ChronoBias: A Benchmark for Evaluating Time-conditional Group Bias in the Time-sensitive Knolwedge of Large Language Models

Kyungmin Kim[1] Youngbin Choi[1] Hyounghun Kim[12] Dongwoo Kim[12] Sangdon Park[12]
[1]POSTECH GSAI [2]POSTECH CSE

EMNLP 2025 Suzhou, China 中国苏州

TL;DR

We propose a novel benchmark, `ChronoBias`, to highlight the importance of ***time-conditional (temporal) group bias*** analysis in fully uncovering the mechanisms of group bias in the ***time-sensitive*** knowledge of large language models (LLMs).

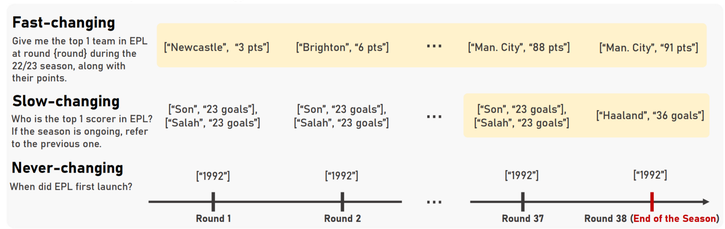
Problem 1: Group Bias of LLMs

- LLMs often show disparate performance across factors such as demographic groups, regions, or levels of popularity.
- This undermines LLMs' fair deployment for diverse groups of users.

Metric	Brazil	Spain	England	Korea	Saudi	Algeria
Model Performance	0.378 (1)	0.275 (2)	0.237 (3)	0.14 (4)	0.136 (5)	0.073 (6)
Monthly Page-views	47,028 (3)	219,607 (2)	622,919 (1)	10,324 (5)	28,197 (4)	4,098 (6)

Problem 2: Time-sensitivity of Knowledge

- LLMs must adapt to *time-sensitive* knowledge that are updated after their knowledge cutoffs, to maintain performance uniformly across all time intervals.
- Knowledge cutoff of an LLM: The most recent date of the knowledge used for its pretraining



Possible Solution: Retrieval-augmented Generation (RAG)

- RAG can be a possible solution to aforementioned problems, since it...
  - supplements knowledge acquired before the knowledge cutoff that was not well retained in model parameters (**Problem 1**);
  - processes new information without further fine-tuning (**Problem 2**).
- **Main Goal**: **Construct RAG-based LLMs, which are fair and accurate uniformly across all time intervals!**

Research Questions (RQ)

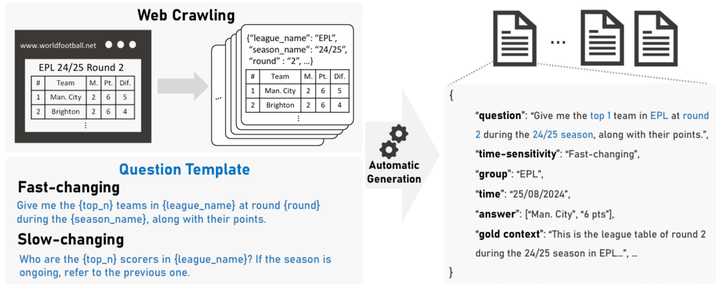
- **RQ 1. Group bias in vanilla LLMs**
- **RQ 2. Rationale for group bias in vanilla LLMs**
- **RQ 3. Guidelines for building fair and accurate RAG-based LLMs**

Main Challenge: Lack of Group Bias Analysis in the Time-sensitive Knowledge of LLMs

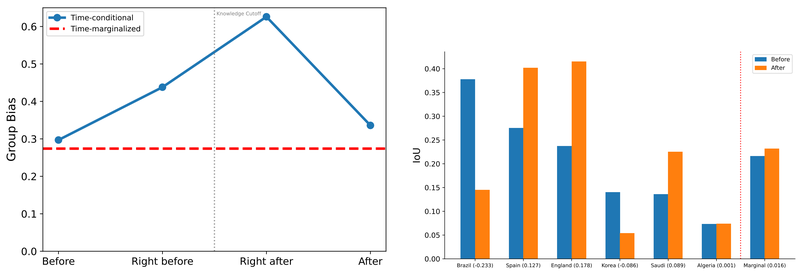
- For questions involving *time-sensitive* knowledge, LLMs provide correct answers only during specific time periods.
- **Limitation of existing group bias analyses**: They average over all time points, thereby overlooking such temporal variability.
- We refer to this type of evaluation as *time-marginalized group bias*.

`ChronoBias` Benchmark

- ChronoBias is tailored to *time-conditional group bias* analysis!
  - Group: Geographical regions
  - Date of knowledge update
- `ChronoBias` consists of QA pairs in the sports domain across 6 geographical regions from 2015 to 2025, each aligned with a gold passage.



RQ 1. Temporal Group Bias in Vanilla LLMs



(**Left**) *Time-marginalized group bias* and *time-conditional group bias*. (**Right**) Group-wise model performance before and after the knowledge cutoff.

RQ 2. Rationale for Temporal Group Bias in Vanilla LLMs

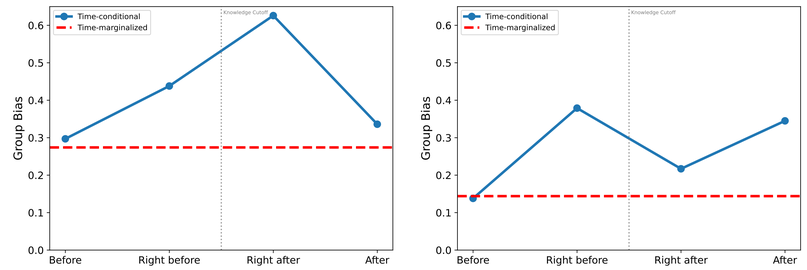
Parametric knowledge bias: **Before** and **After** the knowledge cutoff

Metric	Brazil	Spain	England	Korea	Saudi	Algeria
Model Performance	0.378 (1)	0.275 (2)	0.237 (3)	0.14 (4)	0.136 (5)	0.073 (6)
Monthly Page-views	47,028 (3)	219,607 (2)	622,919 (1)	10,324 (5)	28,197 (4)	4,098 (6)

Time-sensitivity bias: **After** the knowledge cutoff

Parametric Knowledge		High			Low		
Time-varying		Slow		Fast	Slow		Fast
Region		England	Spain	Brazil	Saudi	Korea	Algeria
Top	1	5.42	4.16	**6.42**	4.23	4.59	**6.03**
	3	9.81	9.21	**12.76**	15.5	12.5	**12.4**
	5	13.13	15.08	**16.84**	11.92	17.86	**15.5**
Bottom	1	9.32	8.63	**9.13**	8.42	7.55	**9.77**
	3	21.53	20.37	**21.55**	18.92	15.18	**22.33**
	5	27.58	25.05	**28.66**	23.77	17.95	**26.97**

RAG Partially Mitigates Temporal Group Bias

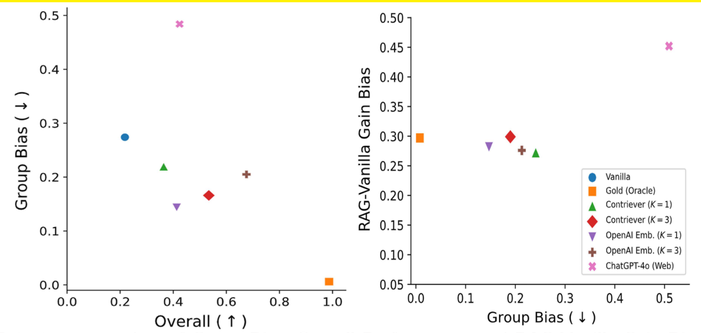


(a) Vanilla LLM (b) RAG-based LLM

Time-marginalized group bias and *time-conditional group bias*.

RQ 3. Guidelines for Building Fair and Accurate RAG-based LLMs

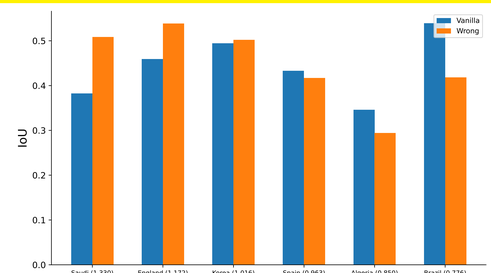
Guideline 1: "UNFAIR" retrieval for "FAIR" model performance



Pairwise scatterplots of **Group Bias**, **Overall Performance**, and **RAG-Vanilla Gain Bias**.

- **Overall Performance**: Group-wise model performance
- **RAG-Vanilla Gain Bias**: Disparity between the maximum and minimum performance improvements from RAG.

Guideline 2: Even when the same type of retrieval error occurs, *time-sensitivity bias* causes variation in how much useful partial information is present in the retrieved document.



Group-wise performance of Vanilla and RAG-based LLMs only on queries with incorrect retrieval.