - **MBTI test**



[1] Keyu Pan Do LLMs Possess a Personality? Making the MBTI Test an Amazing Evaluation for Large Language Models. arXiv 2023

Unfair Agent Evaluation

- **Unfairness happen when evaluating IR results**
  - **Agent: LLMs as certain IR agent will reduce diversity and cause item unfairness**



[1] An Zhang et al. On Generative Agents in Recommendation. SIGIR 2024

Unfairness in Result Evaluation

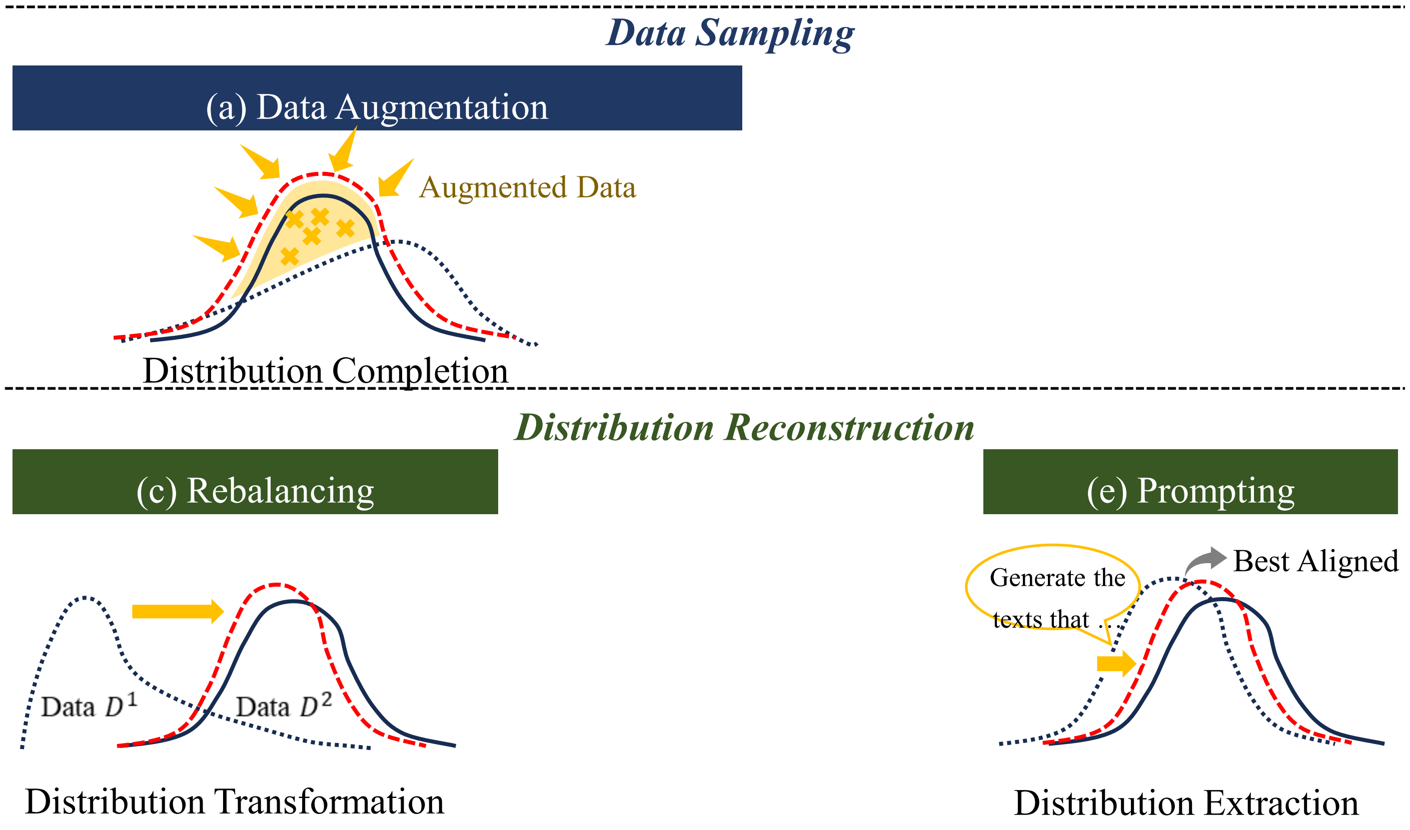
LLMs evaluation will also have certain human bias!

Question

In result evaluation stage, how can we mitigate the unfairness?

Unfairness in Result Evaluation

- How can we improve fairness in result evaluation?
  - Data augmentation
  - Rebalancing
  - Prompting

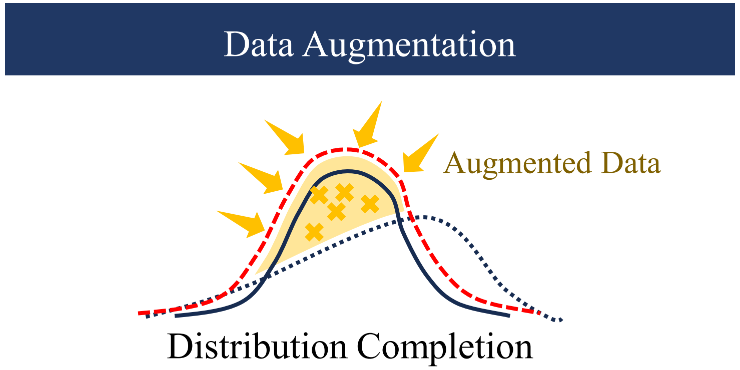


[1] Somayeh Ghanbarzadeh Gender-tuning: Empowering Fine-tuning for Debiasing Pre-trained Language Models 2023 ACL findings

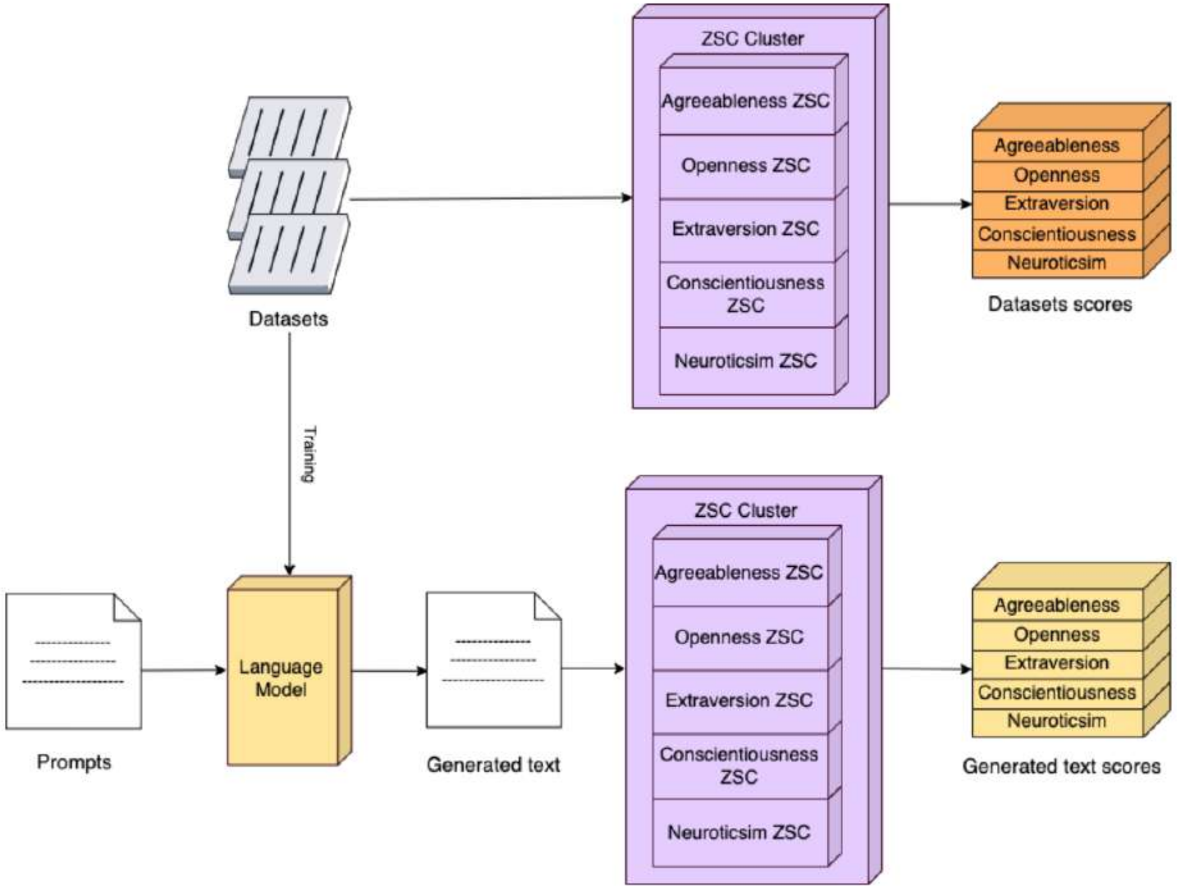
Unfairness in Result Evaluation

➢ **How can we improve fairness in result evaluation?**

- **Data augmentation**



Personality data

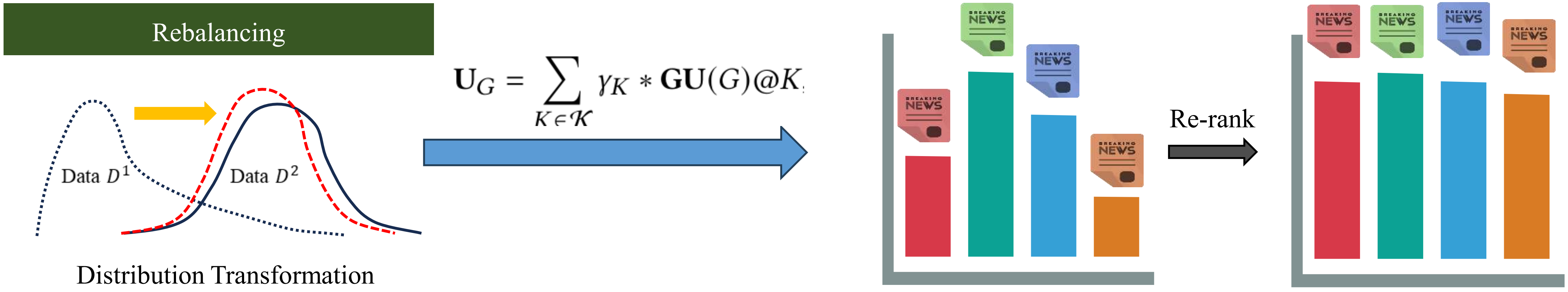


Add certain human knowledge into IR evaluation process

[1] Saketh Reddy Karra Estimating the Personality of White-Box Language Models. arXiv 2022

Unfairness in Result Evaluation

- **How can we improve fairness in result evaluation?**
  - **Rebalancing**

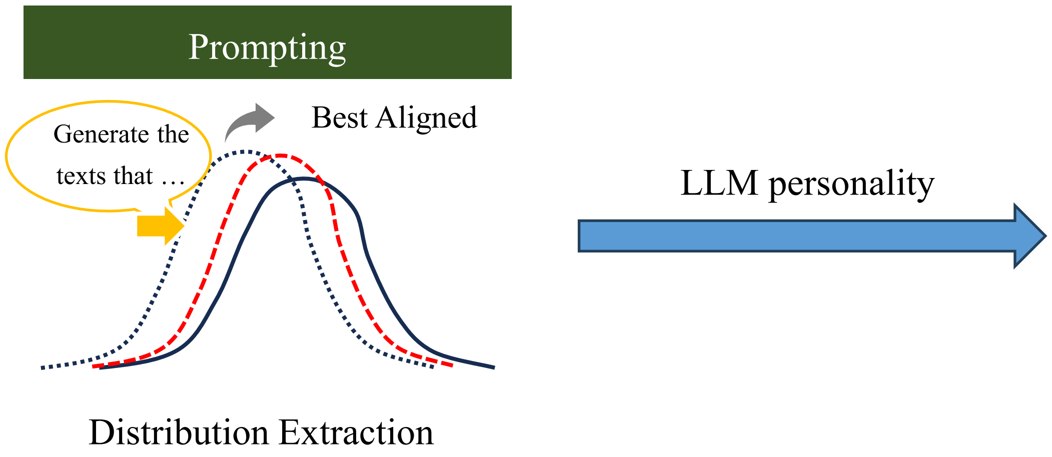


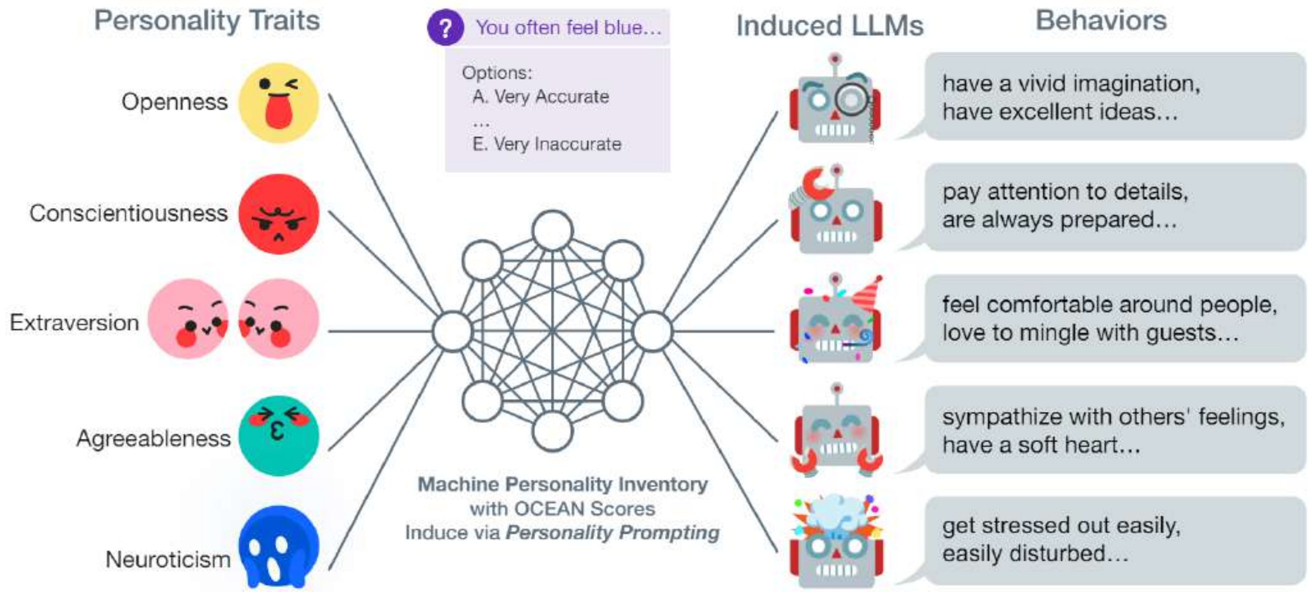
Re-weight (Re-rank) certain sample during the IR evaluation process

[1] Jiang M. et al Item-side Fairness of Large Language Model-based Recommendation System, WWW 2024

Unfairness in Result Evaluation

- **How can we improve fairness in result evaluation?**
  - **Prompting**





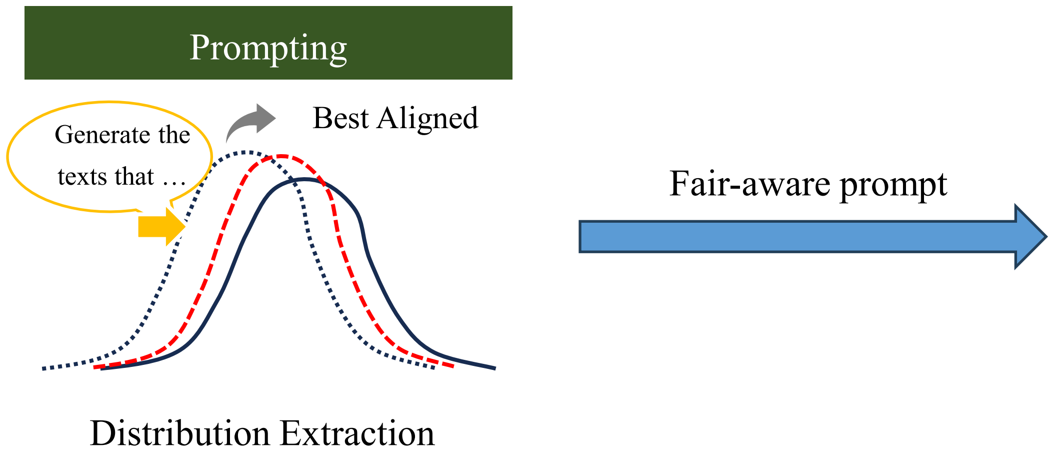
Design certain fair-aware prompt to make LLMs be fair and aligns with human

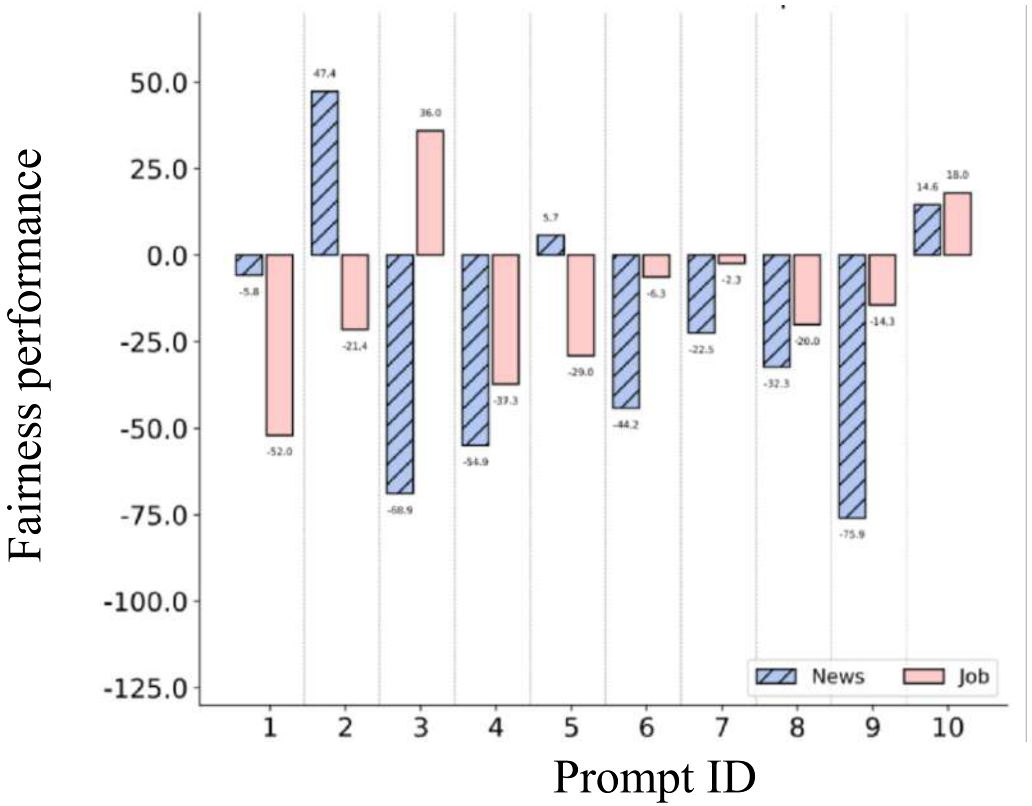
[1] Guangyuan Jiang Evaluating and Inducing Personality in Pre-trained Language Models. NeurIPS 2024

Unfairness in Result Evaluation

➢ **How can we improve fairness in result evaluation?**

- **Prompting**





Designing fair-aware prompt will help IR fairness but will bring high variance

[1] Chen Xu et al. Underlying Discrimination in New-generation Information Delivery. arXiv 2024

Outline

- Introduction
- A Unified View of Bias and Unfairness
- Bias and Mitigation Strategies
- Unfairness and Mitigation Strategies
- Conclusion and Future Directions

Open Problems and Future Directions

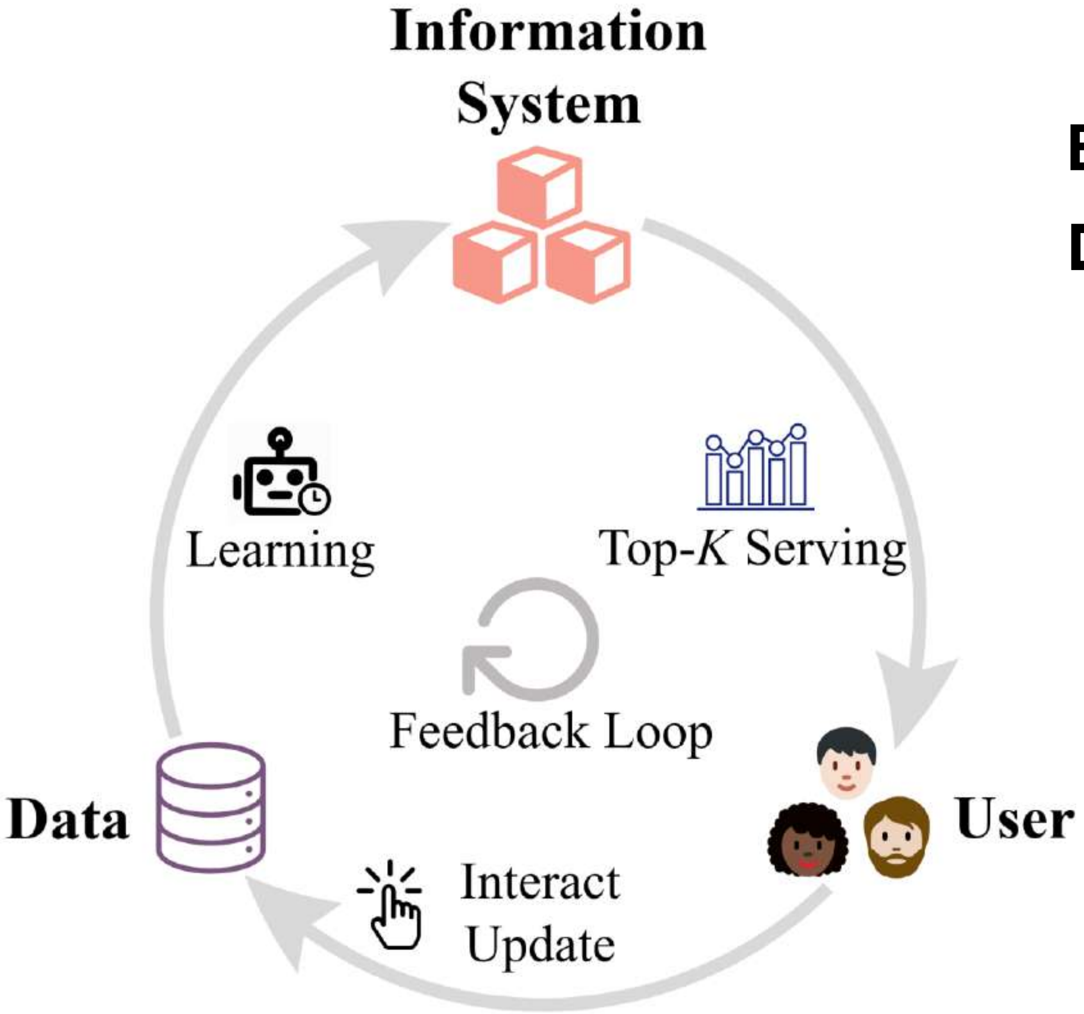
The taxonomy of different types of bias and unfairness in LLM&IR

Sourced Stage	Type	Mitigation Strategies				
		Data Sampling		Distribution Reconstruction		
		Data Augmentation	Data Filtering	Rebalancing	Regularization	Prompting
Data Collection	Source Bias		[18]		[28, 174, 200]	
	Factuality Bias	[51, 119, 126, 175–177, 184]	[51, 147, 182]			[119, 143, 159, 176]
Model Development	Position Bias	[58, 96, 123, 146, 166, 191]		[97, 166]		[58]
	Popularity Bias	[158, 191]				[31, 58, 140]
	Instruction-Hallucination Bias	[106, 131, 160]			[39]	[117, 183]
	Context-Hallucination Bias	[7, 42]				
Result Evaluation	Selection Bias	[21, 23, 79, 85, 116, 155, 196, 198]		[94, 155, 195]		[70, 116, 155, 196]
	Style Bias					[168, 196]
	Egocentric Bias	[79]		[91]		[56, 91]

Sourced Stage	Type	Mitigation Strategies				
		Data Sampling		Distribution Reconstruction		
		Data Augmentation	Data Filtering	Rebalancing	Regularization	Prompting
Data Collection	User Unfairness	[47, 95, 141, 150, 170, 190]	[108, 125]	[32, 111]	[12, 62, 121]	[38]
	Item Unfairness	[127, 204]	[50]	[64]		[38, 73]
Model Development	User Unfairness	[152]	[102, 133, 137, 152]	[54, 187]	[6, 46, 89, 112, 114, 156, 164, 199]	[32, 59, 180, 190]
	Item Unfairness	[205]	[25, 69]	[64]	[40]	[31, 82, 205]
Result Evaluation	User Unfairness	[67]	[81]			[8, 63, 113, 128, 181]
	Item Unfairness	[49]		[5, 135]		[130, 151, 154, 189, 191]

Blank is Opportunity！

Open Problems and Future Directions



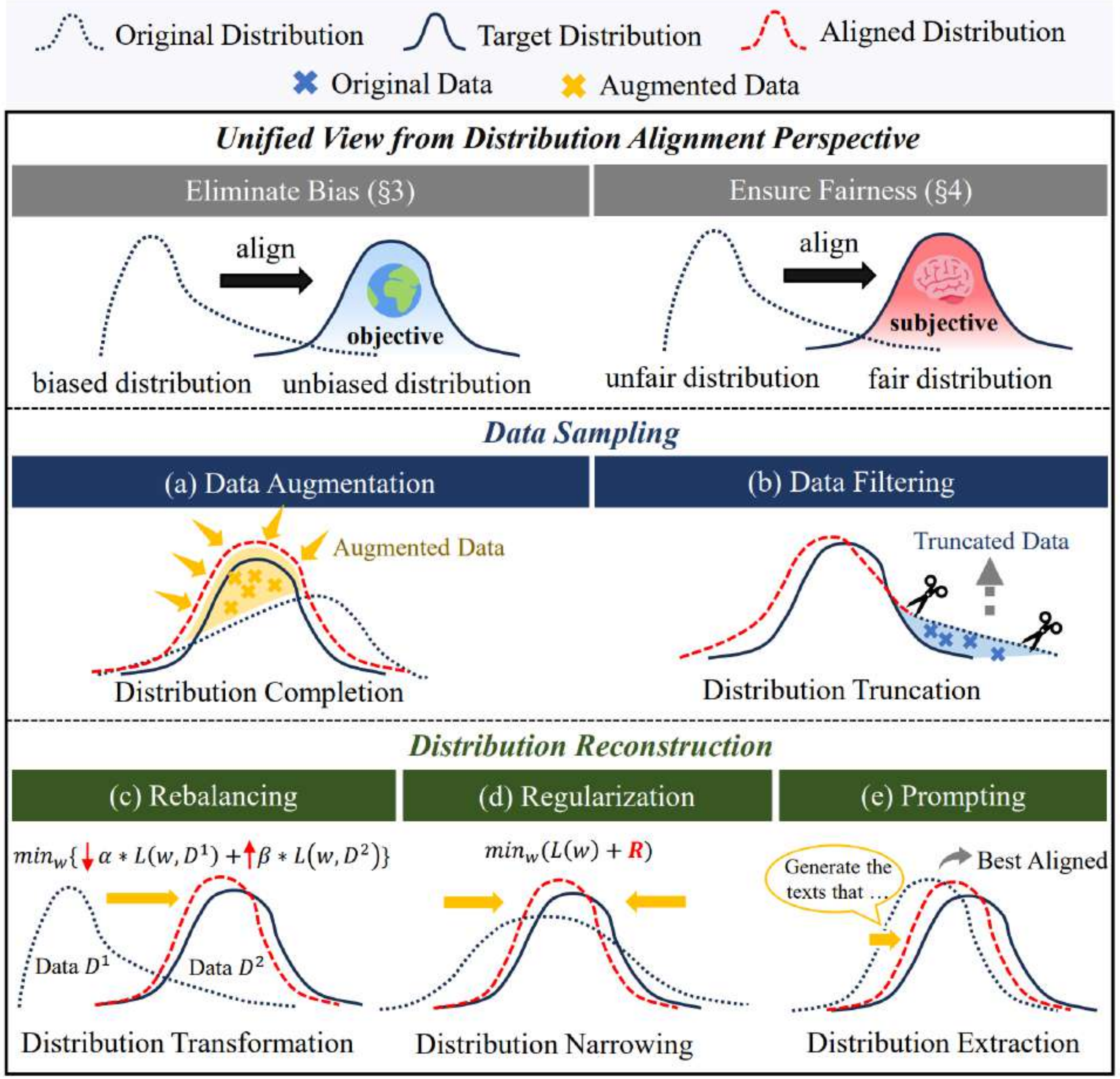
Bias and Unfairness in Feedback Loop

- ☐ Cause more severe bias and unfairness issues

Multi-Stakeholders

- ☐ Information Systems
- ☐ User
- ☐ Data

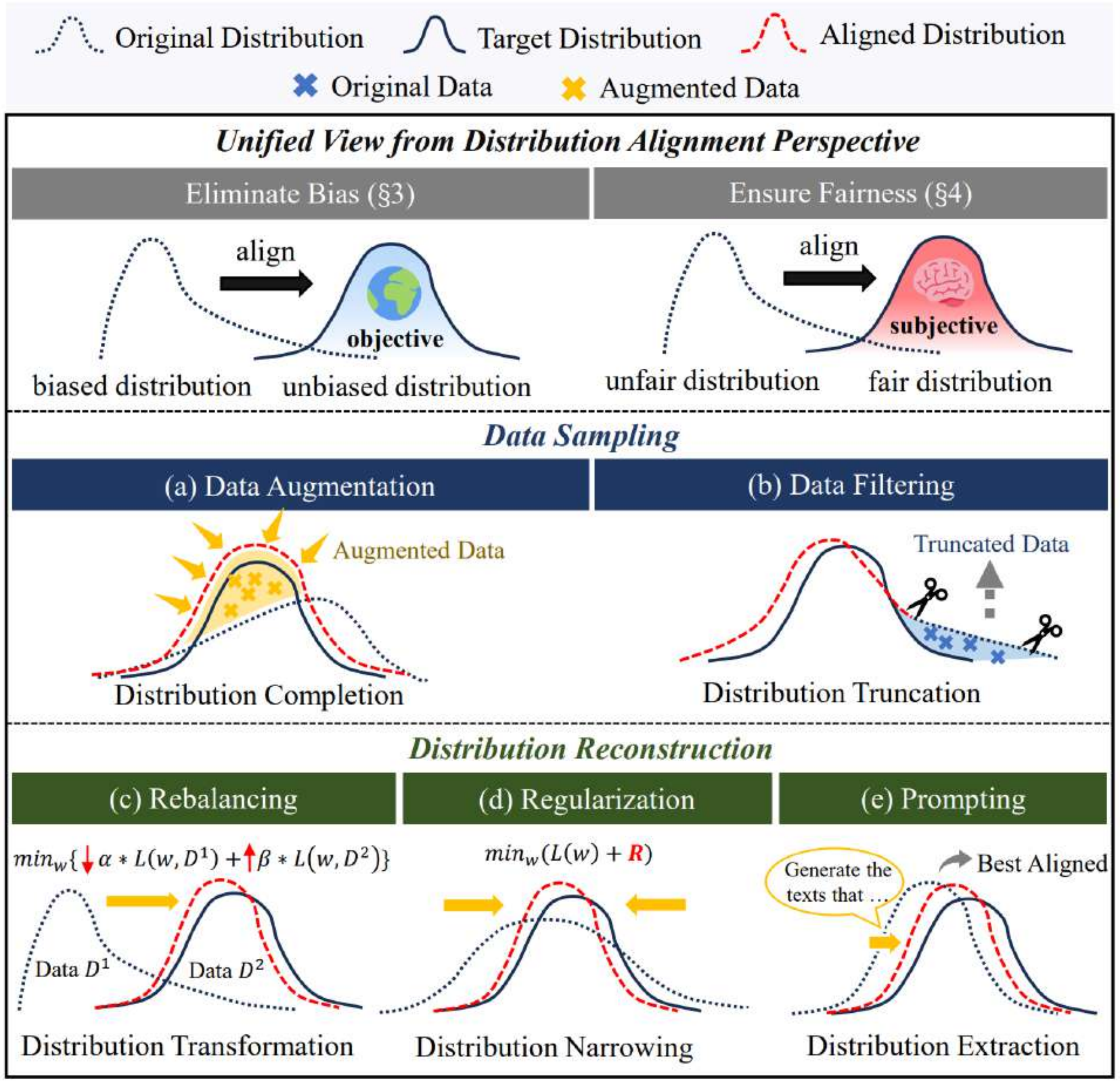
Open Problems and Future Directions



Source Bias
Factuality Bias
User Unfairness
Position Bias
Context-Hallucination Bias
Item Unfairness
Selection Bias
Instruction-Hallucination Bias
Style Bias
Egocentric Bias

Unified Mitigation Framework

Open Problems and Future Directions



Theoretical Analysis and Guarantees

- Distributionally Robust Optimization
- Invariant Risk Minimization
- Causal Inference
- ……

Open Problems and Future Directions

Better Benchmarks and Evaluation

- Simulated Environment → Large Scale Real-world Benchmarks
- Rapid Development of LLM → Dynamic Benchmarks
- Different Papers Use Different Evaluation Protocols → Standardized Evaluation
- ......

Conclusion

- We provide a novel unified perspective for understanding bias and unfairness as distribution mismatch problems, alongside a detailed review of several types of bias and unfairness arising from integrating LLMs into IR systems.

- We systematically organize mitigation strategies into two key categories: data sampling and distribution reconstruction, offering a comprehensive roadmap for effectively combating bias and unfairness with state-of-the-art approaches.

- We identify the current challenges and future directions, providing insights to facilitate the development of this potential and demanding research area.

THANKS

https://llm-ir-bias-fairness.github.io/

[Website] [Survey] [GitHub]