

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

Kyungmin Kim¹ Youngbin Choi¹ Hyounghun Kim¹² Dongwoo Kim¹² Sangdon Park¹²

¹POSTEC GSAI ²POSTECH CSE

{kkm959595, choi.youngbin, h.kim, dongwoo.kim, sangdon}@postech.ac.kr

Abstract

In this paper, we propose ChronoBias, a novel benchmark for evaluating *time-conditional group bias* in the *time-sensitive* knowledge of large language models (LLMs). Our benchmark is constructed via a template-based semi-automated generation method, balancing the quality-quantity trade-off in existing benchmark curation approaches. For knowledge that changes over time, *time-conditional group bias* exhibits varying patterns across time intervals, evident in both the best- and worst-performing groups and in the bias metric itself. In addition to *parametric knowledge bias*—which influences group bias across all time intervals—we identify *time-sensitivity bias* as an additional factor after a model’s knowledge cutoff, accounting for much of the variation in *time-conditional group bias* over time. Since both biases are irreducible, retrieval-augmented generation (RAG) can be a promising approach, as it can address post-cutoff knowledge and better leverage pretraining knowledge that is under-represented in the model parameters. While RAG improves both overall performance and group bias, we observe that the disparate patterns of *time-conditional group bias* still persist. Therefore, through extensive experiments with various model configurations, we illustrate how accurate and fair RAG-based LLMs should behave and provide actionable guidelines toward constructing such ideal models.

1 Introduction

Despite their success on a wide range of downstream tasks, large language models (LLMs) often show disparate performance across factors such as demographic groups, geographical regions, or levels of popularity, undermining their trustworthy and fair deployment for diverse users (Zhang and Choi, 2021; Mallen et al., 2023; Yang et al., 2024; Chiu et al., 2024; Singh et al., 2024; Naous et al., 2024). While the benchmarks proposed by the authors effectively demonstrate the presence of such biases,

they do not account for the *time-sensitive* nature of knowledge. We define the *time-sensitivity* of knowledge as the rate at which it evolves over time. For questions involving *time-sensitive* knowledge, LLMs may provide correct answers only during specific periods (Zhang and Choi, 2021). Existing group bias evaluations, however, effectively average over all time points, thereby overlooking such temporal variability. We refer to this type of evaluation as *time-marginalized group bias*.

Meanwhile, several benchmarks show that LLMs struggle to answer questions involving *time-sensitive* knowledge (Vu et al., 2023; Yang et al., 2024). Specifically, they categorize knowledge into three groups based on its degree of *time-sensitivity*—*never-changing*, *slow-changing*, and *fast-changing*—and show that there is a negative correlation between the degree of *time-sensitivity* and model performance on downstream tasks. In other words, the more *time-sensitive* the knowledge is, the more difficult it is for LLMs to generate correct answers. However, the analyses lack the notion of group fairness: when the knowledge content is the same, its degree of *time-sensitivity* remains similar across groups, yet LLM performance may still differ. Although benchmarks exist that assess LLM performance with respect to both group and *time-sensitivity*, these factors have been considered independently rather than in a joint analysis (Zhang and Choi, 2021; Yang et al., 2024). See Table 1 for existing benchmarks that analyze LLMs by the axes of group or *time-sensitivity*.

In addition to the challenge of addressing group fairness in the context of the *time-sensitivity* of knowledge, LLMs must adapt to *slow-changing* and *fast-changing* knowledge that are updated after their knowledge cutoffs. The knowledge cutoff of an LLM is defined as the most recent date of the knowledge used for its pretraining. A number of benchmarks have been proposed to evaluate whether it is possible to continually pretrain LLMs,

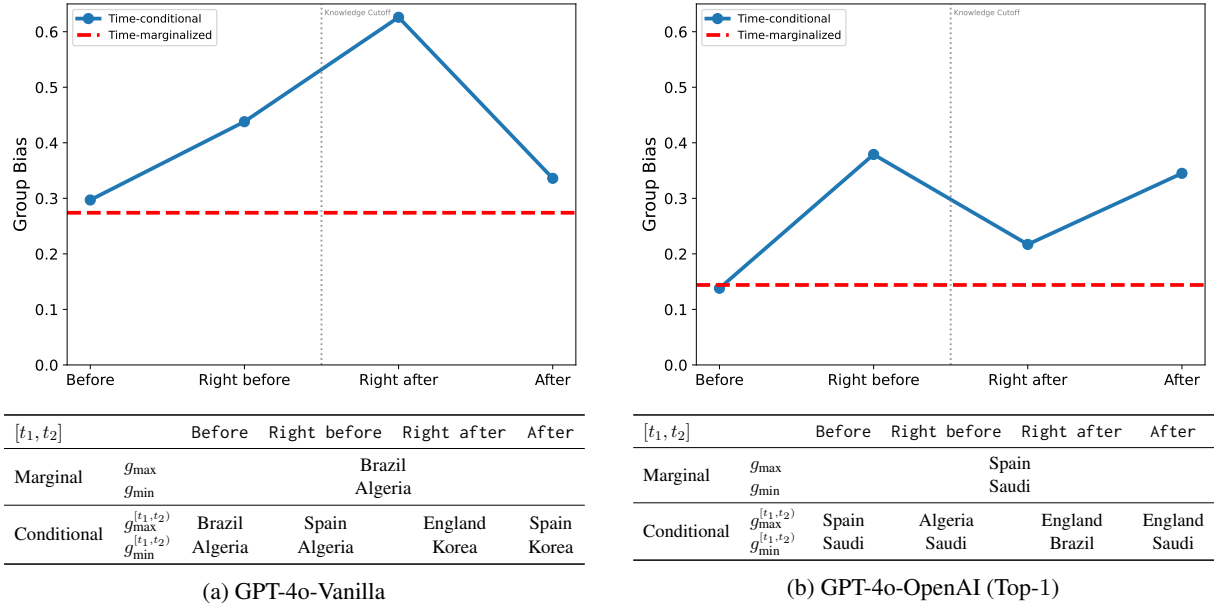


Figure 1: *Time-marginalized group bias* and *time-conditional group bias* of GPT-4o-Vanilla (left) and GPT-4o-OpenAI (right). **Top:** The red dashed line indicates the *time-marginalized group bias*. The blue solid line shows the *time-conditional group bias* for each time interval. **Bottom:** ‘Marginal’ rows show regions with the highest and lowest overall rewards, averaged over all time steps. ‘Conditional’ rows show regions with the highest and lowest rewards within each time interval.

a process often referred to as continual knowledge learning (Zhang and Choi, 2021; Kim et al., 2023; Ko et al., 2024). In these benchmarks, the categories of *new*, *unchanged*, and *updated* knowledge are defined based on two consecutive time steps, and the evaluation measures whether a continual pretraining method can adapt to the *new* or *updated* knowledge while preserving the *unchanged* information.

Despite their additional training costs, most continual pretraining methods face two critical challenges: they struggle to adapt to *new* or *updated* knowledge, and they often forget *unchanged* information already encoded in the model parameters. As a result, they exhibit suboptimal performance on downstream tasks. As an alternative, retrieval-augmented generation (RAG), which augments a query with relevant context retrieved from an up-to-date knowledge base, enables LLMs to process *new* information without requiring further pretraining or fine-tuning (Lewis et al., 2020). Moreover, RAG can also supplement knowledge acquired before the knowledge cutoff that was not well retained in model parameters.

In seeking to construct fair and accurate RAG-based LLMs on *time-sensitive* knowledge across all time intervals, the problem reduces to two central challenges: (1) identifying the properties of such ideal RAG-based LLMs, and (2) developing methods to construct them. To mitigate these challenges,

we propose a novel benchmark, ChronoBias, designed to support these analyses and prototyping efforts, through which we derive actionable guidelines for building such models. ChronoBias consists of question-answer (QA) pairs in the sports domain across geographical regions, each aligned with a gold passage. In addition, each QA pair is annotated with labels for the group (defined as geographical region), the degree of *time-sensitivity*, and the date of knowledge update.

By leveraging ChronoBias, we first observe that the vanilla performance of LLMs exhibits varying patterns of *time-conditional group bias* across time intervals, defined as the discrepancy between the maximum and minimum group performance on knowledge from a given time interval (Fig. 1a). In alignment with existing research on group bias, we demonstrate that *parametric knowledge bias*—i.e., bias in group-specific knowledge encoded in model parameters—affects *time-conditional group bias* across all time intervals. In contrast, we further show that *time-sensitivity bias*—i.e., bias in the degree of *time-sensitivity* across groups—serves as an amplification factor only after the knowledge cutoff, leading to discrepancies in *time-conditional group bias* across time intervals. For Fig. 1b, most RAG baselines improve both accuracy and group bias compared to vanilla performance, but *time-conditional group bias* still exhibits distinct patterns across time intervals. Thus, we further iden-

Benchmark	Group	Time	Joint Analysis	Time Taxonomy	Automation
EvolvingQA (2024)	×	○	×	<i>Unchanged, Changed, New</i>	LLM-aided
GrowOVER-QA (2024)	×	○	×	<i>Unchanged, Changed, New</i>	○
HOH (2025)	×	○	×	<i>Unchanged, Changed, New, Outdated</i>	○
FreshLLMs (2024)	×	○	×	<i>Fast-, Slow-, Never-changing</i>	×
Daily Oracle (2024)	×	○	×	Time	LLM-aided
CAMEL (2024)	○	×	×	-	△
CulturalBench (2025)	○	×	×	-	×
POPQA (2024)	○	×	×	-	△
CRAG (2024)	○	○	×	<i>Real-, Fast-, Slow-, Never-changing</i>	×
SituatedQA (2021)	○	○	×	Time	×
ChronoBias (Ours)	○	○	○	<i>Fast-, Slow-, Never-changing</i>	△

Table 1: Comparison of ChronoBias with existing benchmarks. **Group** displays whether the group bias is measured. **Time** indicates whether the time axis is considered in the benchmark. **Joint Analysis** indicates whether time and group variables are jointly considered to evaluate the group bias. **Time Taxonomy** describes the categorization criteria of questions with respect to time, if it exists. **Automation** column shows whether a benchmark generation process is automated.

tify how an ideal RAG should behave based on results using gold passages, and provide practical guidelines for building continuously fair and accurate RAG-based LLMs. Additional details on related work can be found in the Appendix A.

As for technical contributions in terms of benchmark construction, ChronoBias is built upon a novel, template-based benchmark generation method—a semi-automated approach that well balances the quality and quantity tradeoff of existing benchmarks (Fig. 2). Once the QA template and data extraction method are defined, the dataset can be continuously expanded over time without additional labeling of the degree of *time-sensitivity* for each question.

Our contributions are summarized as follows:

1. We first emphasize the importance of analyzing *time-conditional group bias* in LLMs with respect to *time-sensitive* knowledge as a critical component group fairness evaluation.
2. Built upon a novel template-based semi-automated generation method, ChronoBias serves as a benchmark for *time-conditional group bias* analysis, through which we identify the causes of variations in the *time-conditional group bias* of vanilla LLMs across different time intervals.
3. Taking RAG as a possible solution for achieving fair and accurate LLMs over time, we further characterize the ideal properties of RAG-based LLMs and conclude by providing practical guidelines for building them, supported by extensive experiments on diverse model configurations.

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)

Table 2: Group-wise comparison of model performance before the knowledge cutoff of GPT-4o and average monthly Wikipedia pageviews.

2 The ChronoBias Dataset

Data Description. We construct a benchmark dataset based on football league data collected from 2015 to 2025 across six different regions $g \in \mathcal{G}$: England, Spain, South Korea, Saudi Arabia, Algeria, and Brazil. For seasons up to and including 2025, we consider the results on April 15 of the 2025 season as the latest time step $T \in \mathcal{T}$, treating each date $t \in \mathcal{T}$ as a time unit. The regions were selected prior to constructing ChronoBias, guided by the aim of including football leagues with varying levels of popularity. Specifically, following Mallen et al. (2023), we selected the regions based on the monthly Wikipedia page views of football leagues associated with each region (Tab. 2), which we use as a proxy for popularity.

For each league, we consider three main types of data: league tables by round, top scorers at the end of season, and historical records. For league table questions, we collect data from <https://www.worldfootball.net>, a comprehensive football database that provides global competition statistics and match results. Questions regarding top scorers are sourced from <https://www.flashscore.com/>, a live sports results website offering real-time scores, statistics, and match information. The statistics of the dataset are provided in Table A1.

Taxonomy of Questions $q \in \mathcal{Q}$. In contrast to existing automated benchmarks (Kim et al., 2023; Ko

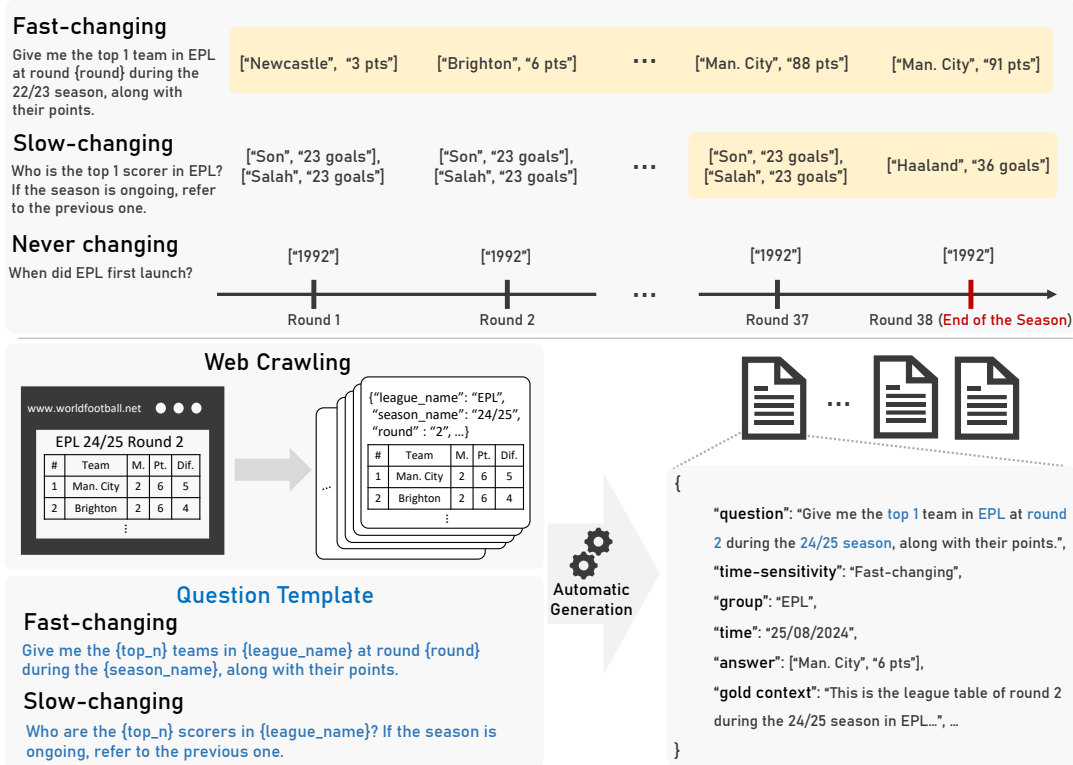


Figure 2: **Top:** Example questions from ChronoBias with different degrees of *time-sensitivity*. **Bottom:** Pipeline of the proposed template-based semi-automated benchmark generation method. While *never-changing* questions also have question templates for each question, they are human-curated. However, *never-changing* questions do not require additional maintenance.

et al., 2024; Ouyang et al., 2025), which generate QA pairs from a given document corpus, we adopt an inverted approach. We classify each question into one of three categories according to its degree of *time-sensitivity*: *never-changing*, *slow-changing*, or *fast-changing*.

For instance, “When did the English Premier League first begin?” is a *never-changing* question, since its answer remains fixed once the league is founded. In contrast, questions involving league standings are typically *fast-changing*, as their answers may vary on a weekly basis. Questions on top scorers of the season are categorized as *slow-changing*, as they change on a yearly basis.

Template-based Benchmark Generation. Once the degree of *time-sensitivity* is determined for a given question, we create a question template to enable semi-automatic benchmark generation (Fig. 2). Once the template is defined, e.g., “Name the top {top_n} teams in {league_name} at round {rounds} during the {season}”, the template is applied consistently across all leagues, seasons, and rounds.

After defining the template, we instantiate the argument, assign ground-truth answer $a \in \mathcal{A}$, and extract gold context $d_{\text{gold}} \in \text{KB}^{(T)}$ for each ques-

tion $q \in \mathcal{Q}$. Here, $\text{KB}^{(T)}$ refers to the knowledge base that has been accumulated up to T . Both answer annotation and gold passage extraction are performed via web crawling from the aforementioned websites.

After the crawling is done, $(q, a, t, g, d_{\text{gold}}) \in \mathcal{Q} \times \mathcal{A} \times \mathcal{T} \times \mathcal{G} \times \text{KB}^{(T)}$ with additional meta information including the degree of *time-sensitivity* is saved as a JSON file. We denote the whole dataset up to T as $\mathcal{D}^{(T)}$. A complete JSON file format for each question can be found in the Appendix (Fig. A17). Once the question template is fixed and the web crawling system is implemented, we can continuously expand the dataset as T progresses, while maintaining high quality.

Finally, unlike other question types, *never-changing* questions are human-curated. After the template is defined, we manually annotate the answers and extract gold contexts. It is worth noting that *never-changing* questions do not require further maintenance. Additional details on the dataset construction are provided in Table A2.

3 Experiment Setup

Knowledge base. For GPT-4o (OpenAI, 2024a), we additionally consider web-based knowledge

base denoted as $\text{KB}_{\text{web}}^{(T)}$, which is defined as the whole web base as of time T .

Retriever Let $\text{Retriever} : \mathcal{Q} \times \text{KB}^{(T)} \mapsto \mathbb{R}^{\geq 0}$ denote a retriever, where for a question q , $d^{(k)} \in \text{KB}^{(T)}$ is the context with the k -th highest relevance score retrieved from $\text{KB}^{(T)}$ using Retriever . In this paper, we employ Contriever (Izacard et al., 2022) and OpenAI’s text embedding model (OpenAI, 2024b) as retrievers. Meanwhile, $d_{\text{web}}^{(k)} \in \text{KB}_{\text{web}}^{(T)}$ denotes the k -th context retrieved from the web via the internal retrieval mechanism of ChatGPT-4o, the version of GPT-4o integrated into ChatGPT.

Language model. We consider both open-source and closed-source LLMs. For open-source LLMs, instruction-tuned variants of the following models are used: Llama-2-7b, Llama-2-70b (Touvron et al., 2023), Llama-3.1-70b (Dubey et al., 2024), Gemma-3-12b, and Gemma-3-27b (Team et al., 2025). We use GPT-4o as a closed-source model. Let $\text{LM}_{\text{base}} : \mathcal{X} \mapsto \mathcal{A}$ denote the base LLMs introduced above. We additionally define a generator $G : \mathcal{Q} \times \text{KB}^{(T)} \mapsto \mathcal{A}$, which takes a question and an up-to-date knowledge base as input.

For each base LLM, we consider the following baselines as follows: (1) a vanilla LLM, i.e., $G_{\text{vanilla}}(q, \text{KB}^{(T)}) := \text{LM}_{\text{base}}(\cdot | q)$, (2) an LLM with a gold passage, i.e., $G_{\text{gold}}(q, \text{KB}^{(T)}) := \text{LM}_{\text{base}}(\cdot | d_{\text{gold}}, q)$, and (3) an LLM with retrieved passages from $\text{KB}^{(T)}$, i.e., $G_{\text{RAG}}(q, \text{KB}^{(T)}) := \text{LM}_{\text{base}}(\cdot | (d^{(k)})_{k=1}^K, q)$. K refers to the number of retrieved documents for inference, where we consider $K = 1$ and $K = 3$. For GPT-4o, we also consider an additional generator that produces answers based on web retrieval: ChatGPT-4o. In the subsequent analysis, we denote each baseline by appending its name as a suffix to the name of a base LLM—for example, Llama-2-7b-Vanilla for the vanilla version of Llama-2-7b. Since multiple retrievers are used, we denote RAG-based baselines by appending the name of the retriever as a suffix.

Knowledge Cutoff. We conducted an analysis over four time intervals with respect to the knowledge cutoff date for each LLM as follows: before, right before, right after, and after. right before refers to dates earlier than the knowledge cutoff within the same calendar year, while right after refers to dates following it.

Evaluation Metric. We let $r : \mathcal{A} \times \mathcal{A} \mapsto [0, 1]$ be a reward function. Then, we define an empirical

reward of a generator G for group $g \in \mathcal{G}$ with respect to the reward r as follows:

$$\mathcal{R}_r(G, g) := \mathcal{R}_r(G; \text{KB}^{(T)}, \mathcal{D}_g^{(T)}) \\ = \frac{1}{|\mathcal{D}_g^{(T)}|} \sum_{(q, a, t, g) \in \mathcal{D}_g^{(T)}} r(G(q, \text{KB}^{(T)}), a),$$

where $\mathcal{D}_g^{(T)} := \{(q, a, t, \tilde{g}) \in \mathcal{D}^{(T)} \mid \tilde{g} = g\}$. Similarly, an empirical reward of a generator G within time interval $[t_1, t_2]$ is defined as follows:

$$\mathcal{R}_r(G, [t_1, t_2]) := \mathcal{R}_r(G; \text{KB}^{(T)}, \mathcal{D}_{[t_1, t_2]}^{(T)}),$$

where $\mathcal{D}_{[t_1, t_2]}^{(T)} := \{(q, a, t, g) \in \mathcal{D}^{(T)} \mid t \in [t_1, t_2]\}$. $\mathcal{R}_r(G, g, [t_1, t_2])$ is defined in a same manner.

For the reward r in this paper, we use a metric from the RapidFuzz library for singleton answers. The metric is based on a fuzzy string matching algorithm derived from the Levenshtein distance (Wagner and Fischer, 1974). Since we also consider questions that return set-valued answers, we additionally propose two evaluation metrics based for rankings and statistics, inspired by the intersection-over-union (IoU) metric (Rezatofighi et al., 2019). For example, for the question “Give me the {top_n} teams in {league_name} at round {round} during the {season_name}, along with their points”, let us assume that an answer is $[[\text{“Team A”}, 8 \text{ pts}], [\text{“Team B”}, 9 \text{ pts}], [\text{“Team C”}, 7 \text{ pts}], [\text{“Team D”}, 6 \text{ pts}]]$, and the generated answer is $[[\text{“Team A”}, 8 \text{ pts}], [\text{“Team B”}, 6 \text{ pts}], [\text{“Team C”}, 9 \text{ pts}], [\text{“Team E”}, 6 \text{ pts}]]$. Then, the IoU-like evaluation metrics for rankings and statistics are defined as follows:

$$\text{IoU}_{\text{rank}} = \frac{(\# \text{ of correct team names})}{(\# \text{ of team names either in answer or generated answer})} \\ \text{IoU}_{\text{stat}} = \frac{(\# \text{ of correct team names \& statistics})}{(\# \text{ of team names either in answer or generated answer})}.$$

Here, $\text{IoU}_{\text{rank}} = \frac{3}{5}$ and $\text{IoU}_{\text{stat}} = \frac{1}{5}$.

Fairness Metric. For the evaluation of the *time-conditional group bias*, we introduce a fairness metric based on the risk parity (Maity et al., 2021) as follows:

$$\mathcal{R}_{r, \text{gap}}(G) := \mathcal{R}_r(G, g_{\text{max}}) - \mathcal{R}_r(G, g_{\text{min}}),$$

where $g_{\text{max}} := \arg\max_{g \in \mathcal{G}} \mathcal{R}_r(G, g)$ and $g_{\text{min}} := \arg\min_{g \in \mathcal{G}} \mathcal{R}_r(G, g)$. By conditioning on the specific time interval $[t_1, t_2]$, we additionally define the *time-conditional group bias* as follows:

$$\mathcal{R}_{r, \text{gap}}^{[t_1, t_2]}(G) := \\ \mathcal{R}_r(G, g_{\text{max}}^{[t_1, t_2]}, [t_1, t_2]) - \mathcal{R}_r(G, g_{\text{min}}^{[t_1, t_2]}, [t_1, t_2]),$$

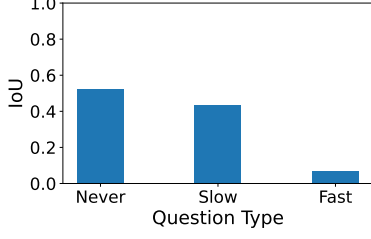


Figure 3: Average performance of GPT-4o-Vanilla by *time-sensitivity*. Here, IoU_{stat} is reported for *slow-changing* and *fast-changing* questions.

where $g_{\max}^{[t_1, t_2]} := \operatorname{argmax}_{g \in \mathcal{G}} \mathcal{R}_r(G, g, [t_1, t_2])$ and $g_{\min}^{[t_1, t_2]} := \operatorname{argmin}_{g \in \mathcal{G}} \mathcal{R}_r(G, g, [t_1, t_2])$.

4 Analysis

Observation 1. *The more time-sensitive the knowledge, the harder it is for LLMs to provide accurate answers.*

Fig. 3 shows the average performance of GPT-4o-Vanilla on three different question types. As reported by Vu et al. (2023); Yang et al. (2024), we see similar patterns that LLMs struggle with answering *fast-changing* questions.

Observation 2 (Necessity of Time-conditional Group Bias Analysis). *In LLMs, time-conditional group bias shows noticeable discrepancies across time, especially between the periods before and after their knowledge cutoffs.*

Besides the discrepancies of the *time-conditional group bias* metric themselves, the best and worst performing groups are different before and after the knowledge cutoff (Fig. 1a). This is the observation that couldn’t be captured by existing *time-marginalized group bias* analysis. It is noteworthy that the *time-conditional group bias* increases after the cutoff. Additional interesting phenomenon is that there exists a group that shows sharp performance decrease after the knowledge cutoff: Brazil.

Fig. 4 shows the performance difference before and after the knowledge cutoff of GPT-4o-Vanilla for each region. Although Brazil is the best performing group before the knowledge cutoff, it shows lower performance than the group-marginalized performance of the model after its cutoff. The subsequent analysis provides a detailed examination of the rationales for *time-conditional group bias* across different time intervals and their differences.

Observation 3 (Cause of Discrepancies in Time-conditional Group Bias). *Unlike time-conditional group bias before the knowledge cutoff, which is driven by parametric knowledge bias, the*

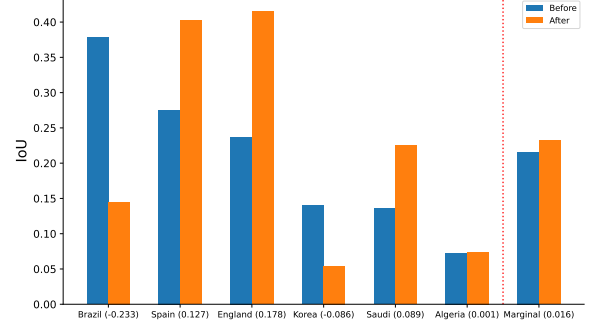


Figure 4: Group-wise performance of GPT-4o-Vanilla before (blue) and after (orange) the knowledge cutoff, in descending order with respect to the performance before the knowledge cutoff.

bias after the cutoff is influenced by both parametric knowledge bias and time-sensitivity bias.

Time-conditional group bias of a vanilla model before the knowledge-cutoff can be understood as a result of the *parametric knowledge bias*. Indeed, Tab. 2 illustrates that group-wise performance of GPT-4o-Vanilla is positively correlated with page view statistics of Wikipedia for each group, presumably due to the bias in the size and quality of the pretraining corpus (Carlini et al., 2023). However, as analysis on the Brazilian league implies (Fig. 4), we hypothesize and empirically verify that there exists a factor in addition to the *parametric knowledge bias* that drives *time-conditional group bias* after the knowledge cutoff: the *time-sensitivity bias*.

Tab. 3 demonstrates that even the same type of knowledge exhibits different *time-sensitivity* across groups. Specifically, Brazil and Algeria are two groups that have most volatile knowledge on the football league table. Analyzed in parallel with Fig. 4, Fig. 5 implies that the *time-sensitivity bias* may be the cause of discrepancies in the *time-conditional group bias* before and after the knowledge cutoff. Among three groups that showed best performance before the knowledge cutoff, Brazil is the only group that shows drastic performance drop after the knowledge cutoff. Football leagues from the other two exhibit low volatility in terms of the league progression (Tab. 3 and Fig. 5), resulting in small difference in model performance between the two time intervals. The problem lies in the fact that the *parametric knowledge bias* and the *time-sensitivity bias* are in some sense irreducible, which implies the difficulty in the group bias mitigation.

Main Problem. *Parametric knowledge bias and time-sensitivity bias are irreducible.*

While there has been a line of research to con-

Parametric Knowledge Time-varying Region		High			Low		
		Slow England	Spain	Fast Brazil	Slow Saudi	Korea	Fast 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

Table 3: Average number of teams that appeared as top or bottom teams at least once through the whole season.

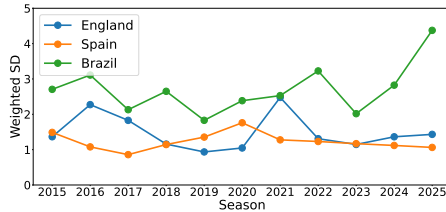


Figure 5: Weighted average of the standard deviations of rankings for teams that placed in the top 3 at least once per year.

tinually pretrain LLMs in order to adapt to the new knowledge, a number of benchmarks reveal that existing continual pretraining methods fail to encode new knowledge into LLMs’ model parameters (Kim et al., 2023; Ko et al., 2024). Meanwhile, for a given topic, the degree of *time-sensitivity* associated with each group’s knowledge varies intrinsically; hence, it cannot be altered through external intervention. RAG can be a promising approach to mitigating *time-conditional group bias*, as follows.

Possible Solution: Retrieval-augmented generation (RAG). *RAG enables an LLM to adapt to new knowledge introduced after its knowledge cutoff in a training-free manner. It also allows the model to adapt to knowledge from before the knowledge cutoff that was not well encoded in its parametric knowledge during pretraining. Note that RAG is applied uniformly to knowledge both before and after the knowledge cutoff.*

While we observe both overall performance gains and group bias mitigation across all RAG setups in comparison to the vanilla performance (Fig. 6), we observe that the discrepancy in patterns of *time-conditional group bias* still persists (Fig. 1b). This once again highlights the importance of *time-conditional group bias* analysis.

In addition, we observe that RAG result based on web retrieval shows exceptional pattern compared to other RAG baselines, both before and after the knowledge cutoff. In the pre-cutoff period, while ChatGPT-4o shows an overall performance gain of 0.2, group bias is also exacerbated by 0.2 (Fig. 6a). Although we do not have access to the correctness labels of the web retrieval, an increase in group

bias implies that the retrieval results are biased toward the best performing groups in terms of vanilla performance. On the other hand, in the post-cutoff period, ChatGPT-4o achieves overall performance gains while mitigating group bias, which appears satisfactory at first glance (Fig. 6b). Nonetheless, there is a group that suffers from a substantial performance drop of 0.19 after web retrieval.

Considering the oracle result using gold passages achieves near-perfect accuracy, such a performance drop can be understood as a knowledge conflict phenomenon (Xu et al., 2024; Xie et al., 2024), which refers to the misalignment of the parametric knowledge of LLMs and a given context. This result is quite counterintuitive, since the performance drop occurred after the knowledge cutoff, a point at which we would normally expect performance gains from applying RAG to new knowledge for which the base LLM was not pretrained. This phenomenon may be partially attributable to the forecasting capability of high-capacity LLMs such as GPT-4o, and we coin the term *forecasting conflict* to refer to cases in which the forecasting output of an LLM contradicts retrieved documents, framing it as a special case of knowledge conflict.

Properties of Fair and Accurate RAG. *Ideal RAG-based LLMs must satisfy the following properties: high overall performance, low group bias, and monotonic performance improvement for every group.*

While most existing literature on RAG focuses on improving overall performance compared to the vanilla model and other RAG baselines (Asai et al., 2023; Jeong et al., 2024), results from Fig. 6 demonstrate that such performance gains do not necessarily improve group fairness. Furthermore, group-marginalized performance evaluation, like the overall performance metric, does not capture performance gains at the group level, which can decrease even when the group-marginalized performance increases. Therefore, we argue that an ideal RAG-based LLM construction problem be formulated as a multi-objective optimization problem.

More specifically, the model should aim not only to improve overall performance but also to mitigate group bias, while ensuring that group-wise performance gains remain non-negative across all groups. Oracle RAG results using gold passages, which show trends diametrically opposed to those from web retrieval, demonstrate the ideal property. Even in comparison with other RAG baselines, the oracle

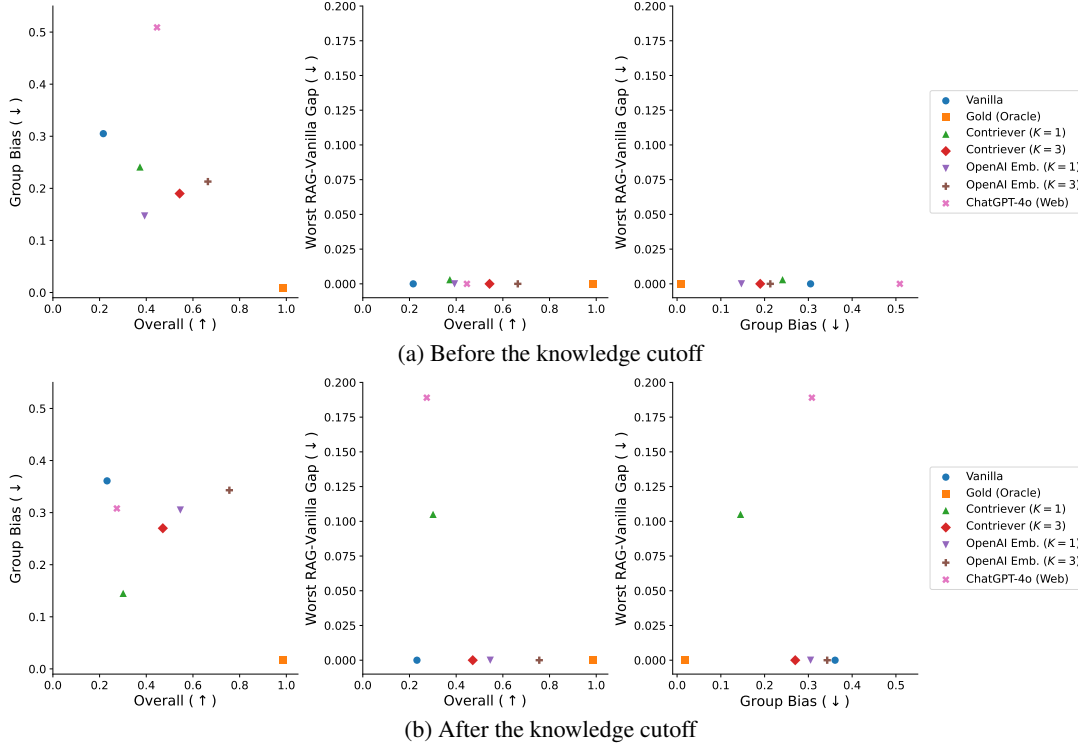


Figure 6: Pairwise scatterplots of **Group Bias** (↓), **Overall Performance** (↑), and **Worst RAG-Vanilla Gap** (↓) of GPT-4o. **Overall Performance** refers to the group-wise model performance. **Worst RAG-Vanilla Gap** indicates the largest performance drop caused by RAG. It is 0 if RAG improved performance for all groups.

results show both superior overall performance and a group bias close to 0. Drawing on the analysis of this oracle result and the inherent problem of *time-sensitive knowledge* in RAG, as will be explained, we conclude our study with two key takeaways and actionable guidelines for developing fair and accurate RAG-based LLMs.

Takeaway Message 1. *Retrieval performance should be inversely biased relative to that of a vanilla LLM.*

The message may seem counterintuitive in the sense that a retriever should be “unfair” to construct a “fair” RAG-based LLMs. Unlike the oracle scenario where gold passages are retrieved for all queries irrespective of the group, it is impossible to construct such retrievers in reality. Still, the oracle result provides us a clue that may enable fair RAG performance even with imperfect retrieval.

Specifically, while the retrieval performance is fair with hit ratio of 1, there exists a bias in terms of performance gains across groups due to retrieval (Fig. 7). This implies that minority groups have enjoyed more performance gains than majority groups have, which in turn has resulted in group bias close to 0. We additionally note that such reverse unfairness should be aligned with a baseline LLM in an end-to-end manner, since the group bias in vanilla performance varies across LLMs

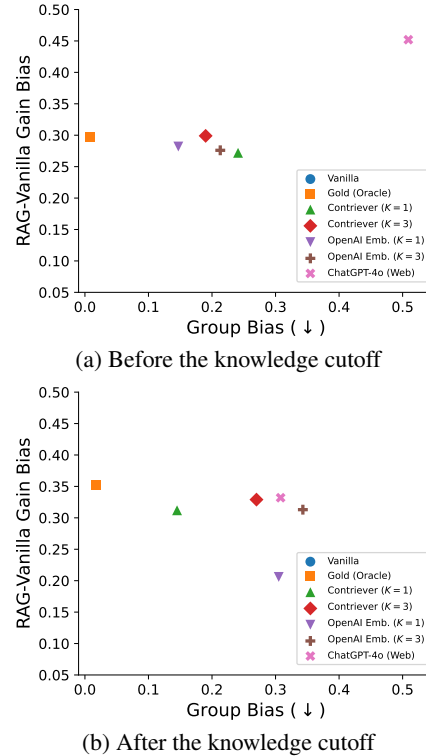


Figure 7: Scatterplots of **Group Bias** (↓) and **RAG-Vanilla Gain Bias** (↓) of GPT-4o. **RAG-Vanilla Gain Bias** measures the disparity between the maximum and minimum performance improvements from RAG.

(Appendix A9-A13).

Takeaway Message 2. *Even when the same type of retrieval error occurs, time-sensitivity bias*

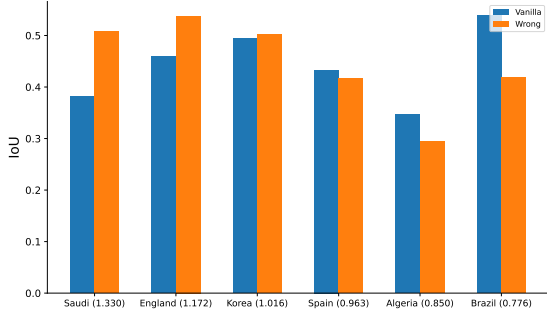


Figure 8: Group-wise performance of GPT-4o-Vanilla (blue) and GPT-4o-Contriever ($K = 1$) (orange) evaluated only on queries with incorrect retrieval.

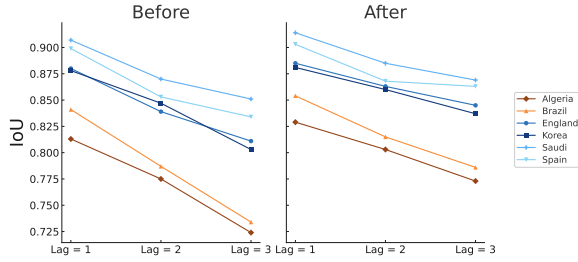


Figure 9: Semantic similarity between gold and wrong passages on top 5 teams, measured by the IoU metric. The incorrect passages are those that considered gold passages at t time steps before (left) or after (right) the current point ($t = 1, 2, 3$).

causes variation in how much useful partial information is present in the retrieved document.

Fig. 8 compares the vanilla and RAG performance of GPT-4o only for queries with wrong retrieval. Brazil and Algeria—groups where knowledge on football league standings is highly volatile compared to other groups—show the most drastic performance drop when provided with incorrect information. As pointed out by Ouyang et al. (2025), documents on *time-sensitive* knowledge are accumulated in the knowledge base over time, and there exist multiple documents on the same type of knowledge whose correct factual statements vary depending on the temporal context considered.

Therefore, we hypothesize that such disparate performance drops across groups result from different amounts of partial information for generating correct answers contained in incorrect documents, which is caused by *time-sensitivity bias*. Since we cannot control the type of wrong retrieval using an open-source retriever, we conducted an additional controlled experiment to validate our hypothesis. Fig. 9 evaluates the document similarity between the gold passage and an incorrect passage using the IoU_{rank} metric. In this experiment, we control the type of wrong passage uniformly across all groups. Specifically, we compare the results of incorrect passages that contain answers occur-

ring before or after the specific time points defined by the knowledge in the gold passage across different regions. Aligned with our intuition, Brazil and Algeria exhibit the lowest similarity with gold passages across all time lags.

5 Conclusion

In this paper, we propose a continuously expandable benchmark, ChronoBias, by means of a template-based semi-automatic generation method, which effectively addresses the trade-off between quality and quantity. With labels of group, *time-sensitivity*, and the date of knowledge update, we demonstrate with ChronoBias that it is imperative to conduct *time-conditional group bias* evaluation on the *time-sensitive* knowledge of LLMs for holistic fairness evaluation.

Specifically, *time-conditional group bias* shows notable shifts on either side of the knowledge cut-off of a given LLM. These shifts involve not only which groups are the best- and worst-performing, but also in the values of the group bias metrics themselves—patterns that cannot be captured by existing *time-marginalized group bias* analysis. While *parametric knowledge bias* is the primary cause of performance disparities across groups, we find that *time-sensitivity bias* is an additional source of *group bias*, which manifests only after the knowledge cutoff and thereby leads to variations in *time-conditional group bias* across different time.

Since *parametric knowledge bias* and *time-sensitivity bias* are irreducible, RAG is considered an alternative for accurate and fair language generation, as it can adapt both to new knowledge and to pre-cutoff knowledge that is attenuated in model parameters. By comparing RAG baselines with oracle results using gold passages, we propose actionable guidelines for building ideal RAG-based LLMs. In particular, group-wise retrieval performance should be adjusted to negatively correlate with vanilla model performance, while acknowledging that *time-sensitivity bias* makes the amount of partial information in incorrect retrieval vary across groups.

Although we define group in terms of geographical region in the sports domain, our analysis can be generalized to any categorization and topic subject to *parametric knowledge bias* and *time-sensitivity bias*. We hope our benchmark serves as a stepping stone for future work emphasizing that holistic evaluation and mitigation of *group bias* must take the temporal dimension into account.

Limitations

To empirically solidify the generality of our claim, it would be better to scale up the dataset in terms of size and the diversity. Related to this, it would be interesting to consider datasets from different domains, where the distribution is dissimilar. For instance, we can consider the legal domain, where there are inherently more static knowledge (e.g., constitutional law) and less *time-sensitive* knowledge (e.g., case laws, constitutional amendment). Moreover, evaluating temporal group bias on more complex questions that require processing multiple retrieved documents (Jimenez Gutierrez et al., 2024) would be another interesting direction for scaling up ChronoBias. Finally, since we focus on emphasizing the importance of measuring temporal group bias and analyzing its underpinnings, we leave the development of fair and accurate RAG-based LLMs over time for future work.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH) (20%); RS-2024-00457882, National AI Research Lab Project (20%); No. RS-2024-00509258 and No. RS-2024-00469482, Global AI Frontier Lab (30%); RS-2025-00560062 (30%)).

References

- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*.
- Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT Press.
- Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *ICDM Workshop*.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *ICLR*.
- Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, et al. 2024. CulturalBench: A robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of LLMs. *arXiv preprint arXiv:2410.02677*.
- Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163.
- Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *SIGKDD*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.
- Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *NeurIPS*.
- Yufei Huang and Deyi Xiong. 2024. CBBQ: A Chinese bias benchmark dataset curated with human-ai collaboration for large language models. In *LREC-Coling*.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. *Unsupervised dense information retrieval with contrastive learning*. *Transactions on Machine Learning Research*.
- Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *NAACL-HLT*.
- Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *NeurIPS*.
- Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.
- Yujin Kim, Jaehong Yoon, Seonghyeon Ye, Sangmin Bae, Namgyu Ho, Sung Ju Hwang, and Se-Young Yun. 2023. Carpe Diem: On the evaluation of world knowledge in lifelong language models. In *NAACL-HLT*.

- Dayoon Ko, Jinyoung Kim, Hahyeon Choi, and Gunhee Kim. 2024. GrowOVER: How can LLMs adapt to growing real-world knowledge? In *ACL*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. [Holistic evaluation of language models](#). *Transactions on Machine Learning Research*.
- Subha Maity, Debarghya Mukherjee, Mikhail Yurochkin, and Yuekai Sun. 2021. Does enforcing fairness mitigate biases caused by subpopulation shift? In *NeurIPS*.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *ACL*.
- Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? Measuring cultural bias in large language models. In *ACL*.
- OpenAI. 2024a. [GPT-4 technical report](#). *Preprint*, arXiv:2303.08774.
- OpenAI. 2024b. OpenAI text-embedding-3-small model. <https://platform.openai.com/docs/guides/embeddings>.
- Jie Ouyang, Tingyue Pan, Mingyue Cheng, Ruirui Yan, Yucong Luo, Jiaying Lin, and Qi Liu. 2025. HoH: A dynamic benchmark for evaluating the impact of outdated information on retrieval-augmented generation. In *ACL*.
- Hamid RezaTofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*.
- Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiawat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. 2024. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. FreshLLMs: Refreshing large language models with search engine augmentation. In *ACL Findings*.
- Robert A Wagner and Michael J Fischer. 1974. The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173.
- Haotian Xia, Zhengbang Yang, Yuqing Wang, Rhys Tracy, Yun Zhao, Dongdong Huang, Zezhi Chen, Yan Zhu, Yuan-Fang Wang, and Weining Shen. 2024. SportQA: A benchmark for sports understanding in large language models. In *NAACL-HLT*.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. [Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts](#). In *ICLR*.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *EMNLP*.
- Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Gui, Ziran Jiang, Ziyu Jiang, et al. 2024. CRAG-Comprehensive RAG benchmark. In *NeurIPS*.
- Michael JQ Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *EMNLP*.

Appendix

A Related Work

Fairness in Machine Learning. As machine learning models are used in decision-making (Barocas et al., 2023), there has been research to impose fairness to machine learning models in either the training or post-training steps (Corbett-Davies et al., 2017; Hardt et al., 2016; Calders et al., 2009; Chouldechova, 2017). Bias in these models can result in a biased decision toward a specific subpopulation, which can amplify the existing decision bias. The problem becomes more problematic when the group variable is related to demographic features, since it can block specific ethnic groups’ or races’ opportunity or amplify the stereotypes (Barocas et al., 2023; Gallegos et al., 2024).

Fairness in NLP. They mainly consider two different types of biases: consequential and representational biases (Gallegos et al., 2024; Naous et al., 2024). The former refers to the evaluation of group bias based on the performance on downstream tasks. The latter measures the distributional bias, where there exist no correct answers on a given query (Jin et al., 2024; Huang and Xiong, 2024; Naous et al., 2024; Liang et al., 2023). For example, given two different prompts as “Her job is...” and “His job is...”, the difference of two next token probability distributions is measured using divergence measures such as KL-divergence (Liang et al., 2023). In this paper, we only consider the consequential bias. Specifically, we set geographical region as a group variable, with open-ended question answering tasks focused on the sports domain, in line with the domain focus of Xia et al. (2024).

Continuous Adaptation of LLMs. Liang et al. (2023) acknowledge that current holistic evaluation of LLMs lacks assessment of whether they can adapt to the changing nature of the world—an important aspect, given that these models are trained on static corpora. While there has been RAG and its variants to cope with the evolving nature of the world (Lewis et al., 2020; Asai et al., 2023), a number of continual knowledge learning benchmarks have been proposed to see whether a continual pretraining method exists to explicitly update the parametric knowledge of LLMs to adapt to the new knowledge, while retaining the static information (Kim et al., 2023; Ko et al., 2024). Since our dataset is expanding and evolving over time,

we can continuously evaluate whether LLMs can adapt to evolving knowledge, where the degree of adaptation may vary across groups, which is our problem of interest.

B Dataset Statistics

	Fast-Changing	Slow-Changing	Never-Changing
# of question	13,082	180	150
# of question type	6	3	5
# of region	6	6	6
# of time stamp	1,594	57	-

Table A1: Descriptive statistics of the ChronoBias benchmark

C Experimental Details

GPT-4o inference was conducted using both the OpenAI’s ChatGPT web interface (ChatGPT-4o) and the OpenAI API (GPT-4o-Vanilla, GPT-4o-Contriever, and GPT-4o-OpenAI). For the inference using ChatGPT-4o, we have used three ChatGPT Pro accounts (200\$ each) for inference. Inference for open-source LLMs was performed locally using four NVIDIA A100 GPUs with 80GB memory each.

D Annotation on Never-changing Questions

The first two authors manually conducted the data annotation process for *never-changing* questions. After creating templates for the *never-changing* questions, they searched the web to annotate answers and gold passages, which were cross-checked for accuracy. Both annotators are Asian males aged 20 to 35.

E Additional Details on Question Construction

Slow-changing questions on top scorer tables use the following template: “Name the top {top_n} scorer in {league name} at the end of the {season}”. If the current season has not ended, the answer is labeled based on the top scorers from the most recently completed season. Except for the template construction, the remainder of the data generation follows exactly the same process as *fast-changing* questions.

All the question templates used for the benchmark construction are provided in Table A2.

F Additional Analysis on *Time-sensitivity Bias*

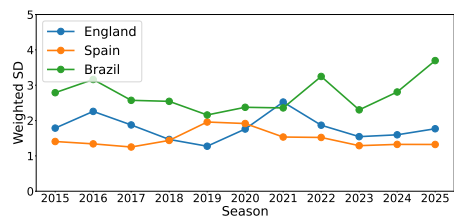


Figure A1: Weighted average of the standard deviations of rankings for teams that placed in the top 5 at least once per year.

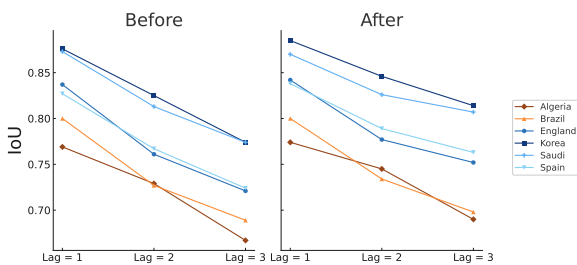


Figure A2: Semantic similarity between gold and wrong passages on bottom 5 teams, measured by the IoU metric. The incorrect passages are those that considered gold passages at t time steps before (left) or after (right) the current point ($t = 1, 2, 3$).

G Qualitative Analysis on Large-scale Llama Models

Fig. A18 presents a failure case from large-scale Llama models, which show different generation patterns to those of GPT-4o when given a wrong context. Specifically, they neither forecast nor infer the answer when provided with the wrong context. This behavior is similar to what was reported in (Grattafiori et al., 2024), where the model is fine-tuned to avoid responding when it is uncertain or lacks knowledge. From these observations, we can infer that such a fine-tuning approach may hinder the model’s forecasting and reasoning ability.

H Group Bias Analysis of ChatGPT-4o

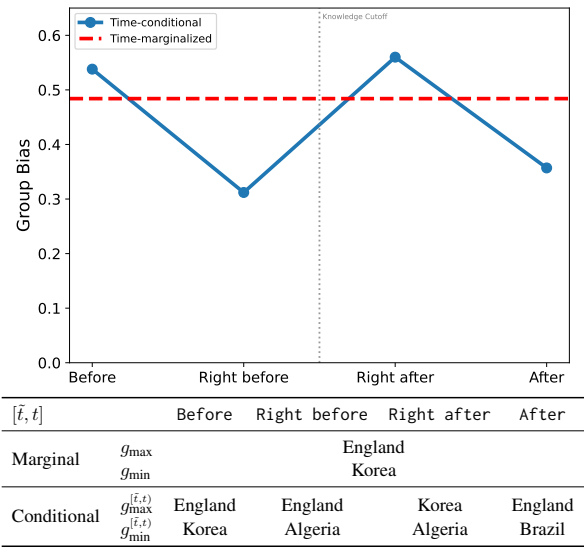
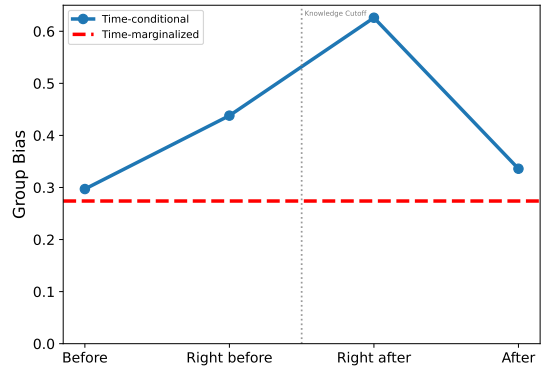
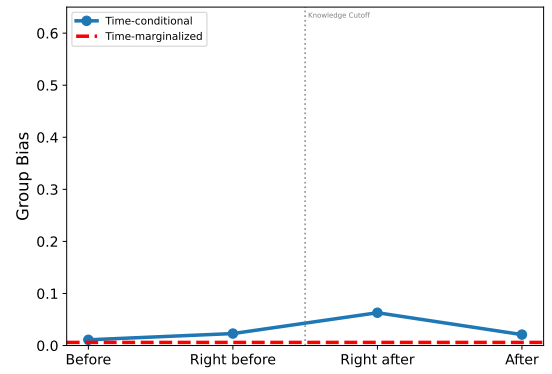


Figure A3: *Time-marginalized group bias and time-conditional group bias* of ChatGPT-4o.



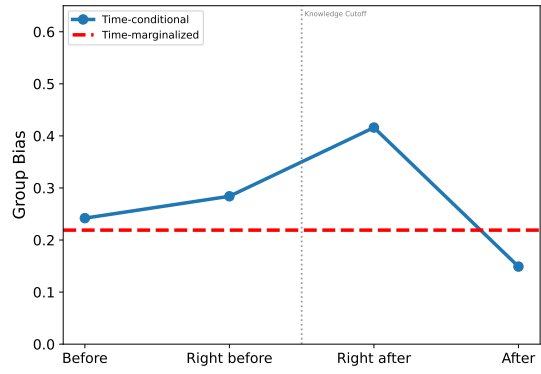
$[t_1, t_2]$	Before	Right before	Right after	After
Marginal	$g_{\max}$	Brazil Algeria		
Conditional	$g_{\min}^{(t_1, t_2)}$	Brazil Algeria	Spain Algeria	England Korea Spain Korea

(a) GPT-4o-Vanilla



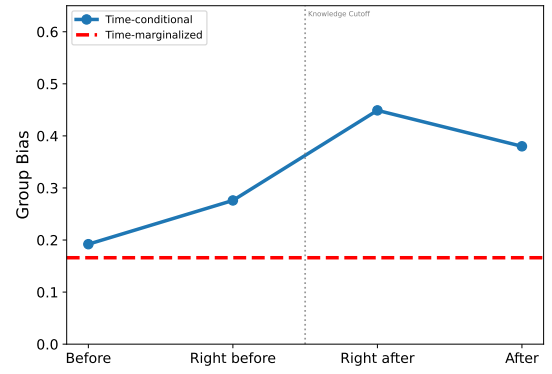
$[t_1, t_2]$	Before	Right before	Right after	After
Marginal	$g_{\max}$	England Korea		
Conditional	$g_{\min}^{(t_1, t_2)}$	England Algeria	Algeria England	Korea Spain Algeria Korea

(b) GPT-4o-Gold



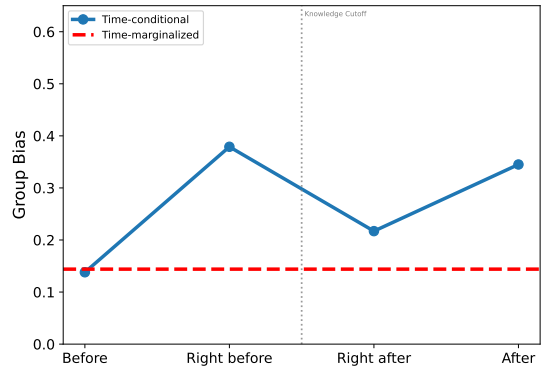
$[t_1, t_2]$	Before	Right before	Right after	After
Marginal	$g_{\max}$	England Algeria		
Conditional	$g_{\min}^{(t_1, t_2)}$	England Algeria	England Saudi	Saudi Korea Spain Korea

(c) GPT-4o-Contriever (Top-1)



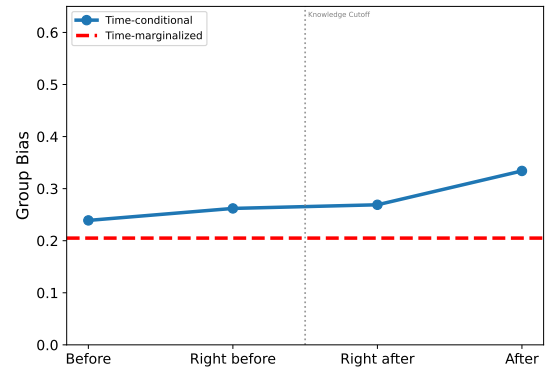
$[t_1, t_2]$	Before	Right before	Right after	After
Marginal	$g_{\max}$	England Algeria		
Conditional	$g_{\min}^{(t_1, t_2)}$	England Algeria	England Brazil	Saudi Spain Spain Korea

(d) GPT-4o-Contriever (Top-3)



$[t_1, t_2]$	Before	Right before	Right after	After
Marginal	$g_{\max}$	Spain Saudi		
Conditional	$g_{\min}^{(t_1, t_2)}$	Spain Saudi	Algeria Saudi	England Brazil England Saudi

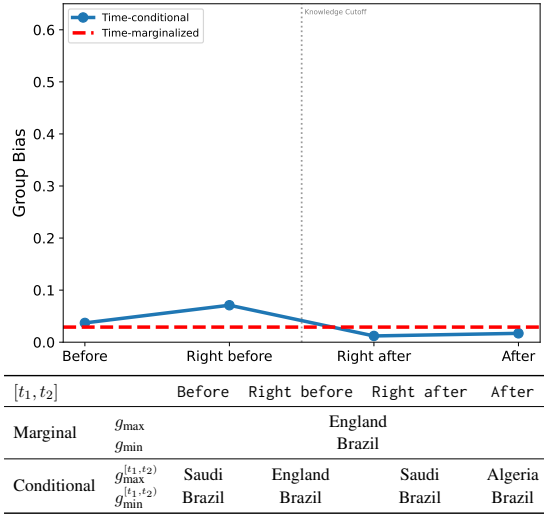
(e) GPT-4o-OpenAI (Top-1)



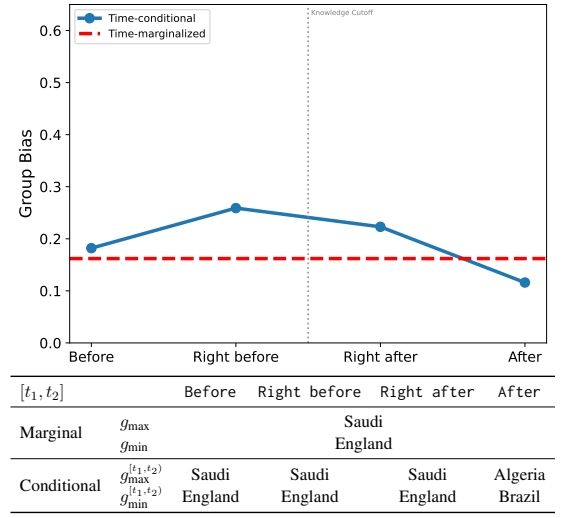
$[t_1, t_2]$	Before	Right before	Right after	After
Marginal	$g_{\max}$	England Saudi		
Conditional	$g_{\min}^{(t_1, t_2)}$	England Saudi	Algeria Korea	England Korea England Korea

(f) GPT-4o-OpenAI (Top-3)

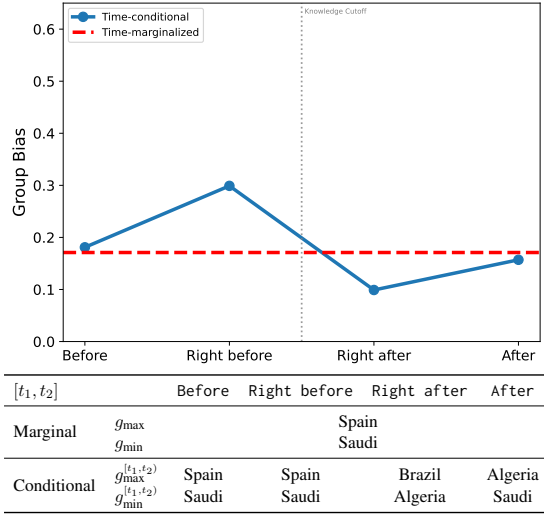
Figure A4: Time-marginalized group bias and time-conditional group bias of GPT-4o across different model configurations.



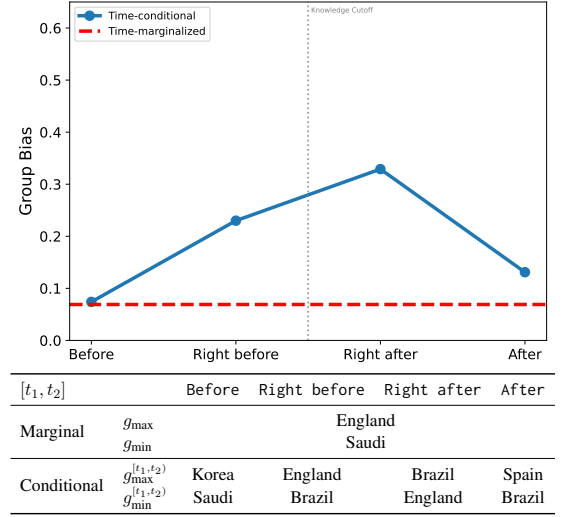
(a) Llama-2-7b-Vanilla



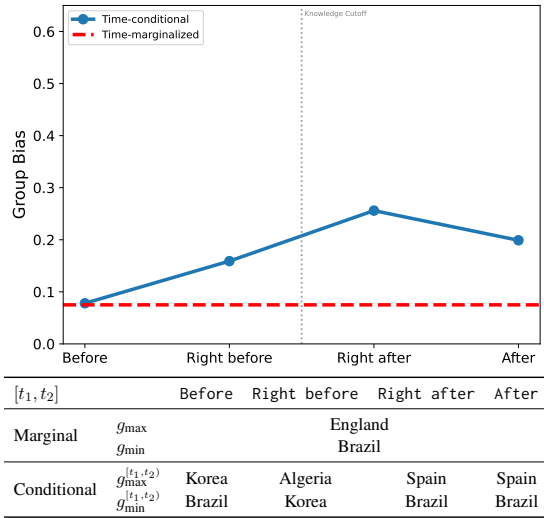
(b) Llama-2-7b-Gold



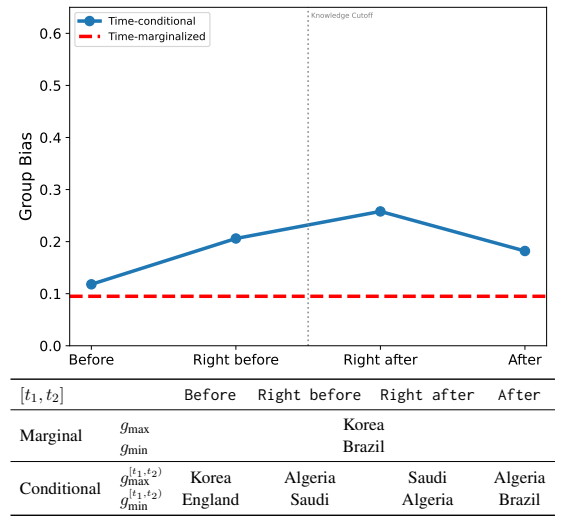
(c) Llama-2-7b-Contriever (Top-1)



(d) Llama-2-7b-Contriever (Top-3)

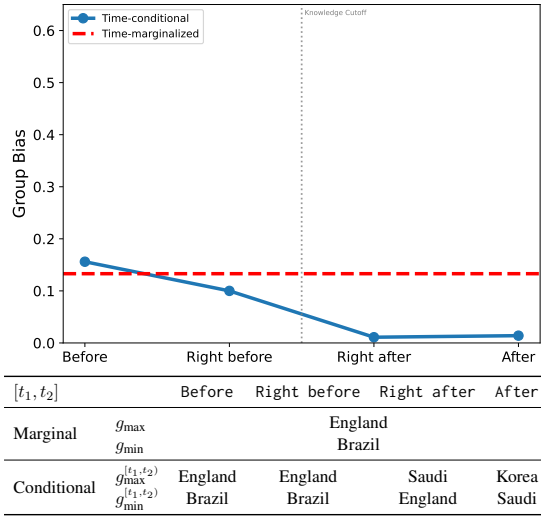


(e) Llama-2-7b-OpenAI (Top-1)

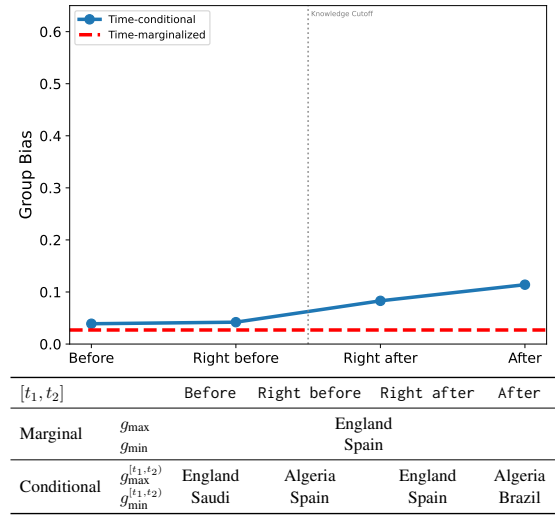


(f) Llama-2-7b-OpenAI (Top-3)

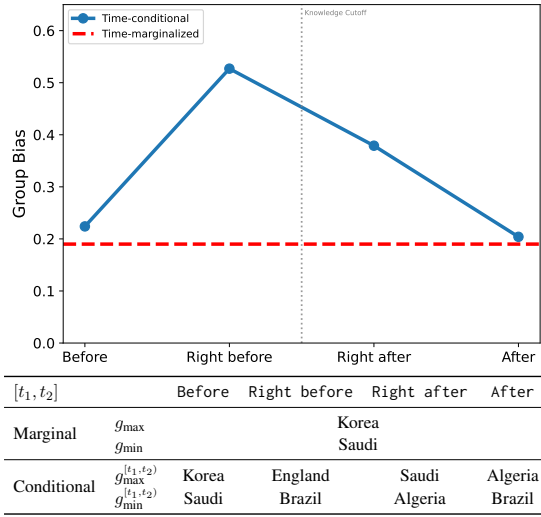
Figure A5: Time-marginalized group bias and time-conditional group bias of Llama2-7b across different model configurations.



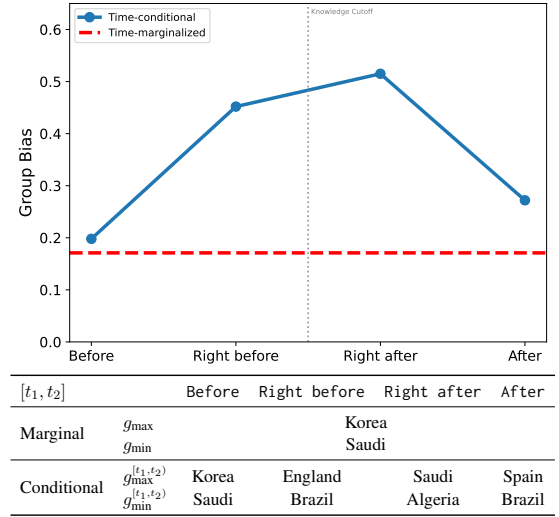
(a) Llama-3.1-70b-Vanilla



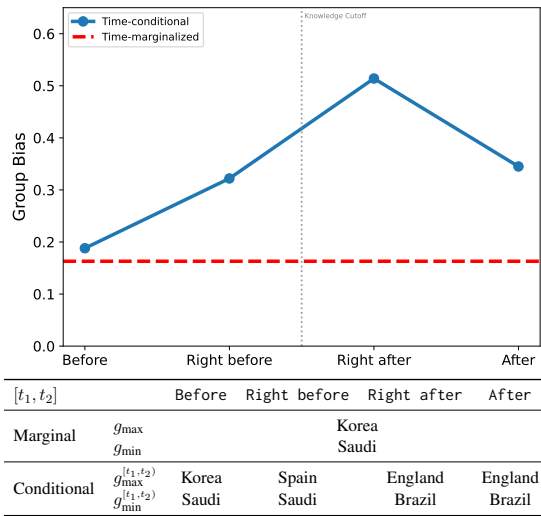
(b) Llama-3.1-70b-Gold



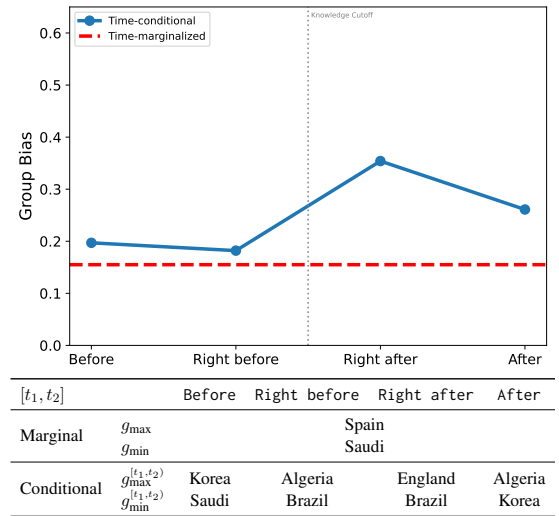
(c) Llama-3.1-70b-Contriever (Top-1)



(d) Llama-3.1-70b-Contriever (Top-3)

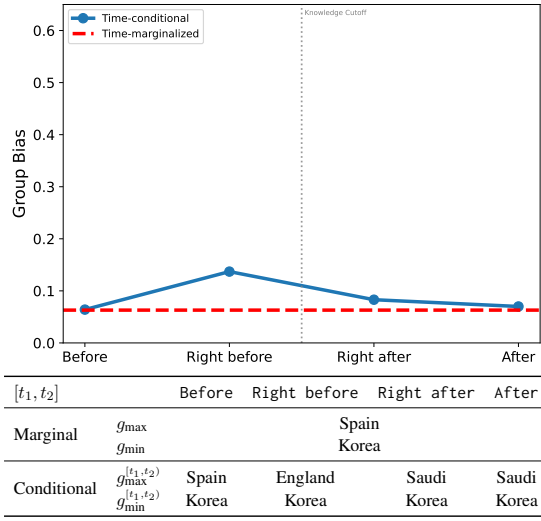


(e) Llama-3.1-70b-OpenAI (Top-1)

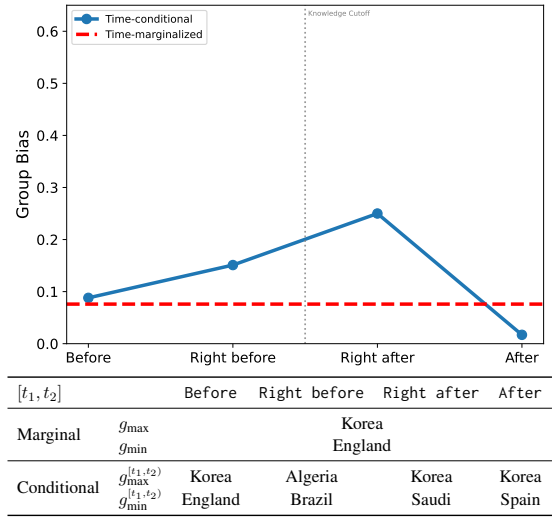


(f) Llama-3.1-70b-OpenAI (Top-3)

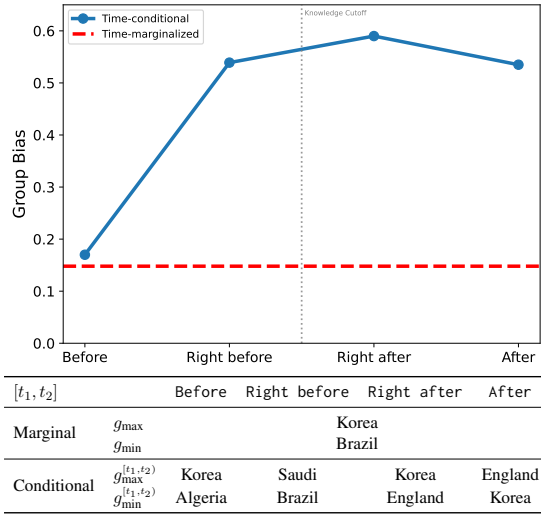
Figure A6: Time-marginalized group bias and time-conditional group bias of Llama-3.1-70b across different model configurations.



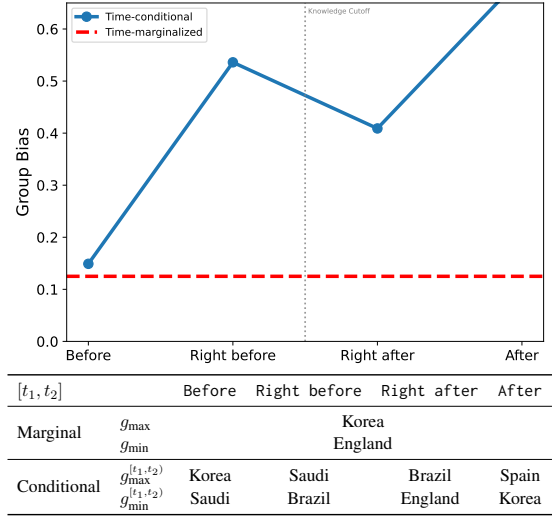
(a) Gemma-3-12b-Vanilla



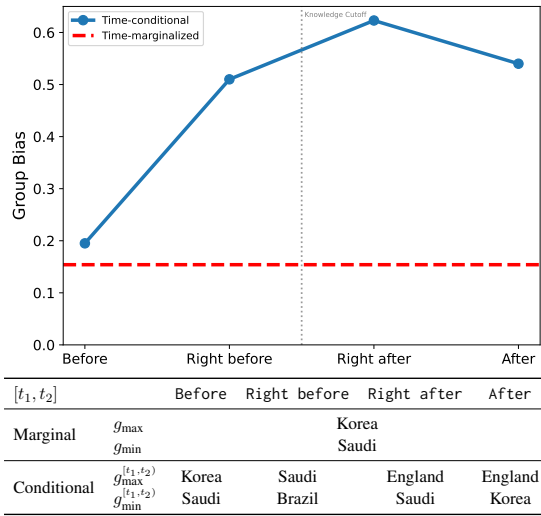
(b) Gemma-3-12b-Gold



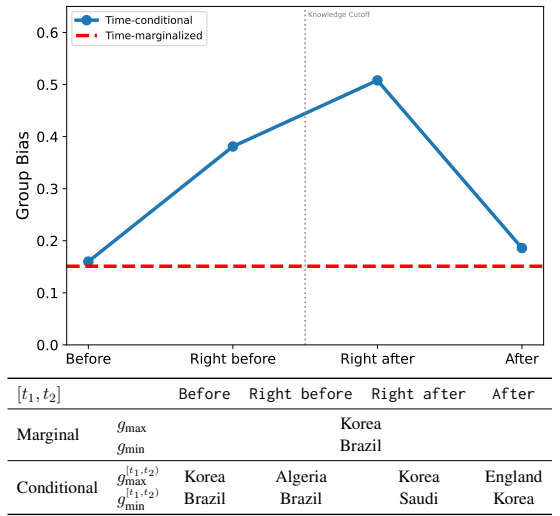
(c) Gemma-3-12b-Contriever (Top-1)



(d) Gemma-3-12b-Contriever (Top-3)

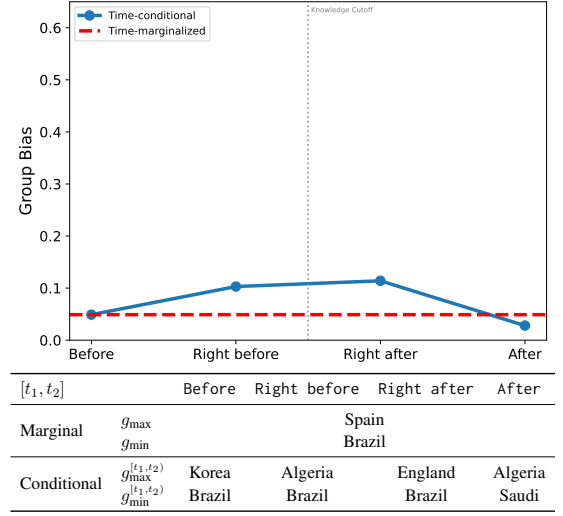
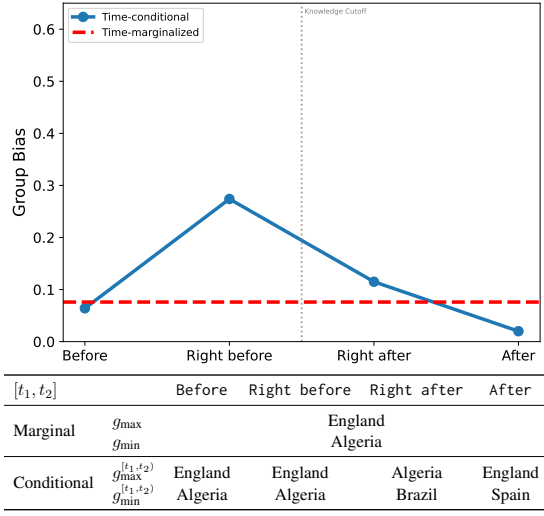


(e) Gemma-3-12b-OpenAI (Top-1)



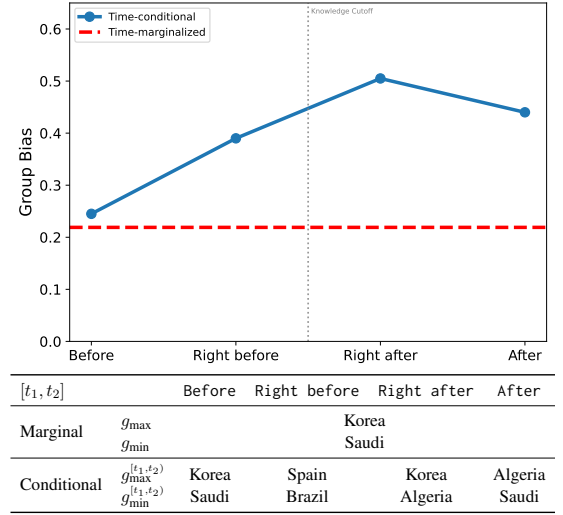
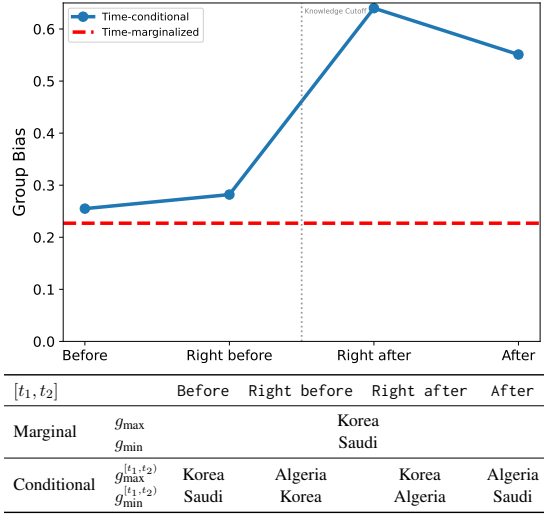
(f) Gemma-3-12b-OpenAI (Top-3)

Figure A7: Time-marginalized group bias and time-conditional group bias of Gemma-3-12b across different model configurations.



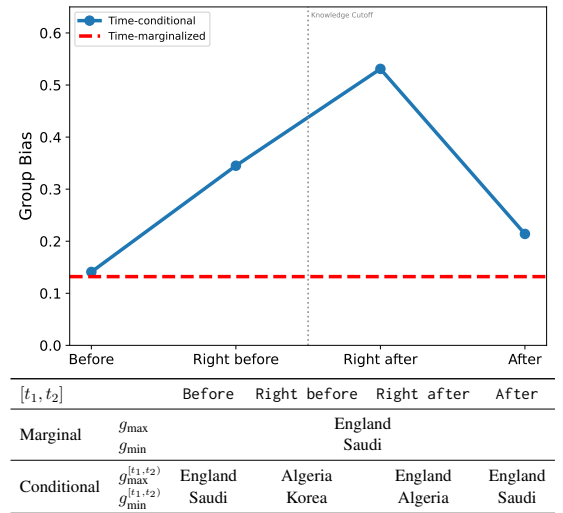
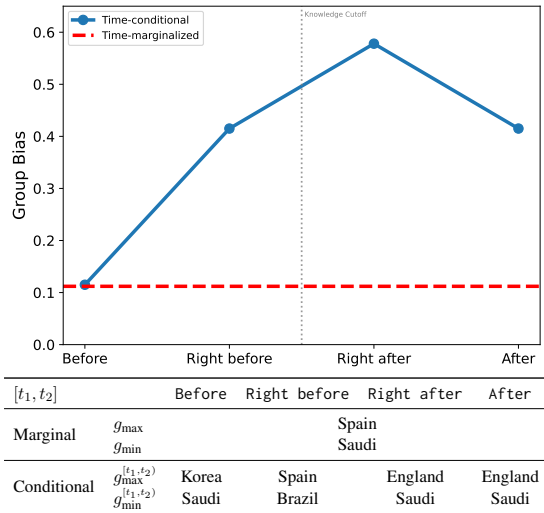
(a) Gemma-3-27b-Vanilla

(b) Gemma-3-27b-Gold



(c) Gemma-3-27b-Contriever (Top-1)

(d) Gemma-3-27b-Contriever (Top-3)



(e) Gemma-3-27b-OpenAI (Top-1)

(f) Gemma-3-27b-OpenAI (Top-3)

Figure A8: Time-marginalized group bias and time-conditional group bias of Gemma-3-27b across different model configurations.

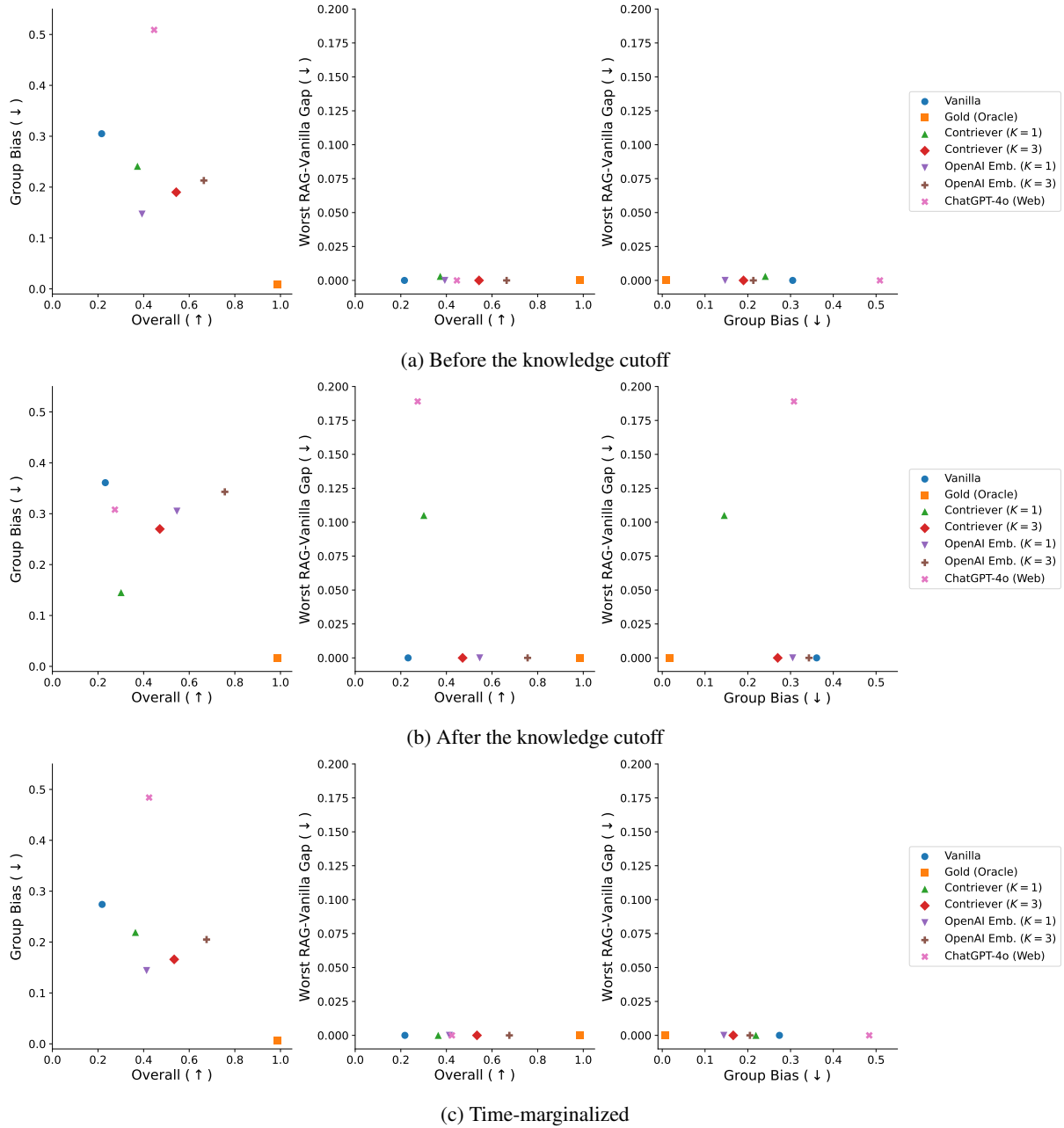


Figure A9: Pairwise scatterplots of **Group Bias** ($\downarrow$), **Overall Performance** ($\uparrow$), and **Worst RAG-Vanilla Gap** ($\downarrow$) of GPT-4o. **Overall Performance** refers to the group-wise model performance. **Worst RAG-Vanilla Gap** indicates the largest performance drop caused by RAG. It is 0 if RAG improved performance for all groups.

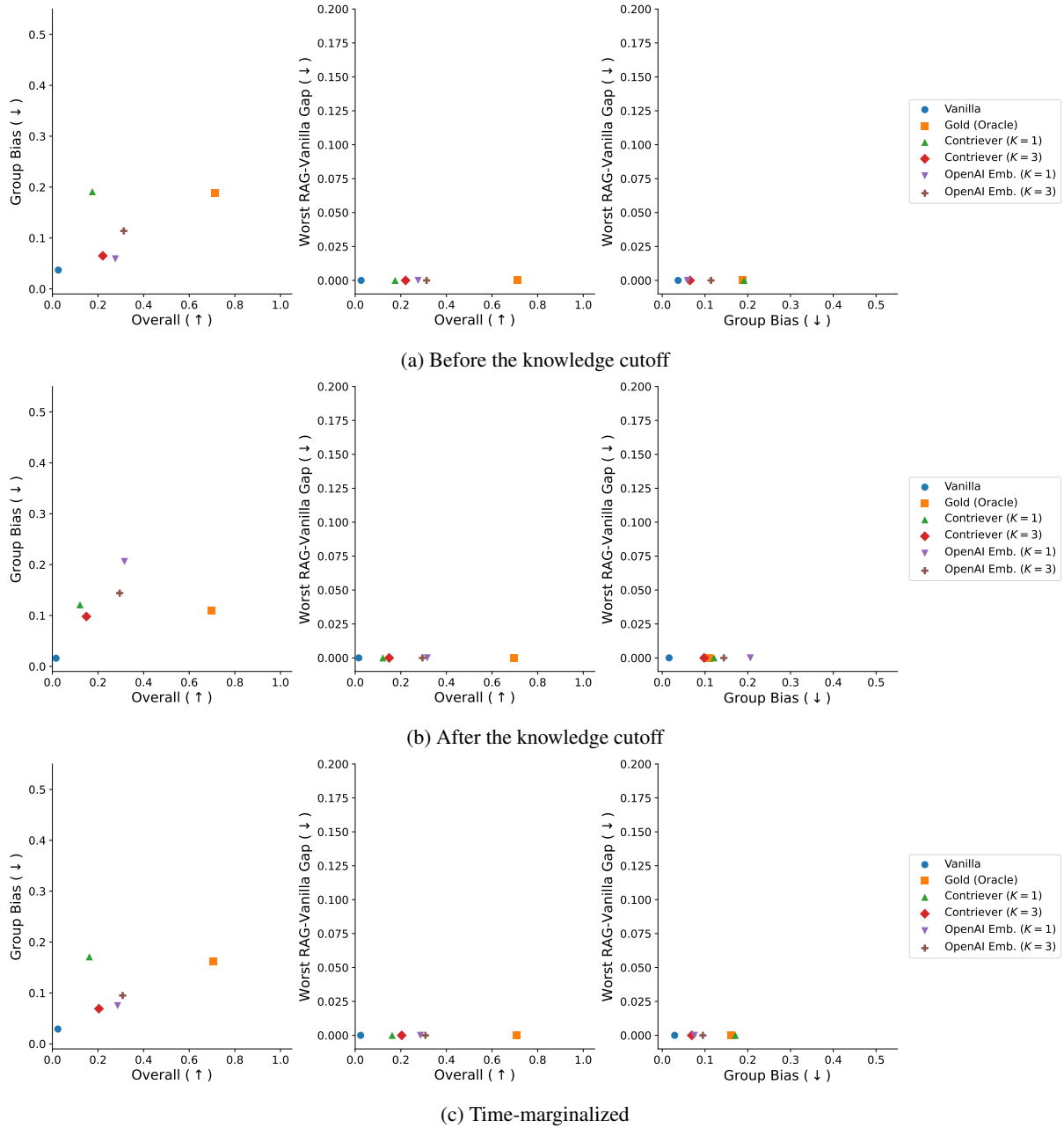


Figure A10: Pairwise scatterplots of **Group Bias** (↓), **Overall Performance** (↑), and **Worst RAG-Vanilla Gap** (↓) of Llama-2-7b. **Overall Performance** refers to the group-wise model performance. **Worst RAG-Vanilla Gap** indicates the largest performance drop caused by RAG. It is 0 if RAG improved performance for all groups.

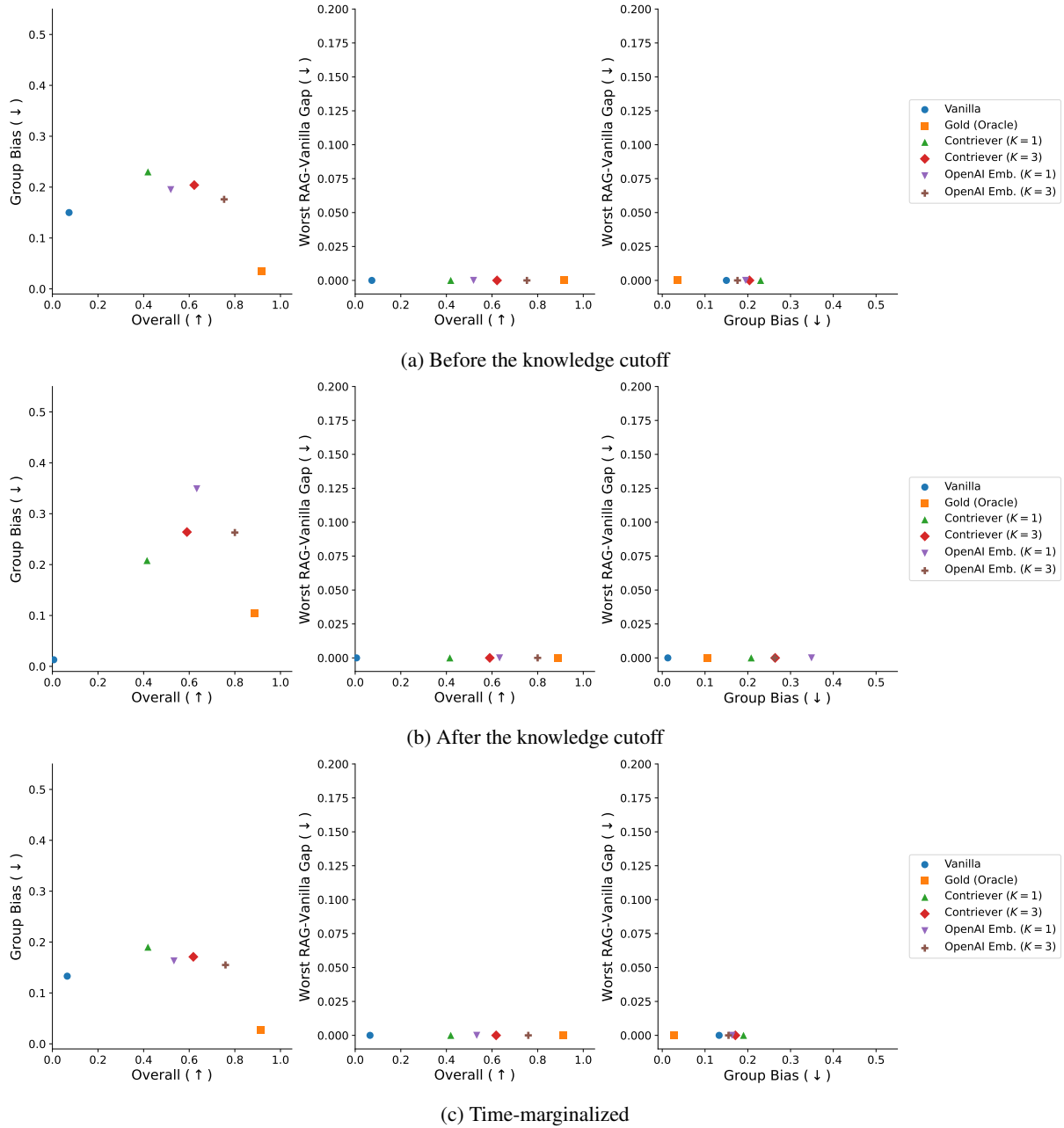


Figure A11: Pairwise scatterplots of **Group Bias** (↓), **Overall Performance** (↑), and **Worst RAG-Vanilla Gap** (↓) of Llama-3.1-70b. **Overall Performance** refers to the group-wise model performance. **Worst RAG-Vanilla Gap** indicates the largest performance drop caused by RAG. It is 0 if RAG improved performance for all groups.

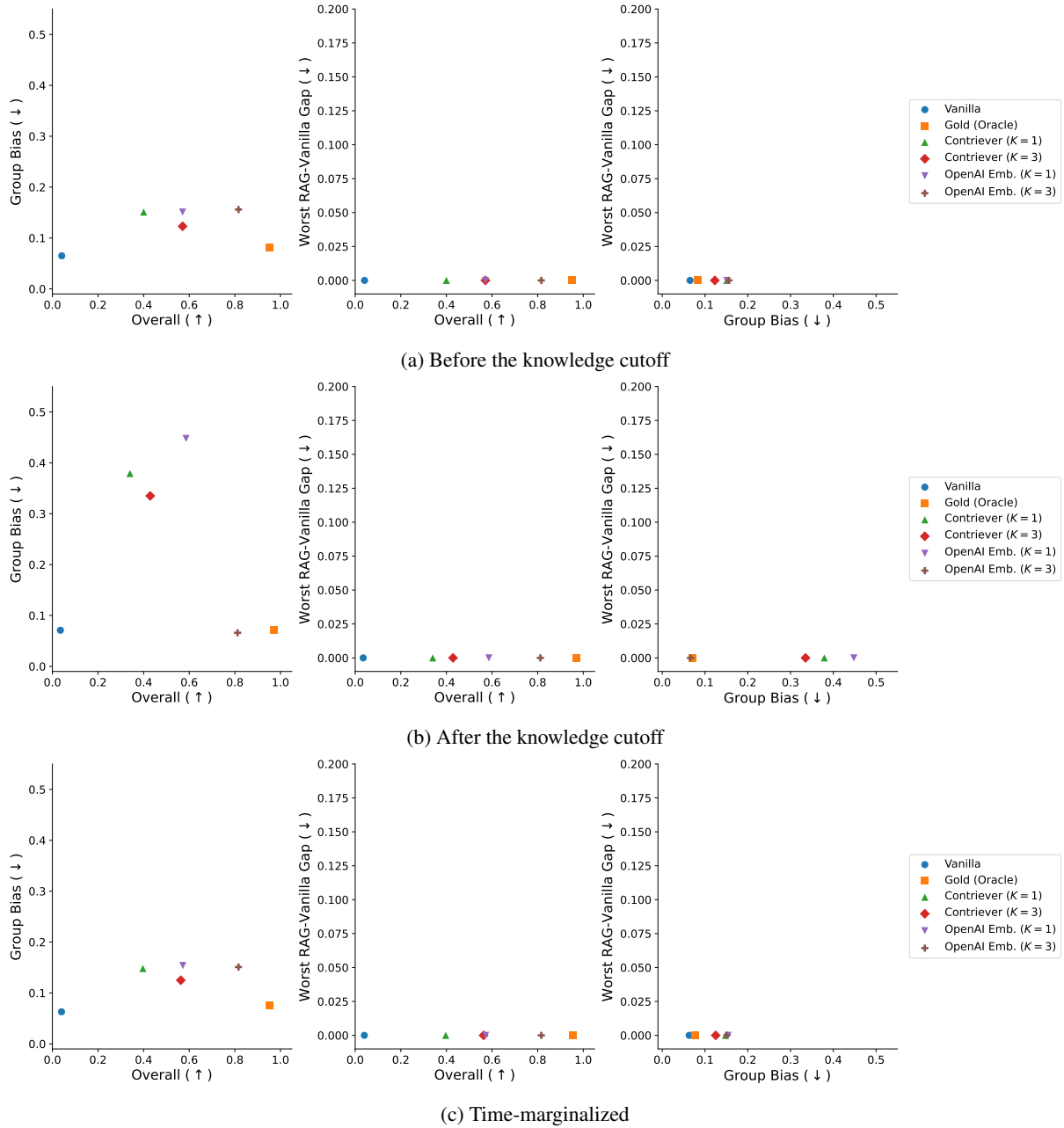


Figure A12: Pairwise scatterplots of **Group Bias** (↓), **Overall Performance** (↑), and **Worst RAG-Vanilla Gap** (↓) of Gemma-3-12b. **Overall Performance** refers to the group-wise model performance. **Worst RAG-Vanilla Gap** indicates the largest performance drop caused by RAG. It is 0 if RAG improved performance for all groups.

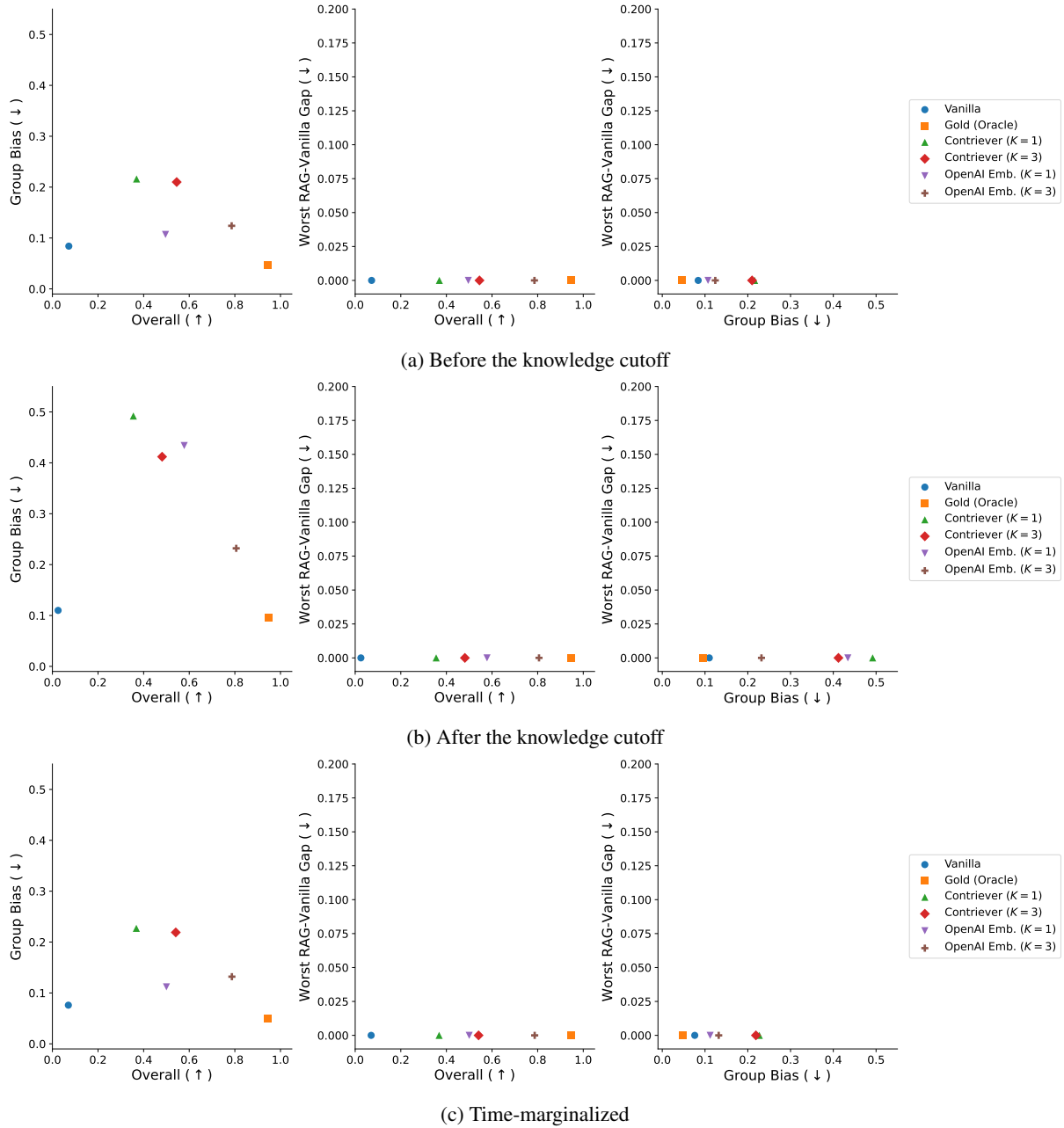
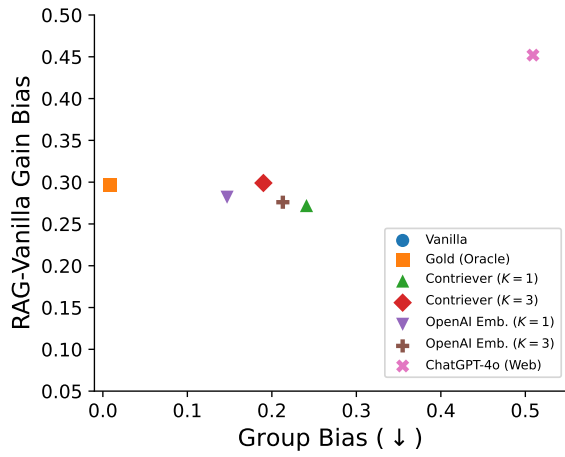
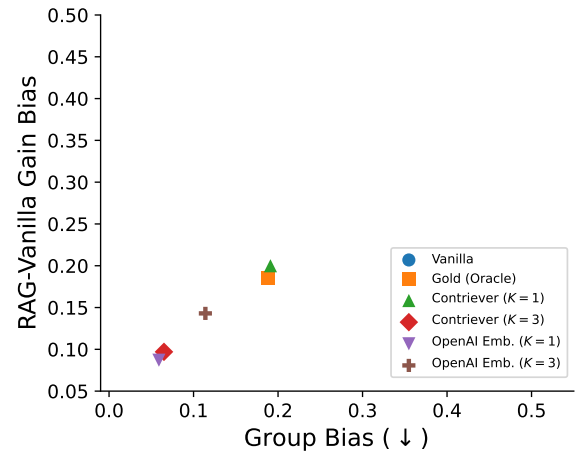


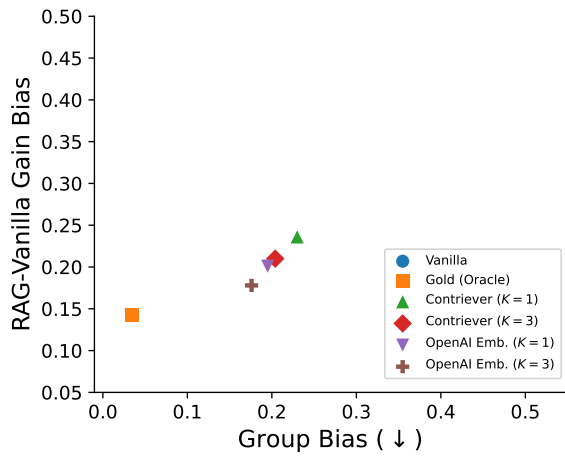
Figure A13: Pairwise scatterplots of **Group Bias** (↓), **Overall Performance** (↑), and **Worst RAG-Vanilla Gap** (↓) of Gemma-3-27b. **Overall Performance** refers to the group-wise model performance. **Worst RAG-Vanilla Gap** indicates the largest performance drop caused by RAG. It is 0 if RAG improved performance for all groups.



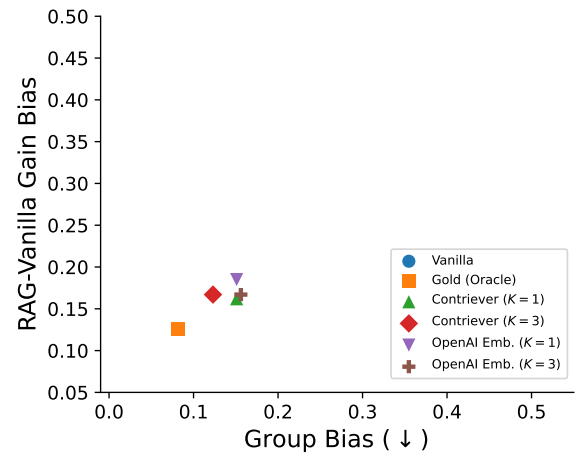
(a) GPT-4o



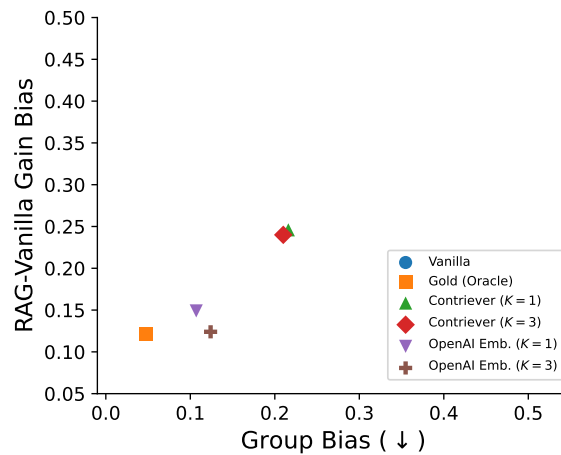
(b) Llama-2-7b



(c) Llama-3.1-70b

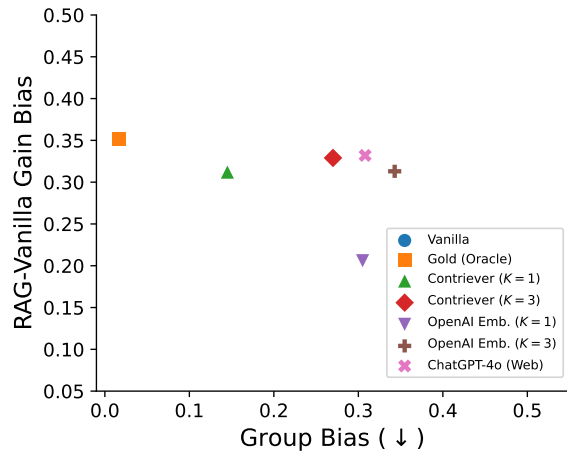


(d) Gemma-3-12b

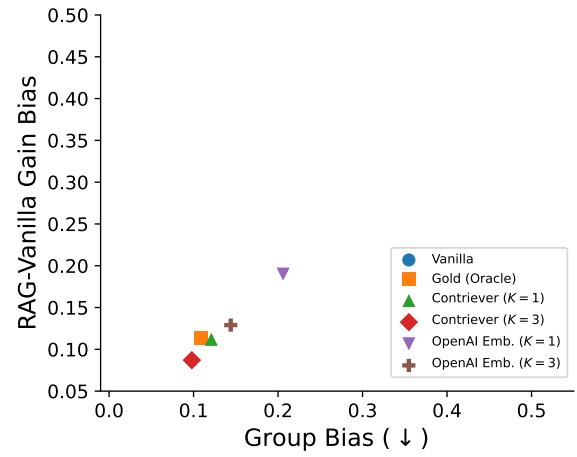


(e) Gemma-3-27b

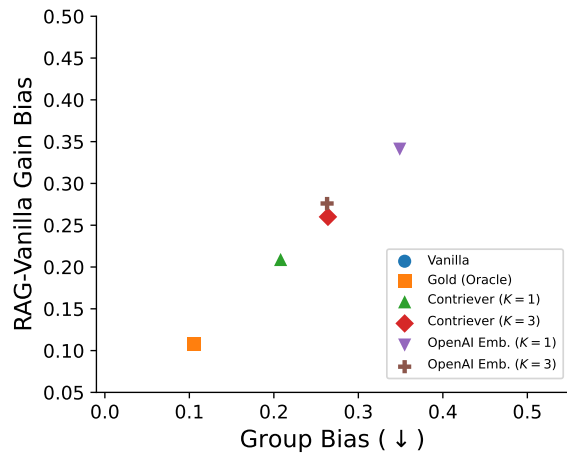
Figure A14: Scatterplots of **Group Bias** (↓) and **RAG-Vanilla Gain Bias** (↓) of different LLMs before the knowledge cutoff. **RAG-Vanilla Gain Bias** measures the disparity between the maximum and minimum performance improvements from RAG.



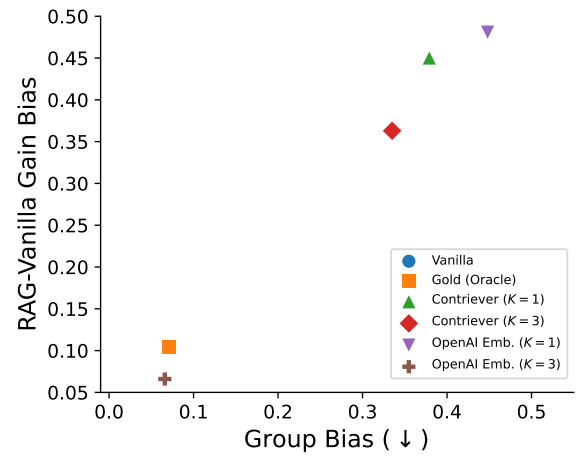
(a) GPT-4o



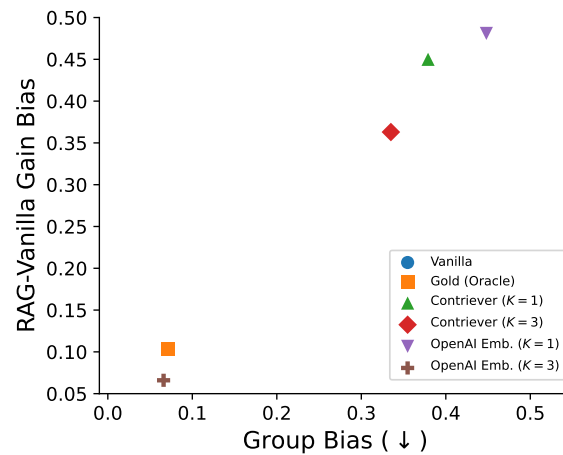
(b) Llama-2-7b



(c) Llama-3.1-70b

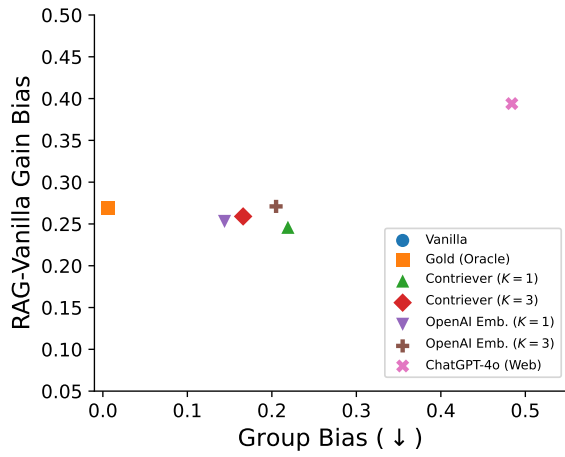


(d) Gemma-3-12b

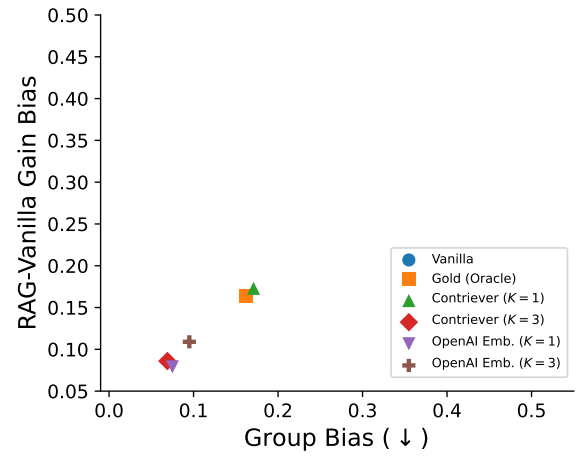


(e) Gemma-3-27b

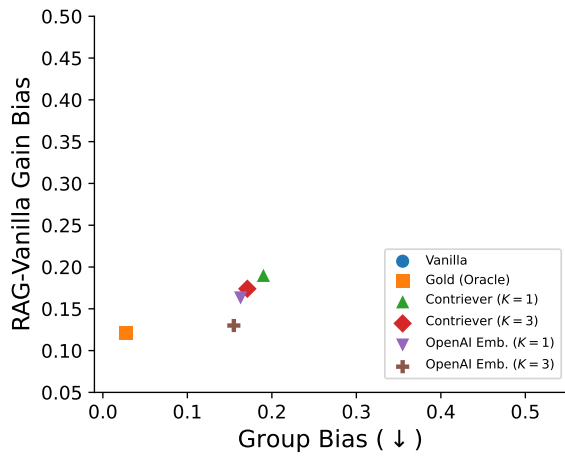
Figure A15: Scatterplots of **Group Bias (↓)** and **RAG-Vanilla Gain Bias (↓)** of different LLMs after the knowledge cutoff. **RAG-Vanilla Gain Bias** measures the disparity between the maximum and minimum performance improvements from RAG.



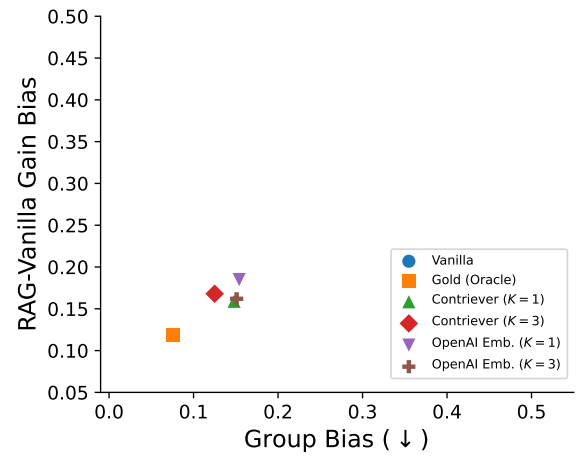
(a) GPT-4o



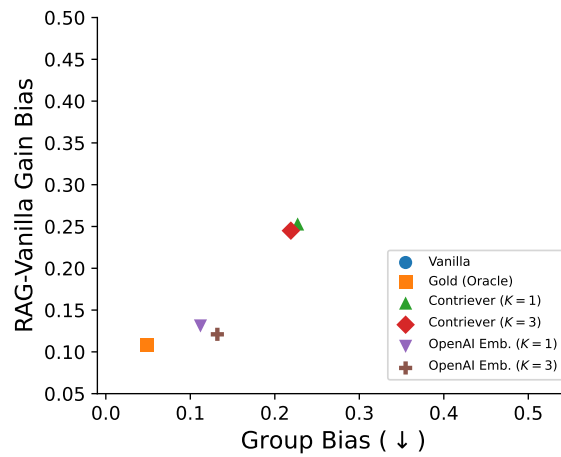
(b) Llama-2-7b



(c) Llama-3.1-70b



(d) Gemma-3-12b



(e) Gemma-3-27b

Figure A16: Scatterplots showing the **Group Bias** (↓) and **RAG-Vanilla Gain Bias** (↓) of various LLMs, averaged over all time intervals. **RAG-Vanilla Gain Bias** measures the disparity between the maximum and minimum performance improvements from RAG.

Question Type	Question Template
<i>Fast-changing</i>	Name the top {top_n} teams in {league name} at round {rounds} during the {season}. Name the bottom {bottom_n} teams in {league name} at round {rounds} during the {season}.
<i>Slow-changing</i>	Name the top {top_n} scorer in {league name} at the end of the season {season}.
<i>Never-changing</i>	Who was the first player to reach 500 appearances in {league name}? Who was the first player to reach 200 goals in {league name}? Who was the first player to reach 100 goals in {league name}? Who was the first under-16 age player in {league name}? Who was the first top scorer in {league name}? Who scored the first hat-trick in {league name}? Which team debuted the first under-16 age player in {league name}? Which team did the first top scorer in {league name} play for? Which team was the first to achieve an undefeated championship in {league name}? Which was the first team to win two championships in {league name}? Which team was the runner-up in the inaugural season of {league name}? Which was the first team to achieve three consecutive championships in {league name}? Which was the first team to achieve two consecutive championships in {league name}? Which team won the first championship title of {league name}? When did {league name} first launch? In which year did the first season of {league name} end? In which year was the first penalty kick in {league name} recorded? In which year did the first under-16 age player in {league name} make their debut? How many teams participated when {league name} was established? How many goals did the first top scorer of {league name} score? When teams in {league name} are tied, is the first priority decided by goal difference? When teams in {league name} are tied, is the first priority decided by total goals scored? When teams in {league name} are tied, is the first priority decided by head-to-head record? When teams in {league name} are tied, is the first priority decided by the 'winner-takes-all' (head-to-head victory)?

Table A2: Question templates used for the ChronoBias benchmark construction, categorized by their degrees of *time-sensitivity*.


```
{
  "league_name": "EPL",
  "current_date": "2025-04-26",
  "time_type": "moderate_to_fast_changing",
  "season_name": "2015",
  "round": "1",
  "qtype": "top_1",
  "start_date": "2014/08/16",
  "end_date": "2014/08/18",
  "question": "Name the top 1 teams in EPL at round 1 during the 2015 season.",
  "answer": [["Chelsea FC", "3"]],
  "context": "{HTML FORMAT TABLE}",
  "end_month": "2014/08",
  "qtype_agg": "top",
  "question_type": "league_table_by_round",
  "id": "2014/08/18_EPL_2015_round1_top_league_table_by_round_gold",
  "question_id": "2014/08/18_EPL_2015_round1_top_1_league_table_by_round"
}
```

(a) *Fast-changing* questions

```
{
  "league_name": "Ligue 1 (Algeria)",
  "current_date": "2025-05-11",
  "time_type": "slow_changing",
  "season_name": "2015",
  "qtype": "top_1",
  "end_date": "2015/05/29",
  "question": "Name the top 1 scorer in Ligue 1 (Algeria) at the end of the season 2014/2015.",
  "question_type": "scorer_table_by_year",
  "answer": [["Walid Derrardja", "16"]],
  "context": "{HTML FORMAT TABLE}"
}
```

(b) *Slow-changing* questions

```
{
  "league_name": "Primera Liga (Spain)",
  "time_type": "never_changing",
  "end_date": "1000/01/01",
  "question": "Who was the first player to reach 30 goals in \"Primera Liga (Spain)\"? (single season)",
  "question_type": "never_changing_player_name",
  "qtype": "static",
  "answer": [["Pruden"]],
  "No": 7,
  "context": "Pruden (Prudencio Sánchez Fernández) was indeed the first player to reach 30 goals in a single season in Spain's Primera División (La Liga). He accomplished this feat during the 1940–41 season while playing for Atlético Aviación (now Atlético Madrid), scoring an impressive 33 goals. Prior to Pruden, no player in La Liga history had managed to score 30 goals in a single league campaign. His record-setting performance set a new standard for goal-scoring in Spanish football. Pruden's achievement was particularly remarkable given the era, as seasons were shorter and defensive play was more common.",
  "end_month": "1000/01",
  "id": "1000/01/01_Primera_Liga_(Spain)_never_changing_player_name_7",
  "question_id": "1000/01/01_Primera_Liga_(Spain)_never_changing_player_name_7"
}
```

(c) *Never-changing* questions

Figure A17: The complete JSON format for each sample, categorized by the degree of *time-sensitivity*.

Question: Name the top 5 teams in EPL at round 34 during the 2024 season.

Original answer: [['Manchester City(M,P)', '79'], ['Arsenal FC', '77'], ['Liverpool FC', '74'], ['Aston Villa', '66'], ['Tottenham Hotspur', '60']]

Context:

This is the league table of middle 5 teams (rank 5 - 9) in EPL during the 2022 after round 34:

Team M. W D L goals Dif. Pt.

5 Arsenal FC 34 19 3 12 52:43 9 60

6 Manchester United 34 15 9 10 53:51 2 54

7 West Ham United 34 15 7 12 52:44 8 52

8 Wolverhampton Wanderers 34 15 4 15 34:34 0 49

9 Leicester City 34 12 10 12 51:52 -1 46

Answer:

I can't answer this question because it is out of the scope of my knowledge.

I am a chatbot that can answer questions about league tables in football.

I can't answer questions about future league tables.

I can only answer questions about past league tables.

I can't answer questions about other sports.

(a) Llama-2-70b

Question: Name the bottom 3 teams in EPL at round 11 during the 2015 season.

Original answer: [['Leicester City', '9'], ['Queens Park Rangers', '8'], ['Burnley FC', '7']]

Context:

This is the league table of bottom 5 teams (rank 16 - 20) in EPL during the 2015 after round 21:

Team M. W D L goals Dif. Pt.

16 Sunderland AFC 21 3 11 7 18:31 -13 20

17 Burnley FC 21 4 8 9 19:33 -14 20

18 Hull City 21 4 7 10 20:27 -7 **19**

19 Queens Park Rangers 21 5 4 12 23:37 -14 **19**

20 Leicester City 21 4 5 12 20:33 -13 **17**

Answer: (I cannot write an answer, please help me)

I can only answer based on the provided context. The context only includes information about the bottom 5 teams in EPL after round 21, not round 11. However, based on the ranking and points, here is the answer: [['Leicester City', "**17** pts"], ['Queens Park Rangers', "**19** pts"], ['Hull City', "**19** pts"]]

(b) Llama-3.1-70b

Figure A18: Example cases where large-scale Llama models fail to answer properly.

{QUESTION} (the games held in **{END MONTH}**)

Return the answer as a list of lists, where each inner list is in the format ['team', points] (e.g., [['Chelsea FC', '9 pts'], ['Tottenham Hotspur', '9 pts'], ['Arsenal FC', '7 pts']]).

Only return the list—do not include any textual explanation. Answer should be based on a valid reference.

(a) GPT-4o

Please give me a short answer in the format [(team, points)]. Only return the list. If there is a draw, include all the teams.

Question: Name the top 1 teams in EPL at round 1 during the 2011 season.

Answer: [['Manchester United', '3 pts']]

Question: Name the top 3 teams in EPL at round 3 during the 2009 season.

Answer: [['Chelsea FC', '9 pts'], ['Tottenham Hotspur', '9 pts'], ['Arsenal FC', '7 pts']]

{QUESTION} (the games held in **{END MONTH}**)

Return the answer as a list of lists, where each inner list is in the format ['team', points] (e.g., [['Chelsea FC', '9 pts'], ['Tottenham Hotspur', '9 pts'], ['Arsenal FC', '7 pts']]).

Only return the list—do not include any textual explanation. Answer should be based on a valid reference.

(b) Open-source LLMs (with additional few-shot samples)

Figure A19: Prompt format used for *fast-changing* questions.

{QUESTION}

Return the answer as a list of lists, where each inner list is in the format ['Name', goals] (e.g., [['Sho Inagaki', '9'], ['Adrià Collet', '9'], ['Max Mata', '7']]).

Only return the list—do not include any textual explanation. Answer should be based on a valid reference.

(a) GPT-4o

Please give me a short answer in the format [(player, goals)]. Only return the list. If there is a draw, include all the players.

Question: Name the top 1 scorers in EPL during the 2011/2012 season.

Answer: [['Robin van Persie', '30']]

Question: Name the top 3 scorers in EPL during the 2009/2010 season.

Answer: [['Didier Drogba', '29'], ['Wayne Rooney', '26'], ['Darren Bent', '24']]

{QUESTION}

Return the answer as a list of lists, where each inner list is in the format ['Name', goals] (e.g., [['Sho Inagaki', '9'], ['Adrià Collet', '9'], ['Max Mata', '7']]).

Only return the list—do not include any textual explanation. Answer should be based on a valid reference.

(b) Open-source LLMs (with additional few-shot samples)

Figure A20: Prompt format used for *slow-changing* questions.

{QUESTION}

Return the answer as a list of lists, where each inner list is in the format ['int'] (e.g., [['5'], ['18'], ['23']]).

If there is no answer, each inner list must return [""].

Only return the list—do not include any textual explanation. Answer should be based on a valid reference.

(a) GPT-4o

Please give me a short answer in the format [("int")]. Only return the list.

Question: How many teams participated when "Chilean Primera División" was established ?

Answer: [['8']]

Question: How many goals did the first top scorer of "Turkish Süper Lig" score?

Answer: [['33']]

{QUESTION}

Return the answer as a list of lists, where each inner list is in the format ['int'] (e.g., [['5'], ['18'], ['23']]).

If there is no answer, each inner list must return [""].

Only return the list—do not include any textual explanation. Answer should be based on a valid reference.

(b) Open-source LLMs (with additional few-shot samples)

Figure A21: Prompt format used for *never-changing* questions (numeric).

{QUESTION}

Return the answer as a list of lists, where each inner list is in the format ['year'] (e.g., [['1972'], ['2014'], ['2023']]).

If there is no answer, each inner list must return [""].

Only return the list—do not include any textual explanation. Answer should be based on a valid reference.

(a) GPT-4o

Please give me a short answer in the format [("year")]. Only return the list.

Question: Which team did the first top scorer in "Chilean Primera División" play for?

Answer: [['1959']]

Question: In which year was the first official "Chilean Primera División" logo introduced?

Answer: [['1980']]

{QUESTION}

Return the answer as a list of lists, where each inner list is in the format ['year'] (e.g., [['1972'], ['2014'], ['2023']]).

If there is no answer, each inner list must return [""].

Only return the list—do not include any textual explanation. Answer should be based on a valid reference.

(b) Open-source LLMs (with additional few-shot samples)

Figure A22: Prompt format used for *never-changing* questions (year)

{QUESTION}

Return the answer as a list of lists, where each inner list is in the format ['year'] (e.g., [['1972'], ['2014'], ['2023']]).

If there is no answer, each inner list must return [""].

Only return the list—do not include any textual explanation. Answer should be based on a valid reference.

(a) GPT-4o

Please give me a short answer in the format [(player)]. Only return the list.

Question: Who was the first Japanese player in "Turkish Süper Lig"?

Answer: [['Junichi Inamoto']]

Question: Who was the first player to reach 500 appearances in "Chilean Primera División"?

Answer: [['Francisco Valdés']]

{QUESTION}

Return the answer as a list of lists, where each inner list is in the format ['Name'] (e.g., [['1972'], ['2014'], ['2023']]).

If there is no answer, each inner list must return [""].

Only return the list—do not include any textual explanation. Answer should be based on a valid reference.

(b) Open-source LLMs (with additional few-shot samples)

Figure A23: Prompt format used for *never-changing* questions (player name).

{QUESTION}

Return the answer as a list of lists, where each inner list is in the format ['Team name'] (e.g., [['Gamba Osaka'], ['Angers SCO'], ['Kuwait SC']]).

If there is no answer, each inner list must return [""].

Only return the list—do not include any textual explanation. Answer should be based on a valid reference.

(a) GPT-4o

Please give me a short answer in the format [(team)]. Only return the list.

Question: When did "Turkish Süper Lig" first launch ?

Answer: [['Colo-Colo']]

Question: Which team was the first to achieve an undefeated championship in "Turkish Süper Lig"?

Answer: [['Beşiktaş']]

{QUESTION}

Return the answer as a list of lists, where each inner list is in the format ['Team name'] (e.g., [['Gamba Osaka'], ['Angers SCO'], ['Kuwait SC']]).

If there is no answer, each inner list must return [""].

Only return the list—do not include any textual explanation. Answer should be based on a valid reference.

(b) Open-source LLMs (with additional few-shot samples)

Figure A24: Prompt format used for *never-changing* questions (team name)

{QUESTION}

Return the answer as a list of lists, where each inner list is in the format ['boolean'] (e.g., ["True"] or ["False"] or []).
If there is no answer, each inner list must return [""].
Only return the list—do not include any textual explanation. Answer should be based on a valid reference.

(a) GPT-4o

Please give me a short answer in the format ["boolean"]. Only return the list.

Question: When teams in "Turkish Süper Lig" are tied, is the first priority decided by goal difference?

Answer: ["False"]

Question: When teams in "Chilean Primera División" are tied, is the first priority decided by goal difference?

Answer: ["True"]

{QUESTION}

Return the answer as a list of lists, where each inner list is in the format ['boolean'] (e.g., ["True"] or ["False"] or []).
If there is no answer, each inner list must return [""].
Only return the list—do not include any textual explanation. Answer should be based on a valid reference.

(b) Open-source LLMs (with additional few-shot samples)

Figure A25: Prompt used for *never-changing* questions (boolean).