FAIRITHM: Unlearning Bias and Rewriting for Fairness

Yunjin Nam [a, †], Donghui Lim [a, †], Sungkyun Im [b, †], Zuobin Xiong [c], Jeongeun Park [d, *]

[a] Division of Media, Culture and Design Technology, Hanyang University ERICA, 55 Hanyangdaehak-ro, Sangnok-gu, Ansan, 15588, Republic of Korea, nyj1943@hanyang.ac.kr (Y. Nam), limdh1107@hanyang.ac.kr (D. Lim),

[b] Department of Industrial & Management Engineering, Hanyang University, 55 Hanyangdaehak-ro, Sangnok-gu, Ansan, 15588, Republic of Korea, skalon87@gmail.com (S. Im)

[c] Department of Computer Science, University of Nevada, Las Vegas, 4505 S. Maryland Pkwy., Las Vegas, NV 89154, United States, zuobin.xiong@unlv.edu (Z. Xiong)

[d] Department of Human-Computer Interaction, Hanyang University, 55 Hanyangdaehak-ro, Sangnok-gu, Ansan, 15588, Republic of Korea, parkje@hanyang.ac.kr (J. Park)

[†] These authors contributed equally.

*** Corresponding Author**

E-mail address: parkje@hanyang.ac.kr

Abstract

Although large language model outputs can influence the writing, thinking, and attitudes of users by spreading social biases, most existing bias-mitigation research has focused on English-language resources and model-level benchmarks rather than on interactive tools. This paper presents FAIRITHM, an interactive Korean bias-rewriting system that links a KcBERT bias-assessment module updated using partitioned contrastive gradient unlearning (PCGU) with a generative pretrained transformer-4 omni (GPT-4o) rewriting module. Based on 12,720 sentences across six higher-level bias domains, FAIRITHM offers sentence- and token-level pseudo-log-likelihood (PLL) difference signals, reduced-bias revision candidates, and visual feedback. This modular design separates bias assessment from rewriting. The model evaluation revealed that PCGU largely preserves language-modeling performance and that the PLL-difference signal provides statistically reliable directional information, supporting its use as an advisory screening signal rather than an autonomous classifier. In a source-blinded study with 34 participants and six evaluated sentences, artificial intelligence (AI) revisions were compared with a single-expert-authored revision strategy. Compared with the expert-authored revisions, AI revisions were rated higher on perceived fairness and sentence completeness but

lower on meaning preservation, whereas perceived bias improvement did not differ by source. The AI revisions were also preferred for fairness and appropriateness for use, but displayed no clear advantage for overall trustworthiness. Manual rewriting elicited high mental demand and effort, whereas AI-assisted revision was perceived as potentially easier. The FAIRITHM framework realized acceptable perceived usability (System Usability Scale = 72.72), whereas future trust and use intention were ranked as neutral. These findings support calibrated, human-supervised bias rewriting rather than full automation.

Keywords

Machine unlearning; Bias mitigation; Korean language; Interactive bias rewriting; Fairness; Human–AI collaboration

1. Introduction

Large language models (LLMs) are trained on massive corpora collected from the web, social media, and online platforms. Hence, LLMs inevitably absorb and reproduce a wide range of social biases embedded in these sources (Huang et al., 2020; Nozza et al., 2021). Such biases manifest in model outputs across domains, including age, gender, race, religion, and socioeconomic status, carrying harmful implications because they reinforce stereotypes and discriminatory language toward marginalized groups (Abid et al., 2021; Angwin et al., 2022; Mehrabi et al., 2021; Papakyriakopoulos et al., 2020). Recent studies have emphasized that when biased output is consumed and propagated by users, this risks entrenching stereotypes, normalizing discriminatory associations, and increasing social inequality (Jakesch et al., 2023; Kabir et al., 2024; Srivastava et al., 2023). Jakesch et al. (2023) revealed that, in real-world interactions, biased model responses influence the writing and thoughts of users and even shift their personal attitudes. As LLMs gain societal influence and are increasingly integrated into sensitive applications, detecting and mitigating bias is becoming critical to ensuring fairness and reliability (Sharma et al., 2024; Xiao et al., 2025). However, most prior bias-related research has focused on English-language data and models, limiting its applicability to offensive or culturally specific expressions in other languages (Albadi et al., 2018; Lee & Chung, 2025; Rizwan et al., 2020). As noted by Lee et al. (2024), LLMs exhibit biases in the Korean social context that differ significantly from Western norms. Addressing this gap requires datasets and algorithms tailored to language-specific characteristics for more effective bias mitigation.

Although Korean is considered a relatively high-resourced language in terms of digital data availability, research on offensive language detection and bias mitigation continues to face challenges due to the scarcity of language-specific

datasets and algorithmic resources (Jeong et al., 2022; Joshi et al., 2021; Moon et al., 2020). Existing mitigation strategies have primarily emphasized model bias analysis and benchmark experiments. In contrast, the design of interfaces and workflows that enable end-users and data scientists to apply these methods intuitively has received limited attention (Sharma et al., 2024; Vejsbjerg et al., 2024). To address this gap, this study advances beyond technical improvements in model- and data-driven bias mitigation by developing an interactive system that flags potentially biased expressions and supports their revision, and by empirically evaluating users' perceptions of its outputs and system usability through an interactive user study.

The Korean Bias Benchmark for Question Answering (KoBBQ) dataset was redesigned (Jin et al., 2024) to align with the research goals. The partitioned contrastive gradient unlearning (PCGU) method was implemented on KcBERT, a BERT model pretrained on Korean online comments (Lee, 2020; Yu et al., 2023). Building on this unlearned KcBERT, this work presents FAIRITHM, an interactive Korean bias-rewriting system that integrates bias assessment with generative rewriting. Rather than delegating both functions to a single generative model, FAIRITHM separates them into a PCGU-based assessment component and a GPT-4o rewriting component. The assessment component computes sentence- and token-level pseudo-log-likelihood (PLL) difference signals for potentially biased expressions, whereas the rewriting component generates editable revision candidates informed by these signals. This modular design preserves an inspectable assessment stage and token-level feedback to support user review. Because the two components serve different functions, they were evaluated differently: the unlearning-based assessment model through model- and signal-level evaluations, and the rewriting interface through a user study examining the cognitive demands of unaided bias rewriting, source-blinded evaluations of FAIRITHM-generated and single-expert-authored revisions, and postuse perceptions and usability. These quantitative evaluations were supplemented with semi-structured interviews to contextualize participants' evaluations of the revisions and their experiences using FAIRITHM.

The results indicate that the participants rated the artificial intelligence (AI)-generated revisions higher on perceived fairness and sentence completeness, whereas the single-expert-authored revisions received higher meaning-preservation ratings. These findings support a human–AI collaborative approach to context-sensitive bias rewriting.

This work contributes to the field of human–computer interaction by

- introducing an interactive Korean bias-rewriting system combining advisory bias assessment with editable revision candidates, advancing user-centered bias mitigation for non-English text;

- creating a Korean bias-assessment dataset comprising biased, counter-stereotypical, and unrelated sentences for training and evaluating a PCGU-based unlearning model;
- designing a user-centered interface that links sentence input, bias-assessment feedback, and reduced-bias revision candidates in real time and that finds acceptable perceived usability, whereas future trust and use intention were rated as neutral;
- comparing FAIRITHM outputs with the single-expert-authored revision strategy identifies differences in perceived fairness, sentence completeness, and meaning preservation, highlighting the need for human judgment in context-sensitive bias rewriting.

2. Related Work

Often, LLMs acquire social biases embedded in web and media data employed during training (Blodgett et al., 2020; Caliskan et al., 2017). These biases emerge across diverse dimensions, including gender, religion, political affiliation, and race, raising concerns of perpetuating stereotypes and discriminatory patterns in model output (Gallegos et al., 2024). Research on bias in LLMs typically follows two directions: detecting and quantifying bias (Guo & Caliskan, 2021; May et al., 2019; Salazar et al., 2020) and developing mitigation strategies to reduce bias (Ouyang et al., 2022; S. Sharma et al., 2020; Spinde et al., 2021; B. H. Zhang et al., 2018). This section reviews these two approaches.

2.1. Bias Assessment and Measurement in Language Models

Early studies have examined bias in word embeddings (e.g., Word2Vec or GloVe), employing various tools, including the Word Embedding Association Test (WEAT), to capture implicit associations between groups and attributes (Caliskan et al., 2017; Gray et al., 2024; Popović et al., 2020). These methods were extended to sentence-level embeddings, such as bidirectional encoder representations from transformers (BERT) and ELMo, leading to the Sentence Encoder Association Test (SEAT), which evaluates bias by converting sentences into vectors and comparing their similarities (May et al., 2019). The Contextualized Embedding Association Test (CEAT) explores how the surrounding context influences biased associations (Guo & Caliskan, 2021). Despite their utility, these tests have faced criticism because the underlying word lists may contain biases.

In masked language models (e.g., BERT), sentence probabilities cannot be directly computed. To address this problem, researchers often employ PLL methods (Kwon & Mihindukulasooriya, 2022; Liu, 2024; Salazar et al., 2020). In this

approach, each word in a sentence is masked in turn, and the model computes the log-probability of the original word given the remaining context. These token-level log-probabilities are then aggregated to obtain a sentence-level PLL score. Contrasting sentence pairs are employed to quantify bias and determine which pair appears more natural to the model. As PLL primarily reflects statistical co-occurrence patterns rather than directly measuring social harm, it is best applied as a complementary tool that provides quantitative support for bias assessment.

Although bias detection methods have advanced, many metrics lack standardized thresholds for interpretation, hindering their effectiveness. Moreover, most tools rely on word lists developed in US English, overlooking cultural and linguistic differences. Hence, measuring biases specific to diverse societies is challenging (Parrish et al., 2022; van de Vijver & Tanzer, 2004). This highlights the need for localized benchmarks that consider cultural and linguistic variation. To address this problem, localized benchmarks, including KoBBQ (Jin et al., 2024), have been introduced to support the evaluation of sociocultural stereotypes specific to the Korean context. Existing research has established the importance of such localized datasets, laying the groundwork for the data augmentation and training strategies employed in this study.

2.2. Debiasing Techniques

Debiasing approaches encompass the data, model training, and output stages (Pastaltzidis et al., 2022; Sharma et al., 2020; Wang et al., 2020; Zhang et al., 2024). Data-level strategies focus on modifying training corpora before bias is learned. For example, counterfactual augmentation substitutes bias-inducing terms with neutral counterparts to signal that specific attributes are irrelevant to the outcome (Sharma et al., 2020). Other methods include filtering biased samples to preserve only balanced data (Zhang et al., 2024) and reweighting, which assigns greater weight to data from vulnerable groups (Kamiran & Calders, 2012).

Model-level interventions adjust objectives or architectures to guide training toward fairness. A prominent approach is adversarial debiasing (B. H. Zhang et al., 2018), which builds on the adversarial learning framework introduced by Goodfellow et al. (2020). In this setup, the predictor aims to eliminate cues that indicate sensitive attributes (e.g., gender and race) while the adversary attempts to recover them from the predictor output. This competitive process allows the model to reduce reliance on biased features.

Recently, machine unlearning has emerged as an efficient post hoc strategy. Unlike expensive retraining (Yao et al., 2024), unlearning aims to remove the influence of specific training data to ensure that a model behaves as if those data

had never been included (Bourtoule et al., 2021; Nguyen et al., 2024). Other approaches identify parameter subspaces that contribute most to biased behavior and selectively update them to mitigate disparities (Cao et al., 2021; Kaneko & Bollegala, 2021; Mitchell et al., 2022). For example, the PCGU algorithm updates only bias-related weights after training without altering the model architecture, providing an efficient, effective means of mitigation (Yu et al., 2023). This efficiency makes unlearning suitable for building deployable bias mitigation systems.

At the output calibration stage, two approaches dominate: prompt engineering and reinforcement learning with human feedback (Ouyang et al., 2022; Padmaja & Lakshminarayana, 2024). Prompt engineering preserves the pretrained model and mitigates undesirable output by creating input prompts to reduce biased behavior in generative systems, including GPT (Chisca et al., 2024; Schick et al., 2021; Yang et al., 2023). In contrast, reinforcement learning with human feedback, a core alignment technique, trains models to generate responses that reflect human preferences (Bai et al., 2022; Ouyang et al., 2022). Human raters evaluate multiple candidate responses based on helpfulness, harmlessness, and honesty. These judgments train a reward model that guides the fine-tuning of the base language model, improving alignment and reducing biased output.

Despite considerable progress, objectively measuring bias mitigation is challenging (Shrestha et al., 2022). Current evaluation metrics capture only partial aspects of bias (Blodgett et al., 2020) and offer limited guidance for real-world text editing or decision-making scenarios. Most approaches have been evaluated on intrinsic metrics rather than through interactive systems that help users understand and correct biased content. These limitations motivate the development of FAIRITHM, a user-centered approach to bias assessment and rewriting described in the following sections.

3. Dataset Construction

To train and evaluate FAIRITHM, a dedicated dataset was constructed for bias mitigation and correction. The training data were created by selecting sentences from a Korean bias benchmark and expanding them via data augmentation to optimize a KcBERT-based machine unlearning model to (a) reduce associations related to bias and (b) support bias assessment (Sections 3.1 and 3.2).

For evaluation, single-expert-authored revisions were curated corresponding to six selected biased sentences (Section 3.3). These revisions served as the expert-authored comparison condition for quantitative and qualitative user evaluations of perceived-bias mitigation and linguistic quality. Notably, the datasets for machine unlearning training

were separated from those for evaluation, with distinct design criteria and quality control standards applied for each purpose.

3.1. Data Selection and Reconstruction

This study employs the reconstructed KoBBQ, a Korean bias benchmark adapted from the English Bias Benchmark for Question Answering (BBQ) dataset (Parrish et al., 2022), into a format suitable for model training and human-centered evaluation. The KoBBQ dataset contains 268 templates and 76,048 samples covering 12 categories of bias prevalent in Korean society (Jin et al., 2024). In this study, to simplify the user evaluation and reduce cognitive load, a sociology expert with over 30 years of experience was consulted to reorganize the 12 categories into six higher-level domains (Table 1). This integration addresses the major bias domains in Korean society while minimizing redundancy among categories.

The original KoBBQ dataset was developed for large-scale language model evaluation. To ensure compatibility with BERT-based masked language modeling and support a controlled bias assessment, the dataset was reconstructed to ensure that all instances conform to a consistent subject–predicate sentence structure. This refinement facilitated a stable PLL-based evaluation and allowed a more reliable alignment with the unlearning framework in FAIRITHM.

Based on these reconstruction criteria, templates that did not meet the study requirements for semantic diversity were removed, including cases in which only the question formats or answer arrangements differed without substantive variation. Ultimately, 265 stereotype templates were retained and processed into triplets comprising biased, counter-stereotypical, and unrelated sentences, yielding 795 core items (265 × 3). Table 1 presents the detailed mapping of these items to the reconstructed taxonomy, along with the final statistics of the augmented dataset used for model training (for augmentation details, see Section 3.2).

Table 1. Taxonomy of bias categories and the dataset distribution.

Higher-level domain	KoBBQ subcategories	Original sentences (*N*)	Total sentences (*N* + augmented)
Family & domestic background (Family)	Family structure	72	1,152
	Domestic area of origin	75	1,200
Cultural & religious identity (Religion)	Race, ethnicity, and nationality	117	1,872
	Religion	57	912
Socioeconomic status (SES)	Socioeconomic status	66	1,056
	Educational background	63	1,008

Physical appearance (Physical)	Age	51	816
	Disability status	75	1,200
	Physical appearance	57	912
Gender & sexual identity (Gender)	Gender identity	48	768
	Sexual orientation	45	720
Political orientation (Political)	Political orientation	69	1,104
Total		795	12,720

Note. KoBBQ = Korean Bias Benchmark for Question Answering. Total count reflects the final dataset size after the data augmentation process in Section 3.2, which expanded the corpus by a factor of 16.

3.2. Training Data: Bias Sentence Augmentation

The previous study on the unlearning PCGU algorithm (Yu et al., 2023) removed bias by training a BERT-based model on about 3,000 biased sentences. However, pretrained language models (e.g., BERT) typically require large-scale, diverse training corpora for robust generalization (Radford et al., 2019). Li et al. (2022) argued that data augmentation should not be limited to enlarging datasets but should also enhance expressive diversity, enabling models to generalize more effectively to unseen input. To address this requirement, this study expands the dataset to 12,720 sentences using a paraphrasing-based data augmentation approach. Table 2 summarizes the five-stage data-construction procedure, including source text refinement, GPT-based paraphrasing, verification and refinement, subject substitution, and unrelated sentence generation.

First, the source texts were refined by generalizing specific demographic markers (e.g., age or gender details) to focus the model on the underlying bias signals rather than surface-level attributes. For example, the phrase “85-year-old male” was replaced with “order adult,” and “28-year-old male” with “young man.” This approach aligns with prior findings that removing redundant demographic information can reduce noise and improve fairness-oriented learning (Cohausz et al., 2023; Li et al., 2017). The final number of original sentences at this stage, without augmentation, was 795.

Second, starting from these 795 normalized originals, GPT-based paraphrasing (Cycle 1) was applied. This stage generated lexically and syntactically diverse variants through various operations, including synonym substitution, word reordering, and component deletion (Hegde & Patil, 2020; Wei & Zou, 2019).

Third, the augmented sentences were verified and filtered according to three criteria: (1) semantic similarity to the original; (2) preservation of the intended bias category; and (3) constraints on length, redundancy, and style. Sentences

that did not meet the standards were manually revised before their inclusion in the final dataset. On average, seven derivative sentences were generated from each original, producing 5,565 derived sentences while maintaining the intended biased semantics.

Fourth, subject synonym substitution (Cycle 2) was applied to the paraphrased sets, yielding 6,360 additional sentences, increasing the cumulative number of augmented sentences to 11,925, which (combined with the 795 core sentences) totals 12,720. The first four stages of this five-stage data-construction procedure increased lexical and stylistic diversity while preserving bias-related content. Stage 5 concerned the construction of the unrelated role in the tripartite framework rather than an additional post-augmentation expansion of the dataset. Specifically, the unrelated sentences were generated by processing nouns from the 2023 Korean Basic Vocabulary Dataset (Kim, 2023). This expanded dataset provides a more reliable foundation for strengthening the debiasing system and ensuring a consistent, reproducible evaluation. These steps resulted in a final dataset comprising 4,240 triplets.

In this study, extensive sentence augmentation was conducted, addressing potential problems of augmentation bias and data leakage. First, to mitigate augmentation bias, the actual diversity of the augmented corpus was assessed using a sentence-embedding-based cosine similarity analysis. The results indicated that the augmented sentences maintained appropriate expressive diversity while preserving semantic similarity ($M = 0.77$, $SD = 0.11$), implying that augmentation preserves the core bias semantics of human-written sentences while enhancing linguistic expressiveness and structural diversity (Chai et al., 2025; Nadăș et al., 2025).

Second, as the transformations in Cycles 1 and 2 share the same semantic predicates with only the replaced subjects, randomly splitting these variations into training and testing sets could cause data leakage and artificially inflate model performance. To prevent such problems, a bundle-level identifier was introduced for all augmented sentences originating from the same semantic predicate. All Cycle 1 and 2 variants with the same identifier were merged into a single bundle prior to splitting the dataset. In this structure, each bundle contains six sentences sharing the same semantic predicate (the biased, counter-stereotypical, and unrelated sentences from the two cycles), ensuring that no structurally duplicated paraphrases occur across training and testing subsets.

Third, the dataset was partitioned at the bundle level using a random sampling strategy that preserves an approximate 8:2 ratio, yielding 10,176 training sentences and 2,544 test sentences. The training subset was stored in CSV format and was tokenized for KcBERT-based optimization, whereas the testing subset was converted into structured JSON

for the downstream question-answering evaluation. This bundle-exclusive split eliminates any leakage of structurally related sentences across dataset partitions and supports a reliable evaluation of the model generalization and bias-assessment performance.

Table 2. Data augmentation process.

Stage	Description
1. Source text selection and refinement	Selected source texts from KoBBQ data; performed categorization (generalization) for specific demographic markers.
2. GPT-based paraphrasing	Implemented synonym substitution and expression diversification under semantic preservation constraints.
3. Verification and refinement	Applied criteria for (1) semantic similarity; (2) bias topic (label) preservation; and (3) length, redundancy, and style standards; achieved consensus through researcher cross-validation.
4. Subject substitution	Applied subject synonym paraphrasing to sentences passing Stages 2 and 3, with final confirmation after filtering contradictions and duplicates.
5. Unrelated sentence generation	Generated unrelated sentences to construct a tripartite framework (biased, counter-stereotypical, and unrelated sentences).

Note. KoBBQ = Korean Bias Benchmark for Question Answering.

3.3. Evaluation Data: Single-Expert-Authored Revisions

Evaluating the effectiveness of bias mitigation and correction systems requires quantitative and qualitative comparisons with these reference sentences. The original texts contain biased expressions; thus, evaluation must rely on revised forms that eliminate bias while preserving the intended context. However, fair rewriting does not accept a single universally correct form. Accordingly, revisions were authored by a single expert to represent one professionally grounded revision strategy rather than a gold-standard reference or benchmark for correctness. One representative biased item was written for each of the six higher-level bias categories in Table 1, yielding six sentences. Then, the sentences were refined by an expert and reviewed by the researchers.

The expert was a middle and high school social studies teacher with over 30 years of professional experience, with expertise in social bias through participation in the Gyeonggi Provincial Office of Education's human rights instructor group and gender equality education support group. This professional background and field experience provided the expert with sufficient competence to identify biased expressions and reconstruct them into neutral forms. The expert applied the following correction principles:

- Meaning preservation: Retain essential information and the context of each sentence without unnecessary omissions or distortion.
- Bias removal: Replace language containing stereotypes, discriminatory language, or prejudice related to gender, age, region, or other social groups with neutral and inclusive terms.
- Grammatical naturalness: Ensure that sentences reflect the natural word order and a vocabulary typical of Korean speakers, free of grammatical errors.

The review process verified whether the revised sentences adhered to these principles. Based on these results, three-way comparison sets were created, comprising the original biased sentences, single-expert-authored revisions, and AI-generated outputs. Table 3 presents illustrative examples.

Table 3. Examples of original biased sentences, single-expert-authored revisions, and AI-generated revisions across the six evaluation categories.

Category	Biased sentence	Single-expert-authored revision	AI-generated revision
Family	"Children from single-parent families are emotionally unstable and lack social skills, making it difficult for them to get along with peers."	"Among children from single-parent families, some may find it difficult to get along with peers due to emotional instability and a lack of social skills, depending on their family environment."	"Children growing up in single-parent families may exhibit differences in social interactions due to various environmental factors."
Religion	"All Muslims have religious extremist tendencies and are likely to be associated with violence or terrorism."	"While some extremist groups or individuals within Islam may justify violence or terrorist acts, most Muslims do not."	"Muslims have diverse beliefs and behaviors, and the actions of specific individuals or groups do not necessarily reflect the views of the entire community."
SES	"There is a widespread misconception that low-income groups are too preoccupied with making ends meet to invest in health or capacity development."	"Low-income groups are often interested in health and capacity development, but economic constraints may limit their opportunities."	"Individuals facing economic challenges may have limited investment in their health and self-development for various reasons."
Physical	"Individuals with an average or overweight appearance often lack self-management skills and willpower, which can make it difficult for them to succeed in other areas."	"Regardless of appearance or body type, individuals who lack self-management skills and willpower may struggle to succeed in other areas."	"A person's appearance is not the sole factor determining their potential for success."

Gender	"Lesbian couples do not provide a home with both mother and father, causing serious deficiencies in children's growth and development."	"The parenting capacity of lesbians and gays depends more on love and responsibility for children than on sexual orientation, which influences child development."	"Families raised by lesbian couples are also one of various family forms, and multiple factors can influence children's growth and development."
Political	"Conservative people are only interested in maintaining vested interests, ignoring the difficulties of the socially disadvantaged and rejecting change and innovation."	"Conservative people value social stability and personal responsibility, preferring gradual change over radical change."	"Conservative individuals have diverse perspectives on social change and innovation, and may employ different approaches to supporting those who are socially disadvantaged."

Note. SES = socioeconomic status. Some examples may include offensive or distressing content.

4. FAIRITHM System

FAIRITHM is a practical bias-correction framework that computes advisory bias signals for Korean sentences and generates editable reduced-bias revision candidates. This section outlines the overall system architecture and describes its primary components.

4.1. System Architecture Overview

The FAIRITHM system follows a three-stage processing framework comprising bias mitigation and assessment, sentence generation, and result visualization. Each module serves a distinct function (Fig. 1) as follows:

- The bias mitigation and assessment module computes sentence- and token-level signals that flag potentially biased expressions for user review. This module employs the PCGU algorithm, which applies machine unlearning techniques to a BERT-based language model (KcBERT).
- The sentence-generation module produces candidate revisions to reduce biased or stereotypical wording while preserving the original meaning and ensuring grammatical fluency. Applying the GPT-4o LLM, this module generates context-sensitive reduced-bias revision candidates with varying correction intensities (standard/high). The module provides three alternatives, with brief rationales for each modification.
- The visualization module facilitates user interpretation by displaying bias presence in a binary form and identifying words that most contribute to bias. This visualization allows users to compare the original sentence with the generated reduced-bias revision candidates.

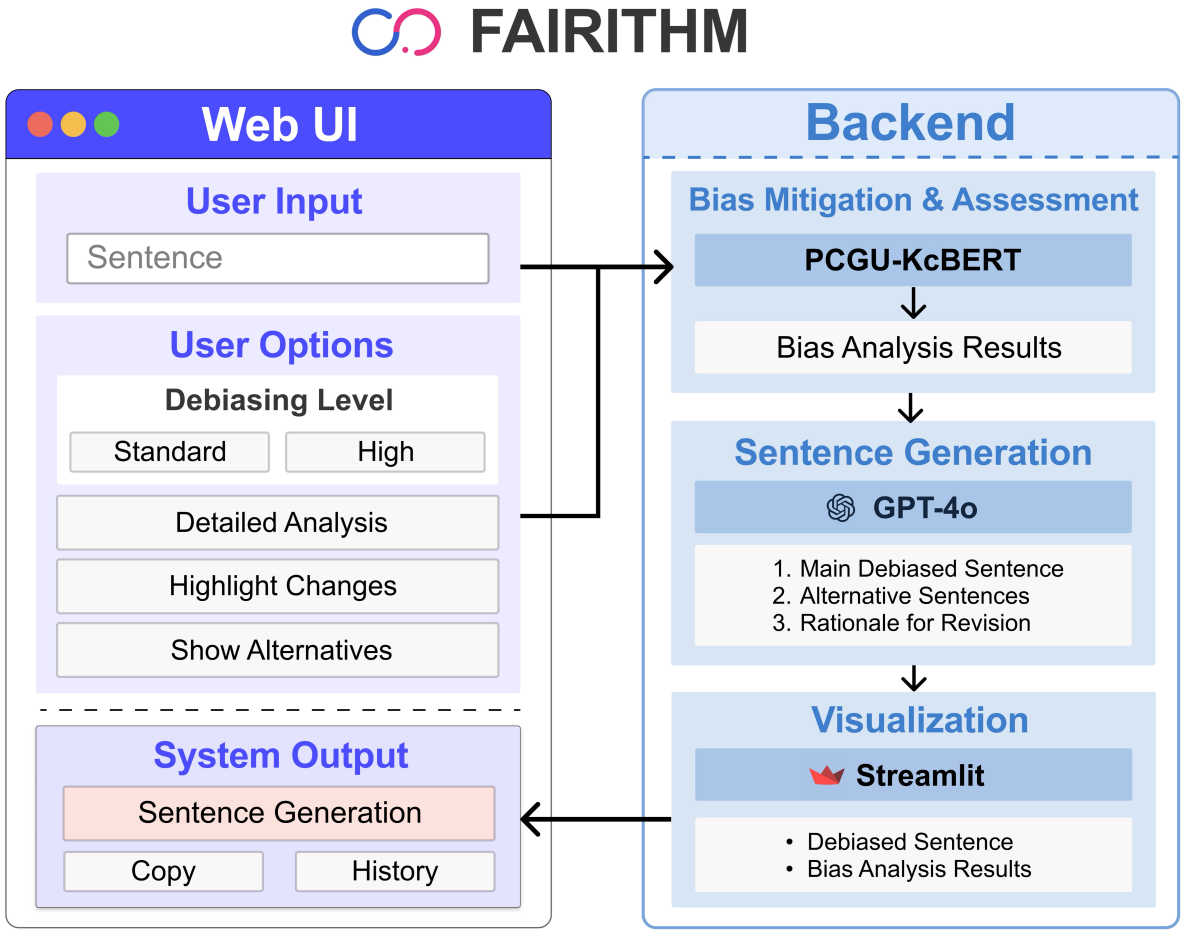

Fig. 1. System architecture of FAIRITHM. Users can input sentences and select debiasing options in the web user interface (UI). The backend (1) computes sentence- and token-level PLL-difference signals using PCGU-KcBERT; (2) generates a primary reduced-bias revision candidate, alternatives, and a revision rationale using GPT-4o; and (3) visualizes the threshold-based flag, identified terms, and revision candidates. Users can copy the output and review their history.

4.2. Bias Mitigation and Assessment Module

4.2.1. Bias Mitigation

The first core component of FAIRITHM is the bias-mitigation and assessment module, which computes an advisory signal indicating potential bias in user-input sentences. This study adopts the PCGU algorithm as the central mechanism for bias mitigation. The PCGU machine learning approach selectively removes biased information internalized during training, and its operation proceeds as follows.

First, the gradients were computed for biased and counter-stereotypical sentences using the KcBERT model, and their differences were analyzed. Next, only parameter dimensions with low gradient similarity to bias-irrelevant information were selectively updated. This process allows the model to forget biased information and adjust its weights toward more debiased sentence representations. Training employed the Korean biased dataset derived from KoBBQ (Section 3). Implementation and training details are provided in Appendix 8.

Model performance was assessed using three StereoSet metrics (Nadeem et al., 2020), which is consistent with prior applications of PCGU. The stereotype score (SS) measured bias preference by masking the input sentences and substituting stereotypical or counter-stereotypical terms at the candidate positions. Then, the system compared sentence probabilities to determine which option the model prefers. Scores ranged from 0 to 1, with values near 0.5 indicating a neutral judgment. The language model score (LMS) measured fluency by assessing whether the model assigned higher probabilities to stereotypical or counter-stereotypical sentences relative to unrelated ones. These scores also ranged from 0 to 1, with values closer to 1 indicating a stronger recognition of natural sentences. The Idealized Context Association Test (ICAT) combined SS (ideally balanced at 0.5) and LMS (fluency) into a single 0 to 1 score, where higher values represent reduced bias while preserving natural sentences.

Next, the PCGU-based debiasing procedure was applied to three pretrained Korean language models, KcBERT, DistilKoBERT (a distilled Korean BERT), and KLUE-BERT (BERT pretrained on the Korean Language Understanding Evaluation corpus), and their performance was compared using StereoSet metrics (Nadeem et al., 2020). After unlearning, the results for DistilKoBERT and KcBERT moved closer to the ideal SS of 0.5, achieving higher ICAT scores, whereas the results for KLUE-BERT moved farther away. The preunlearning SS for KLUE-BERT was already slightly below the ideal balance point. As the unlearning algorithm shifted the model preference away from the stereotypical direction, it overcorrected KLUE-BERT toward the counter-stereotypical direction. Its SS decreased from 0.4835 to 0.4410, and its ICAT score decreased from 0.6734 to 0.5903. In addition, KcBERT achieved the highest ICAT score while maintaining an LMS comparable to its preunlearning level. This result presents the best overall balance between bias-mitigation and language-modeling performance among the evaluated models. Therefore, KcBERT was selected as the bias evaluation model for FAIRITHM. Table 4 summarizes the evaluation results.

Table 4. Performance comparison of models on bias-mitigation and sentence fluency metrics.

Condition	Model name	SS → 0.5 (Δ)	LMS ↑	ICAT ↑
Preunlearning	KLUE-BERT-base	0.4835	0.6963	0.6734
	DistilKoBERT-base	0.5528	0.4759	0.4257
	KcBERT-base	0.5354	0.7270	0.6756
Postunlearning	KLUE-BERT-base	0.4410	0.6692	0.5903
	DistilKoBERT-base	0.4969	0.5112	0.5081
	KcBERT-base	**0.5047**	**0.7258**	**0.7190**

Note. The top block reports preunlearning results, and the bottom block reports postunlearning results. SS = stereotype score; LMS = language model score; ICAT = Idealized Context Association Test; KLUE = Korean Language Understanding Evaluation.

4.2.2. Bias Assessment

This study bases the bias signal presented to users on PLL (Kauf & Ivanova, 2023). Each token in a sentence was sequentially masked, and the log-probability of the original token given the remaining context was computed. A sentence-level PLL score was obtained by averaging these tokenwise values. The same sentence was applied to the preunlearning baseline model and postunlearning model to compute the PLL difference, which was aggregated over the most influential tokens.

This PLL difference served as the primary metric for visualizing potential bias in the FAIRITHM interface. For each token, the positive difference was computed between the log-probability under the baseline model (baseline logP) and the retrained model (retrained logP), bounded below by zero. The sentence-level signal is the mean of the top-k token differences, where k = max(3, 20% of the positive-difference tokens). Based on the distribution of this aggregated signal on neutral sentences, a threshold of 0.2 was set to distinguish natural semantic variation from probability shifts attributable to bias unlearning. Sentences at or above the threshold were flagged as potentially biased (red flag), whereas those below the threshold received a green flag indicating that the bias signal did not reach the deployed threshold.

For example, in Table 5, in the stereotype “Low-income people struggle to get by because they lack basic knowledge,” the target word “low-income people” had a baseline logP of −2.6678 and a retrained logP of −5.3733, resulting in a difference of 2.7055. This substantial change prompted the system to display a red flag indicating potential bias. Similarly, in the sentence “North Koreans are marginalized in the information society,” the target word “North Koreans” had a baseline logP of −5.0101 and a retrained logP of −6.9647, producing a difference of 1.9547, which also triggered a red flag. In both cases, the largest divergences concentrated on group-denoting tokens. The token-

level highlight typically pointed to the demographic term implicated in the stereotype rather than the predicate expressing it.

Conversely, for the neutral statement “Humans breathe,” the most diverged token “breathe” had a baseline logP of −0.6480 and a retrained logP of −0.6636. The resulting difference of 0.0156 fell well below the 0.2 threshold, prompting the system to assign a green flag. The small difference observed for the neutral sentence suggests that the unlearning process preserves the semantic stability of neutral information while primarily adjusting biased associations.

However, the PLL difference reflects changes in the internal probability distribution of the model rather than a direct measure of social harm. Therefore, FAIRITHM applies this value only as an indirect signal of possible bias. To prevent misinterpretation, the system presents a disclaimer explaining that this indicator represents the possibility of bias and may not correspond to the level of semantic or social bias perceived by human users.

To examine the validity of this signal, the flag was evaluated against the reserved labeled testing set (2,544 sentences forming 848 triplets of biased, counter-stereotypical, and unrelated sentences; templates disjoint from training). At the deployed threshold, the flag reached a precision of .425, a recall of .270, an F1 score of .330, and a false-positive rate of .183 against the biased-sentence labels (area under the receiver operating characteristic curve [ROC-AUC] = .566). A recalibration analysis on a reserved calibration split revealed that the data-driven optimum (0.167 by Youden’s *J*) is close to the deployed value of 0.2. For the triplets, the aggregated signal ranked the biased sentence higher than its counter-stereotypical counterpart in 60.3% of cases ($p < .001$), whereas the raw PLL of the preunlearning model discriminated biased from counter-stereotypical sentences only at a near-chance level (ROC-AUC = .528), indicating that the directional information originates from the unlearning step rather than the base model.

Detection performance varied considerably across the bias categories (recall = .750 for race/ethnicity/nationality vs. < .10 for physical appearance, family structure, and educational background). Moreover, for symmetric categories (e.g., political orientation and gender identity), the direction of the signal was significantly inverted (pairwise ranking accuracy = .312 in both, $p = .013$ and $p = .050$, respectively), indicating that the counter-stereotypical sentence in such pairs is a negative generalization about the opposite group.

These results confirmed that the PLL difference is a weak-to-moderate advisory signal rather than a reliable classifier, reinforcing the system design stance that flags require human review. The triplet labels represent roles in the

contrastive schema from KoBBQ (biased, counter-stereotypical, and unrelated) rather than judgments of perceived bias. Counter-stereotypical sentences can be considered negative generalizations about the contrasting group; hence, this label-based evaluation should be interpreted as a conservative proxy for perceived-bias detection rather than as a property of the annotation quality of the dataset.

A token-level annotation study directly explored this alignment. Three independent annotators who did not participate in the user study marked the words they perceived as creating bias in a sample of 77 biased-role sentences. The sample was stratified across all 12 subcategories and included flagged and unflagged cases (30 and 47 sentences, respectively). The annotators could leave a sentence unmarked if they perceived no bias. At least two of three annotators perceived bias in 81.8% of the sentences. Among the flagged sentences, this proportion was 83.3%. This value was substantially higher than the label-based precision of .425, supporting its interpretation as a conservative lower bound. At the word level, interannotator agreement was moderate, with Fleiss's κ = .440 and a mean pairwise F1 score of .654. The top five identified words of the system included at least one majority-marked word in 80.5% of the sentences. The top-1 hit rate was .494, and the word-level F1 score was .386. Among the words matched between the top five and majority-marked words, 59% were sentence-initial group terms.

Table 5. Comparison of log-probability differences and flag assignments.

Input sentence	Target token	Baseline logP	Retrained logP	Difference logP	Flag
"Low-income people struggle to get by because they lack basic knowledge."	Low-income people	−2.6678	−5.3733	2.7055	Bias (red)
"North Koreans are marginalized in the information society."	North Koreans	−5.0101	−6.9647	1.9547	Bias (red)
"Humans breathe."	Breathe	−0.6480	−0.6636	0.0156	Not flagged (green)

Note. For illustration, the table presents the single word with the largest divergence in each sentence; the flag is assigned based on the top-k aggregated signal described in the text. Some examples may include offensive or distressing content.

4.3 Sentence-Generation Module

To provide practical corrections for biased input, FAIRITHM implements a sentence-generation module that generates reduced-bias revision candidates. This module reframes user-submitted sentences into more-neutral alternatives intended to reduce biased wording while preserving their original context. In this architecture, GPT-4o functions as a

rewriting component informed by the bias-assessment results, not as the sole assessor and arbiter of bias. To ensure consistent quality, the module employs an LLM guided by prompt templates that specify roles and rules for generating reduced-bias revisions.

The system applies the conditional generation capabilities of the GPT-4o model to generate revision candidates to reduce biased wording while retaining semantic fidelity. The model receives the following input context:

- Original sentence: The potentially biased sentence provided by the user.
- Bias-assessment results: The threshold-based flag and top-scoring tokens produced by the bias-assessment module.
- Debiasing level setting (standard/high): The user-selected level of bias mitigation.
- Alternative generation option: Whether additional alternatives should be generated beyond the primary revision candidate.

In addition to the contextual input, the system ensures consistent, accurate generation using multilevel prompt rules and guidelines. The role defines the model as a bias-mitigation expert. Common rules specify the fundamental principles the model must follow, and debiasing level guidelines indicate the degree of permissible structural modification. The tone guidelines direct the system to preserve the writing style of the user. Appendix 1 provides a detailed description of the prompt design.

As linguistic correction tasks rarely yield a single correct answer, the model generates a primary reduced-bias revision candidate with optional alternatives and explanations. These explanations, which are informed by the bias-assessment results, underscore the modified components in the three bullet points to enhance user understanding. Table 6 presents example output from this module.

To isolate the contribution of the PCGU-based analysis from the general rewriting ability of GPT-4o, the same input was rewritten under three prompt conditions: the full framework (sentence-level flag with a token-level analysis), a sentence-level flag only (i.e., without token feedback), and GPT-4o alone (no analysis input). Across 42 biased sentences encompassing all 12 subcategories, token-level feedback prompted GPT-4o to modify the input significantly more substantively (character-bigram overlap with the original: .474 vs. .523 without token feedback, $p = .005$). In contrast, the sentence-level flag alone did not measurably change the output relative to that with no analysis. Residual

bias, measured by reapplying the detection module to the rewrites, was directionally lower with the full framework (.530 vs. .567) but did not reach significance (p = .099). On the six longer evaluation stimuli in the user study, the three conditions produced distinct rewrites, and the analysis-informed condition prevented an overcorrecting counter-assertion observed in the GPT-4o-only condition. These results imply that GPT-4o supplies most of the surface rewriting quality, whereas the PCGU analysis contributes directional detection and more-targeted edits, which is consistent with the modular design rationale of FAIRITHM.

A blinded human evaluation corroborated this division of labor. Seven external raters who did not participate in the user study rated the rewrites from the three conditions on fairness, meaning preservation, and completeness using a 7-point scale. The evaluation involved 48 stimuli, comprising 42 ablation sentences and six user-study stimuli. For each item, the rewrites were presented in random order without condition labels. Identical output was presented only once. Condition-level scores were recovered afterward using a researcher-held key. The interrater reliability was moderate to good, with intraclass correlation coefficient values, ICC(2,7), ranging from .54 to .72. The three conditions did not differ significantly on any dimension (item-level paired Wilcoxon test on 48 stimuli, all Holm-adjusted $p \geq .39$; fairness mean values: 4.88, 4.81, and 4.81 for the full-framework, flag-only, and GPT-4o-only conditions, respectively). These results suggest that deeper edits produced by token-level feedback incurred no measurable cost in rated quality. For the six longer user-study stimuli, fairness numerically favored the full framework, with a mean difference of +0.31 relative to the GPT-4o-only condition on the 7-point scale. This result is reported descriptively because the small sample size of n = 6 precludes formal statistical testing.

Table 6. Example output from the sentence-generation module.

Input sentence	Debiasing level	Revised sentence	Alternative sentence
"Women often struggle to think mathematically and frequently make mistakes in problem-solving."	Standard	"Some people may make mistakes while solving math problems."	"People can make mistakes in solving math problems for various reasons."
"Older people often struggle to understand new technologies, which can make it difficult for them to adapt to digital environments."	High	"Some older people may take more time to adapt to new technologies."	"There can be differences across age groups in how new technologies are understood."
"People with low levels of education often lack basic intellectual abilities and cultural knowledge, and therefore cannot handle complex tasks, often being	Standard	"Individuals with limited educational backgrounds can demonstrate abilities across various fields and may be	"People with low levels of education are also able to acquire the skills needed to perform complex tasks."

confined to manual labor throughout their lives."		capable of handling complex tasks."	
"People from low-income backgrounds often lack the diligence and discipline necessary to succeed."	High	"Individuals from low-income backgrounds may sometimes struggle to achieve success for various reasons."	"Individuals from low-income groups may face circumstances that make achieving success more challenging."

Note. The system provides a revised sentence and alternatives based on the input. Users may also adjust the debiasing level applied to the generated sentences. Some examples may include offensive or distressing content.

4.4 Visualization Module

To enable user-centered bias assessment and revision, the FAIRITHM system was developed around three core principles: intuitiveness, repeatability, and user-driven flexibility. The interface was designed to allow users with little or no prior knowledge of linguistic bias to enter sentences, perform iterative correction tasks, and explore alternative outputs easily. The framework presents multiple reduced-bias revision candidates, allowing users to compare, evaluate, and select the most suitable expression. The interface was implemented using Streamlit (Streamlit Developers, 2021), a Python-based web application framework that ensures seamless accessibility for nonexperts in a web environment.

In Figs. 2 and 3, the FAIRITHM interface includes a sidebar configuration panel, input area, results area, analysis summary area, and history preview area. Figure 2 provides an overview of the layout, and Fig. 3 illustrates how sidebar options (e.g., the debiasing level, detailed analysis, highlight changes, and alternative generation) control the interface behavior and amount of information presented to the user.

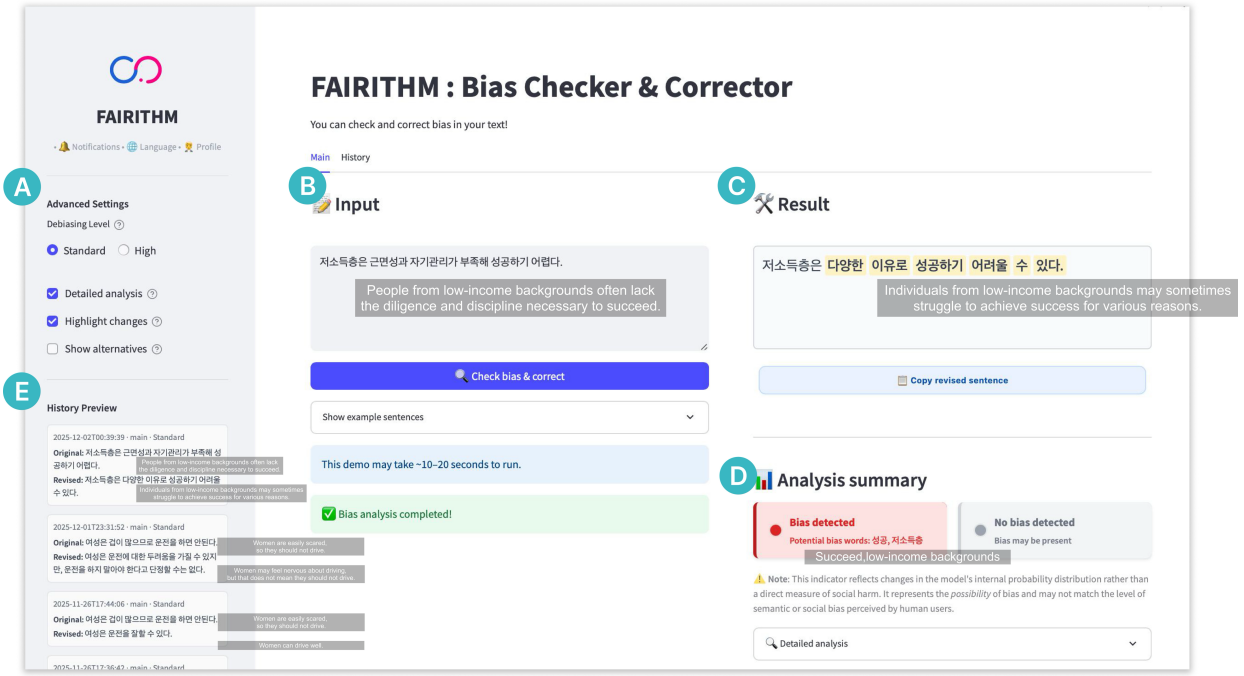

Fig. 2. Main user interface of FAIRITHM. (A) Sidebar controls for the debiasing level, detailed analysis, highlighted changes, and alternative generation. (B) *Input*. (C) *Result* displaying the primary reduced-bias revision. (D) *Analysis summary* displaying the threshold-based flag, highlighted terms, and revision rationale. (E) *History preview* listing recent runs and allowing users to copy previous inputs and revisions.

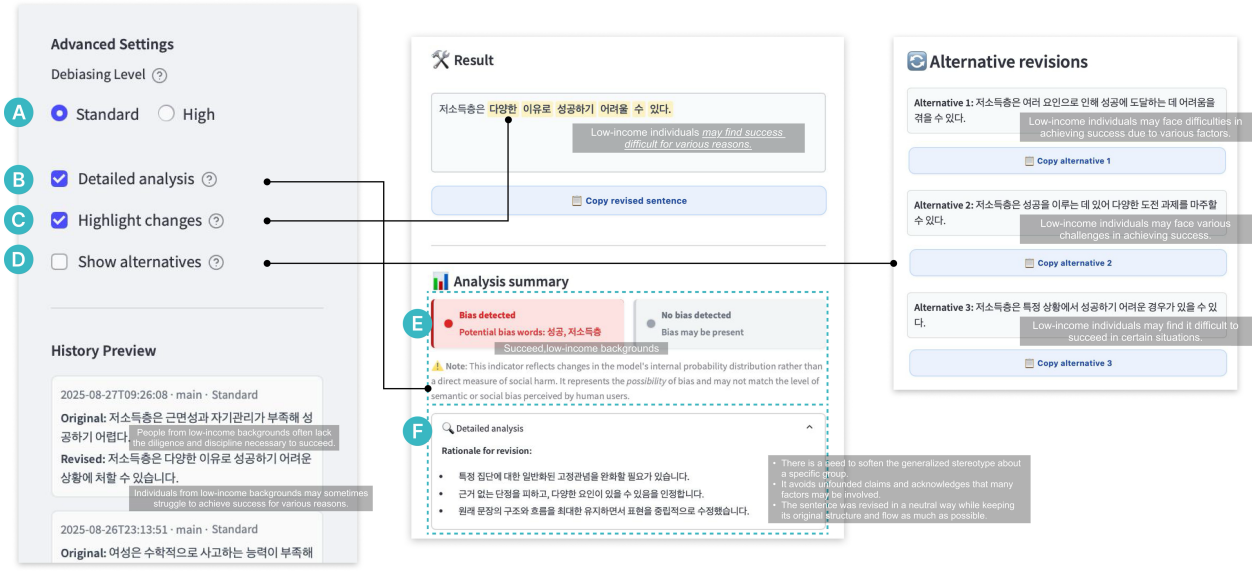

Fig. 3. FAIRITHM option settings and corresponding interface functions. (A) *Debiasing level* with *Standard* or *High* options. (B) *Detailed analysis* displays the revision rationale. (C) *Highlight changes* marks modified terms in the

revised sentence. (D) *Show alternatives* opens the *Alternative Revisions* panel. (E) *Analysis summary* displays the threshold-based red/green flag. (F) *Rationale for revision* explains the basis for the proposed changes.

5. Evaluation Method

A four-part evaluation procedure was employed to assess perceptions of the output of FAIRITHM and the experience of using the system. The evaluation addressed four complementary dimensions: changes in AI-bias perception using the AI Bias Perception Scale (ABPS); the cognitive demands of unaided bias rewriting using time-on-task, the raw form of the NASA Task Load Index (RTLX), and the single-ease question (SEQ); comparative perceptions of original, single-expert-authored, and AI-generated revisions using the Comparative Assessment of Bias-Reduced Sentences (CABRS); and postuse perceptions and usability of FAIRITHM, including the System Usability Scale (SUS). Appendices 2 and 3 provide the complete ABPS and postuse items, and semi-structured interviews were conducted to contextualize the quantitative results.

5.1. Participants

Adults in their 20s to 50s were recruited through university bulletin boards and online communities. All participants provided informed consent. In total, 34 valid responses were analyzed, and Table 7 summarizes the demographics. The sample comprised 18 men (52.9%) and 16 women (47.1%), with most in their 20s or 30s (88.2%). Undergraduates comprised the largest group (38.2%), with over half (55.9%) majoring in engineering. Students comprised 55.9% of the sample, followed by employees (29.4%).

Table 7. Participant demographics (*N = 34*).

Variable	Categories (*n*, %)
Gender	Male: 18 (52.9%), Female: 16 (47.1%)
Age	20s: 21 (61.7%), 30s: 9 (26.5%), 40s: 2 (5.9%), 50s: 2 (5.9%)
Education	Undergraduate (in progress): 13 (38.2%), Bachelor's degree: 12 (35.3%), Graduate (in progress): 5 (14.7%), Graduated (completed): 4 (11.8%)
Field	Engineering: 19 (55.9%), Social sciences: 8 (23.5%), Natural sciences: 3 (8.8%), Humanities: 2 (5.9%), Other: 2 (5.9%)
Occupation	Student: 19 (55.9%), Employee: 10 (29.4%), Unemployed: 4 (11.8%), Homemaker: 1 (2.9%)
AI usage freq.	Daily: 18 (52.9%), 1–3/week: 7 (20.6%), 4–6/week: 3 (8.8%), 1–3/month: 3 (8.8%), Rarely/Never: 3 (8.8%)
AI knowledge	None: 1 (2.9%), Low: 8 (23.5%), Moderate; 15 (44.1%), High: 8 (23.5%), Very high: 2 (5.9%)
Interest in bias	None: 2 (5.9%), Low: 7 (20.6%), Moderate: 3 (8.8%), High: 15 (44.1%), Very high: 7 (20.6%)

5.2. Procedure

The study was conducted remotely and lasted about 30 to 40 min. First, the participants provided informed consent and completed the presurvey (ABPS-Pre). Next, they performed the biased-sentence editing task, during which time-on-task was recorded. The RTLX and SEQ were administered afterward to assess cognitive load. In CABRS, the participants evaluated the original sentence for each selected stimulus and assessed the single-expert-authored and AI revisions individually without source labels. They completed a side-by-side comparison of the two revisions. Afterward, the participants employed FAIRITHM and completed the postuse perception items and SUS. Finally, the participants completed the postsurvey (ABPS-Post) and provided demographic information. Before the semi-structured interview, the participants were informed of which revisions were generated by FAIRITHM and which were authored by the single expert. Then, they participated in a brief semi-structured interview. The experiment was conducted in accordance with the Declaration of Helsinki and the regulations of the institutional review board (IRB) of Hanyang University in the Republic of Korea (IRB approval number: HYUIRB-202511-016).

5.3. Measures

5.3.1. AI Bias Perception Scale

The ABPS was developed to assess changes in the perceptions of the participants regarding generative AI bias before and after using the system. The scale included four core items administered in pre- and postuse surveys: perceived negative social influence of biased expressions (BP1), perceived need for bias detection and mitigation (BP2), perceived difficulty of humans detecting bias (BP3), and perceived effort and time required for humans to revise biased sentences fairly (BP4). The presurvey also included a binary item assessing prior experience with AI bias (EXP), and the postsurvey included one item measuring the perceived ease of AI-assisted revision (BP5). All items, except EXP, were rated on a 7-point Likert scale. Appendix 2 provides the complete set of items.

5.3.2. Biased-Sentence Correction Task and Workload Measurement

Participants were presented with a biased sentence (“Young people are said to be impulsive in their judgments, losing motivation easily after minor failures, and unable to carry out long-term plans.”) and were asked to revise it to a fair and neutral form. The time-on-task was recorded in seconds. Following the task, the subjective workload was

measured using RTLX, including the following six subscales rated on a 7-point Likert scale: mental demand, physical demand, temporal demand, performance (reverse-coded), effort, and frustration. In addition, overall task ease was assessed using an SEQ ("Overall, how easy was it to revise the biased sentence?").

5.3.3. Comparative Assessment of Reduced-Bias Sentences

The CABRS was designed to examine the perceptions of the quality and practical utility of reduced-bias sentences. The CABRS compared AI-generated revisions with a single-expert-authored revision across the six higher-level domains defined in Table 1: (1) family and domestic background, (2) cultural and religious identity, (3) socioeconomic status, (4) physical appearance, (5) gender and sexual identity, and (6) political orientation. Table 3 presents the complete set of six original sentences and their corresponding single-expert-authored and AI-generated revisions. The CABRS survey comprised the following three parts (Table 8):

- Part 1: Evaluation of the original biased sentences. Participants rated the original sentences on the following five criteria using a 7-point scale: grammatical accuracy (O1), naturalness of expression (O2), fairness (O3), absence of stereotypes (O4), and absence of discomfort caused by the content or expression (O5).
- Part 2: Individual evaluation of the revised sentences. Participants completed source-blinded evaluations of the revisions authored by the expert and those generated by the AI separately, using five items parallel to O1 to O5 in Part 1: grammatical accuracy (R1), naturalness of expression (R2), fairness (R3), absence of stereotypes (R4), and absence of discomfort caused by the content or expression (R5), respectively. Two items assessed meaning preservation: preservation of the core meaning of the original sentence (R6) and the avoidance of excessive dilution or weakening of the original meaning (R7). A separate item, R8, assessed the perceived bias improvement relative to the original sentence. Sentence orders were randomized for each participant to minimize order effects.
- Part 3: Comparative evaluation. Participants viewed the single-expert-authored and AI-generated revisions side by side without source labels. They selected Sentence A, Sentence B, or no difference according to three criteria: bias mitigation and fairness (C1), appropriateness for use in media or services (C2), and overall trustworthiness of content and expression (C3).

Table 8. Survey items for the Comparative Assessment of Bias-Reduced Sentences.

Evaluation area	Code	Item

Part 1: Evaluation of original biased sentence		
Sentence completeness	O1	The original sentence is grammatically correct.
	O2	Regardless of its content, the expression (e.g., vocabulary and style) of the original sentence is fluent and natural.
Sentence fairness	O3	The original sentence is unbiased and fair.
	O4	The original sentence does not contain stereotypes about any particular group.
	O5	The content or expression of the original sentence does not cause discomfort.
Individual evaluation of revised sentences (single-expert-authored and AI revisions, evaluated separately)		
Sentence completeness	R1	The revised sentence is grammatically correct.
	R2	Regardless of its content, the expression (e.g., vocabulary and style) of the revised sentence is fluent and natural.
Sentence fairness	R3	The revised sentence is unbiased and fair.
	R4	The revised sentence does not contain stereotypes about any particular group.
	R5	The content or expression of the revised sentence does not cause discomfort.
Meaning preservation	R6	The revised sentence effectively preserves the core meaning of the original text.
	R7	The revised sentence does not excessively dilute or weaken the original meaning.
Perceived bias improvement	R8	The revised sentence effectively improved the bias compared with the original text.
Part 3: Comparative evaluation (Sentence A (AI) vs. Sentence B (single-expert-authored revision) vs. no difference)		
Bias mitigation & fairness	C1	Between the two sentences (Sentence A and B), which do you think is more effective in mitigating bias and is fairer?
Appropriateness for use	C2	If used in actual media articles or services, which of the two sentences (Sentence A or B) do you think is more appropriate?
Overall trustworthiness	C3	Based on your overall evaluation, which of the two sentences' (Sentence A or B) content and expression do you find more trustworthy?

The content validity of the CABRS was supported by developing the survey items based on prior research on linguistic quality, bias, and revision effectiveness (Ma et al., 2020; Pryzant et al., 2020; Z. Zhang et al., 2018), followed by an expert review and refinement of the item wording and evaluation procedure via a small pilot test ($N = 6$). The item groupings for sentence completeness, sentence fairness, meaning preservation, and perceived bias improvement were specified a priori based on the item content and conceptual definitions in Table 8.

5.3.4. Usability Evaluation of FAIRITHM

Following their interaction with FAIRITHM, the participants evaluated the usability and their overall user experience via a two-part survey. The first part included four self-developed items assessing their postuse perceptions of the

system. The items assessed the perceived appropriateness of the system bias judgment (T1), the perceived clarity of the reasons and explanations for revision (T2), the expectation that the system could consistently improve other biased sentences (T3), and future trust in and intention to use the system (T4). Responses to each item were recorded on a 7-point Likert scale. The second part measured system usability using the 10-item SUS (Brooke, 1996), which assessed intuitiveness, efficiency, and ease of learning on a 5-point scale. Score on the SUS were converted to a 0 to 100 scale following the standard scoring procedure. Appendix 3 presents the complete set of survey items.

Before completing the survey, the participants viewed a tutorial video introducing the service. Then, they were encouraged to input biased sentences, review the system revisions, and complete the survey afterward.

5.3.5 Qualitative Follow-Up Interviews

Semi-structured follow-up interviews were conducted with all survey participants to identify the dimensions of user experience and bias-related reflections that were not fully captured by the quantitative results. The questions focused on (a) their experiences of revising the biased sentences before and after using FAIRITHM and (b) new insight gained from the task. The participants were encouraged to speak freely, and their responses were analyzed using a thematic analysis (Braun & Clarke, 2006) to identify recurring patterns and key themes. These qualitative findings provide a deeper interpretation and contextualization of the quantitative results.

5.4 Statistical Analysis

All analyses were conducted using the Rj Editor in Jamovi (v. 2.6.44) running R (v. 4.4.1). The primary packages were *lme4*, *lmerTest*, *emmeans*, *car*, and *dplyr*. The sum-to-zero contrasts were applied for Type III fixed-effect tests in the mixed-effect models. Denominator degrees of freedom for the linear mixed models (LMMs) were obtained using the Satterthwaite approximation, whereas fixed effects in the generalized LMMs were evaluated using the Type III Wald chi-square test. This study reports the estimated marginal mean (EMM), contrast estimate or odds ratio (OR), 95% confidence interval (CI), and *p*-value.

The pre–post responses to the four common ABPS items, BP1 to BP4, were compared using the paired Wilcoxon signed-rank test, with the four *p*-values adjusted using the Holm procedure. The post-only item BP5 and the four postuse perception items T1 to T4 were each compared with the midpoint of four on the 7-point scale using the one-sample Wilcoxon signed-rank test. The four *T*-item *p*-values were also Holm-adjusted. The Wilcoxon results are reported with Hodges–Lehmann pseudomedian or paired location-shift estimates, nonparametric 95% CIs, and the

standardized Wilcoxon effect size $r = |Z|/\sqrt{N}$. Prior experience with AI bias (EXP) was presented using frequencies and percentages. The RTLX and SEQ results from the manual bias-rewriting task were reported using mean, standard deviation, median, interquartile range, and range.

Based on the item groupings specified as a priori in Table 8, sentence completeness was calculated as the mean of O1 and O2 for the original sentence and the mean of R1 and R2 for each of the single-expert-authored and AI revisions. Sentence fairness was computed as the mean of O3 to O5 for the original sentence and the mean of R3 to R5 for each revision. Sentence completeness and fairness was analyzed using integrated LMMs that included the original sentence, single-expert-authored revision, and AI revision.

The sentence condition (*condition*) comprised three levels: the original biased sentence, single-expert-authored revision, and AI revision. The bias category (domain) comprised the six higher-level categories in Table 1, which are represented by the evaluation sentences in Table 3. Because each category was represented by a single evaluation sentence, the domain-related findings were interpreted as differences between the sentences selected for the six categories rather than as general category effects.

The fixed effects for sentence completeness and fairness were *condition*, *domain*, and *condition* × *domain* interaction. The maximal random-effect structure included a participant random intercept, a participant-specific random slope for the condition, and a participant-by-category-sentence random intercept to account for repeated evaluations of the same category sentence across the three sentence conditions. When a boundary-singular fit occurred, unsupported random slopes and participant-by-category-sentence random intercepts were removed sequentially following a prespecified simplification procedure. After this procedure, the completeness model retained the participant and participant-by-category-sentence random intercepts, whereas the fairness model retained only a participant random intercept.

The three planned contrasts averaged across the sentences from the six categories were AI versus the single expert, AI versus the original, and the single expert versus the original. The three *p*-values were adjusted using the Holm procedure. When the condition × domain interaction was significant, the same three contrasts were estimated in the evaluation sentence representing each category. For the category-specific analyses, the six category sentences of the same contrast type were treated as a single family for multiple-comparison correction, and their *p*-values were Holm-adjusted.

Residual-versus-fitted and Q–Q plots were inspected to evaluate the residual distribution and homoscedasticity of the LMMs. Because sentence-completeness ratings demonstrated a pronounced ceiling effect, the three overall condition contrasts averaged across the six category sentences were examined using participant-level paired-rank tests. The same participant-level analysis was applied to sentence fairness as a complementary robustness check. These sensitivity analyses were conducted to examine the direction and significance of the overall condition contrasts. They did not independently test the condition × domain interaction or the simple contrasts for individual category sentences.

Meaning preservation (mean of R6 and R7) and the perceived bias improvement (assessed using R8) were analyzed separately from the common CABRS items. For each outcome, an LMM was fitted with the revision *source* (single-expert-authored or AI revision), *domain*, and the *source* × *domain* interaction as fixed effects. The maximal random-effect structure and prespecified simplification procedure for boundary-singular fits were the same as those for the common-item analyses. The meaning-preservation and perceived-bias-improvement models retained the maximal random-effect structure. When the source × domain interaction was significant, the single-expert-versus-AI contrasts were estimated in the evaluation sentence representing each category, and the six *p*-values were Holm-adjusted.

The three-category choices in CABRS Part 3 were analyzed using a two-part mixed-effect logistic regression. The first model distinguished decisive choices (selection of the AI or single-expert-authored revision) from the no-difference responses. The second model distinguished AI from the single-expert selections among decisive choices.

The fixed effects in both models were the evaluation *criteria* (bias mitigation and fairness, appropriateness for use, and overall trustworthiness), *domain*, and the *criterion* × *domain* interaction. A participant random intercept was included to account for repeated responses. Adding a participant-by-category-sentence random intercept produced a boundary-singular fit and nearly degenerate predicted probabilities; therefore, it was not retained in the final models.

When the criterion × domain interaction was not significant, the reduced main-effect models were fitted. These models were employed to evaluate the decisive-choice patterns and estimate the criterion-specific conditional ORs for selecting the AI revision rather than the single-expert-authored revision among the decisive choices. The *p*-values for the three criterion-specific ORs were Holm-adjusted, and a participant-level preference index was analyzed as a complementary robustness check. The SUS scores were calculated using the standard 0 to 100 scoring procedure and were summarized using the mean, standard deviation, median, interquartile range, and 95% CI for the mean.

6. Results

6.1. Survey Results on AI Service Usage

Regarding the frequency of generative AI service use, 18 participants (52.9%) reported using such services almost daily, and more than 80% reported regular use. Regarding knowledge of AI technology, the largest group indicated a moderate level (n = 15, 44.1%), while low and high levels were reported equally (n = 8, 23.5% each). Interest in problems of social bias and fairness was also substantial: 15 participants (44.1%) reported high interest, and 7 (20.6%) reported very high interest, accounting for 64.7% of the sample. In terms of specific services, ChatGPT was the most frequently used (97.1%), followed by Gemini (70.6%), Perplexity (52.9%), Wrtn (47.1%), and Claude (41.2%; Fig. 4).

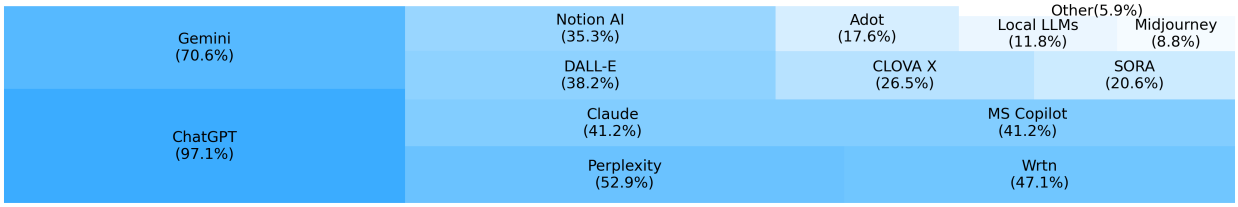

Fig. 4. Generative AI services employed by the participants (multiple responses; N = 34).

6.2. Results of the AI Bias Perception Scale

In response to the pre-only EXP item, 10 participants (29.4%) reported encountering biased or unfair expressions while using generative AI services, whereas 24 participants (70.6%) reported no such experience. None of the four ABPS items administered before and after the study displayed a significant change after Holm correction (all Holm-adjusted $p \geqslant .250$; Table 9). Thus, the study did not find clear evidence of changes in perceived societal influence, the need for bias detection and mitigation, the difficulty of unaided human detection, or the effort and time required for human revision.

For the post-only BP5 item, the mean was 5.76 (SD = 1.26), and the median was 6 (Table 9). The Hodges–Lehmann pseudomedian was 6.00 (95% CI [5.50, 6.50]), and the rating was significantly above the scale midpoint of 4 ($p < .001$, r = .77). Overall, 91.2% of the participants selected values from 5 to 7. This result implies that the participants perceived AI support as potentially making bias rewriting easier than unaided human revision.

Table 9. Pre–post changes in AI bias perceptions and postuse perceived ease of AI-assisted revision (N = 34).

Item	Pre-mean (SD)	Post-mean (*SD*)	HL estimate [95% CI]	Effect size *r*	*p*	Holm-adjusted *p*
BP1. Negative societal influence	5.53 (1.35)	5.76 (1.13)	0.50 [−1.00, 2.00]	.186	.278	.556
BP2. Need for detection and mitigation	5.76 (1.54)	6.09 (1.06)	1.00 [0.00, 1.50]	.319	.063	.250
BP3. Difficulty of unaided human detection	5.91 (1.08)	6.21 (0.84)	1.00 [0.00, 2.00]	.261	.128	.383
BP4. Human effort and time for revision	6.26 (0.71)	6.32 (0.59)	0.00 [−1.00, 1.00]	.072	.674	.674
BP5. Perceived ease of AI-assisted revision	-	5.76 (1.26)	6.00 [5.50, 6.50]	.770	< .001	-

Note. For BP1 to BP4, Hodges–Lehmann (HL) estimates and 95% confidence intervals (CIs) represent paired post–prelocation shifts. For BP5, they represent the one-sample pseudomedian and its CI; BP5 was compared with the scale midpoint of 4. The standardized Wilcoxon effect size was calculated as $r = (|Z|)/\sqrt{N}$, where $N = 34$. The BP1 to BP4 *p*-values were Holm-adjusted across the four comparisons.

6.3. Manual Bias-Rewriting Task and Workload

In the manual bias-rewriting task, mental demand (*M* = 5.65, *SD* = 1.15) and effort (*M* = 5.38, *SD* = 1.02) received the highest RTLX ratings (Fig. 5A). Physical demand was also numerically higher than the scale midpoint (*M* = 5.03, *SD* = 1.34). Temporal demand (*M* = 4.32, *SD* = 1.77) and frustration (*M* = 3.85, *SD* = 1.89) were comparatively lower but exhibited greater between-participant variability. The median time-on-task was 105 s (*M* = 164.75, *SD* = 160.88; range = 30.13 to 694.59), and Fig. 5B presents its distribution. The overall RTLX score was M = 4.76, SD = 1.02, and the SEQ rating was M = 5.38, SD = 1.07 (Fig. 5C). These findings suggest that manual bias rewriting requires substantial judgment and effort, although the participants did not regard the task to be excessively difficult.

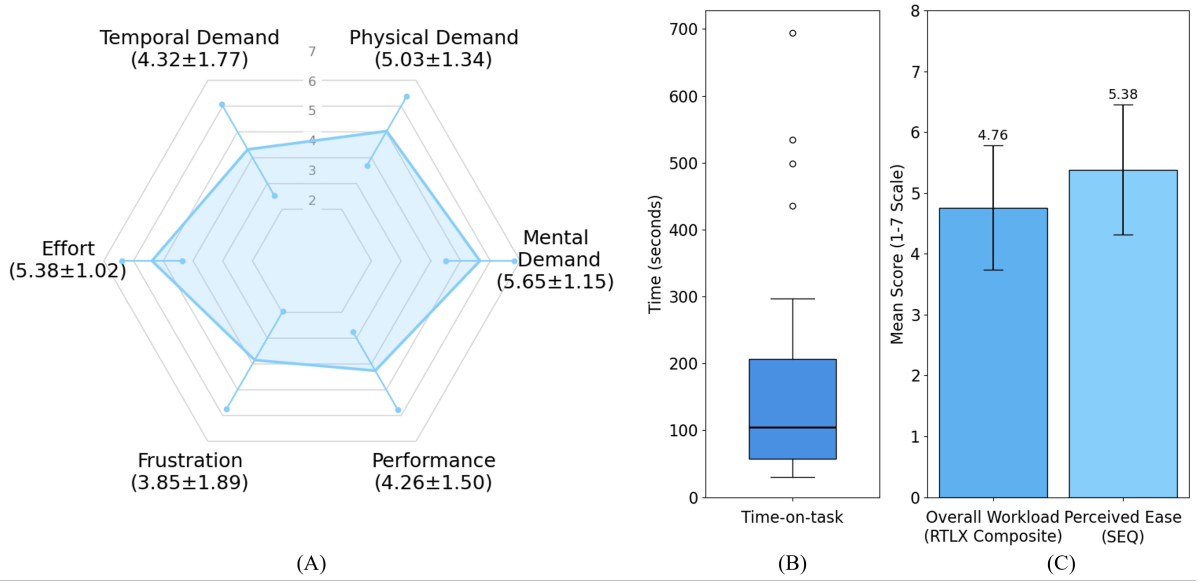

Fig. 5. Workload, time-on-task, and perceived ease in the manual bias-rewriting task (*N* = 34). (A) Radar chart illustrating mean scores (±standard deviation [*SD*]) for the six Raw NASA Task Load Index (RTLX) dimensions (1–

7 scale). (B) Boxplot of time-on-task (s), displaying the median, interquartile range, and outliers. (C) Bar chart of the overall workload (RTLX composite) and perceived ease (single-ease-question, SEQ) with error bars indicating ±*SD*.

6.4. Results of the Comparative Assessment of Bias-Reduced Sentences

The CABRS common items were analyzed using mixed-effect models that included the original sentence, single-expert-authored revision, and AI revision. Appendices 4A and 4B provide composite- and item-level descriptive statistics by condition and category sentence, respectively. Appendix 7 presents the participant-level sensitivity analyses.

6.4.1. Sentence Completeness and Fairness

For sentence completeness, the estimated marginal means (EMMs) were 6.02 for the original, 5.72 for the single-expert-authored revision, and 6.00 for the AI revision (Table 10A). The main effect of sentence condition was significant ($p < .001$), as was the condition × domain interaction ($p < .001$, Table 10C). In the overall planned contrasts, AI revisions were rated 0.277 points higher than the single-expert-authored revisions (95% CI [0.106, 0.448], Holm-adjusted $p = .003$; Table 10B). No significant difference was found between the AI revisions and the original sentences, whereas the single-expert-authored revisions were rated 0.294 points lower than the original sentences (95% CI [−0.465, −0.123], Holm-adjusted $p = .002$). At the category-sentence level, AI revisions received higher completeness ratings for family and religion sentences, whereas the single-expert-authored revisions were rated higher for the political sentence. Appendix 5A presents the complete simple contrasts.

For sentence fairness, the EMMs were 1.91 for the original, 4.58 for the single-expert-authored revision, and 5.46 for the AI revision (Table 10A). The main effect of the sentence condition was significant ($p < .001$), and the condition × domain interaction was also significant ($p = .001$, Table 10C). The AI revisions were rated 0.877 points higher than the single-expert-authored revisions (95% CI [0.629, 1.126], Holm-adjusted $p < .001$; Table 10B). Both revised conditions received higher fairness ratings than the original for all six category sentences. The AI-versus-single-expert difference was statistically significant for the category sentences representing religion and physical appearance, whereas no clear Holm-adjusted difference was found for the socioeconomic status (SES), gender, or political sentences. The family result was treated as an exploratory near-threshold contrast. Appendix 5B presents full simple contrasts.

Participant-level paired-rank sensitivity analyses reproduced the direction and significance pattern of the three overall contrasts for completeness and fairness (Appendix 7A). These analyses evaluated contrasts that were averaged across the six category sentences and did not independently test the condition × domain interactions or the simple contrasts for individual category sentences.

Table 10. Sentence completeness and fairness across the original, single-expert-authored, and AI revision conditions.

Panel A. Estimated marginal means			
Outcome	Original EMM [95% CI]	Single expert EMM [95% CI]	AI EMM [95% CI]
Sentence completeness	6.02 [5.81, 6.23]	5.72 [5.51, 5.93]	6.00 [5.79, 6.21]
Sentence fairness	1.91 [1.67, 2.14]	4.58 [4.35, 4.81]	5.46 [5.22, 5.69]

Panel B. Planned contrasts				
Outcome	Contrast	Estimate	95% CI	Holm-adjusted *p*
Completeness	AI − Single expert	0.277	[0.106, 0.448]	.003
	AI − Original	−0.017	[−0.188, 0.154]	.843
	Single expert − Original	−0.294	[−0.465, −0.123]	.002
Fairness	AI − Single expert	0.877	[0.629, 1.126]	< .001
	AI − Original	3.549	[3.300, 3.798]	< .001
	Single expert − Original	2.672	[2.423, 2.920]	< .001

Panel C. Omnibus fixed effects				
Outcome	Effect	*F*	*df*	*p*
Completeness	Condition	7.23	2, 396	< .001
	Domain	6.03	5, 165	< .001
	Condition × domain	5.04	10, 396	< .001
Fairness	Condition	427.25	2, 561	< .001
	Domain	2.22	5, 561	.051
	Condition × domain	2.97	10, 561	.001

Note. EMM: estimated marginal mean. EMMs and contrasts are estimated using linear mixed models and averaged across the six category-specific sentences. Planned-contrast *p*-values are Holm-adjusted across the three comparisons in each outcome.

6.4.2. Meaning Preservation and Perceived Bias Improvement

The revision-specific items were analyzed separately in terms of meaning preservation (mean of R6 and R7) and perceived bias improvement (R8; Table 11). The EMMs for meaning preservation were 4.75 for the single-expert-authored revisions and 4.01 for the AI revisions. The difference between the single expert and AI was significant (estimate = 0.738, 95% CI [0.413, 1.062], $p < .001$). The source × domain interaction was also significant ($p < .001$). The single-expert advantage was observed for the family, religion, and political category sentences. Appendix 5C presents the complete category-sentence contrasts.

For R8, which assessed perceived bias improvement, the EMMs were 4.84 for the single-expert-authored revisions and 5.06 for the AI revisions. Although the AI mean was numerically higher, the difference between the single-expert and AI revisions was not significant ($p = .283$). The source × domain interaction was also not significant ($p = .163$). Thus, perceived bias improvement did not differ significantly between the AI and single-expert-authored revisions.

Table 11. Meaning preservation and perceived bias improvement for single-expert-authored and AI revisions.

Outcome	Single-expert EMM [95% CI]	AI EMM [95% CI]	Single expert − AI	95% CI	*p*	Source × domain *p*
Meaning preservation (R6 and R7)	4.75 [4.37, 5.12]	4.01 [3.56, 4.46]	0.738	[0.413, 1.062]	<.001	<.001
Perceived bias improvement (R8)	4.84 [4.52, 5.16]	5.06 [4.74, 5.39]	−0.225	[−0.646, 0.195]	.283	.163

Note. AI = artificial intelligence; CABRS: Comparative Assessment of Bias-Reduced Sentences; CI = confidence interval; EMM = estimated marginal mean; R6 to R8 = CABRS revision-specific items.

6.4.3. Comparative Choices in Part 3 of the Comparative Assessment of Bias-Reduced Sentences

Each criterion included 204 responses from 34 participants across six categories of sentences. For bias mitigation and fairness, the observed choices were the AI revision (n = 121, 59.3%), no difference (n = 23, 11.3%), and single-expert-authored revision (n = 60, 29.4%). For appropriateness for use, the corresponding counts were 110 (53.9%), 33 (16.2%), and 61 (29.9%), respectively. For overall trustworthiness, the corresponding counts were 104 (51.0%), 30 (14.7%), and 70 (34.3%), respectively. Figure 6 presents the observed response distributions, and Table 12 reports the conditional ORs estimated by the mixed-effect model.

In the first mixed-effect logistic model, the likelihood of making a decisive choice rather than selecting no difference did not vary significantly by criterion (χ^2 (2) = 2.83, p = .243), by domain (χ^2 (5) = 10.11, p = .072), or by the criterion × domain interaction (χ^2 (10) = 3.52, p = .966). In the second full mixed-effect logistic model, which was restricted to decisive choices, the omnibus criterion effect was not significant (χ^2 (2) = 1.85, p = .397), whereas the domain main effect was significant (χ^2 (5) = 11.17, p = .048). The criterion × domain interaction was also not significant (χ^2 (10) = 5.68, p = .842).

Criterion-specific conditional contrasts estimated by the reduced main-effect model presented higher odds of selecting the AI revision rather than the single-expert-authored revision for bias mitigation and fairness (OR = 2.58, 95% CI [1.43, 4.68], Holm-adjusted p = .005) and for appropriateness for use (OR = 2.21, 95% CI [1.22, 4.01], Holm-adjusted p = .018). For overall trustworthiness, the conditional OR was 1.78 (95% CI [0.99, 3.22], Holm-adjusted p = .054). Although the point estimate favored AI, it did not meet the adjusted significance threshold. The nonsignificant omnibus criterion effect indicates that the magnitude of AI preference did not significantly differ across the three criteria.

The significant domain main effect suggests that the overall tendency to select the AI revision rather than the single-expert-authored revision varied across the six evaluated sentences. However, because the criterion × domain interaction was not significant, individual category sentences are not interpreted as indicating criterion-specific AI or single-expert preferences. Appendix 6 provides the raw frequencies by category and criterion. A complementary participant-level preference-index analysis revealed the same conclusion pattern: AI-oriented results for fairness and appropriateness with no significant result for overall trustworthiness (Appendix 7B).

Table 12. Comparative choices and conditional odds ratios for AI-generated versus single-expert-authored revisions in Part 3 of the Comparative Assessment of Bias-Reduced Sentences.

Criterion	AI *n* (%)	No difference *n* (%)	Single-expert-authored revision *n* (%)	Conditional OR	95% CI	Holm-adjusted *p*
Bias mitigation and fairness	121 (59.3)	23 (11.3)	60 (29.4)	2.58	[1.43, 4.68]	.005
Appropriateness for use	110 (53.9)	33 (16.2)	61 (29.9)	2.21	[1.22, 4.01]	.018
Overall trustworthiness	104 (51.0)	30 (14.7)	70 (34.3)	1.78	[0.99, 3.22]	.054

Note. Conditional odds ratio (OR) compares artificial intelligence (AI) with the single-expert-authored revision among decisive choices after excluding "no difference" responses. CI: confidence interval.

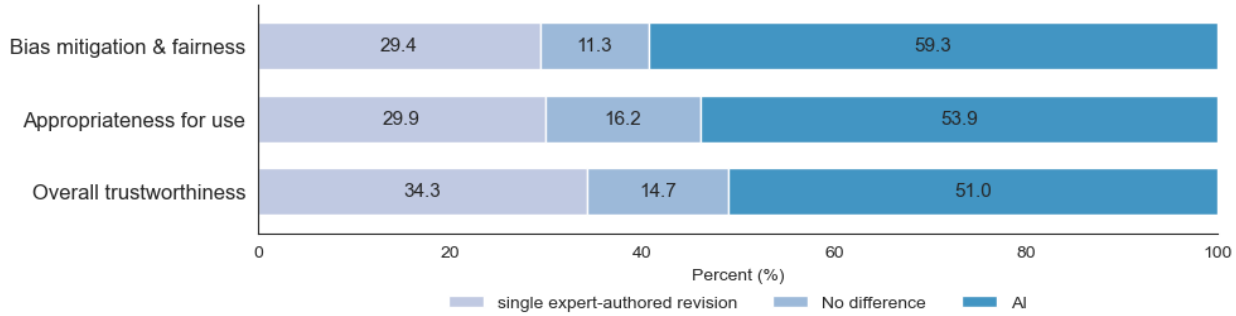

Fig. 6. Observed comparative-choice distributions for Part 3 of the Comparative Assessment of Bias-Reduced Sentences across the three evaluation criteria. Each criterion included 204 responses from 34 participants across six category-specific sentences.

6.5. Postuse Perceptions and System Usability of FAIRITHM

The perceived appropriateness of the bias judgment (T1: $M = 5.38$, $SD = 1.21$), perceived clarity of the revision rationale (T2: $M = 5.74$, $SD = 1.24$), and expected consistency on other biased sentences (T3: $M = 5.71$, $SD = 1.03$) were significantly over the midpoint of 4 on the 7-point scale (Holm-adjusted $p < .001$; Table 13). In contrast, future trust and use intention (T4) had a mean of 3.94 ($SD = 0.98$) and did not differ from the midpoint ($p = .743$). The mean

SUS score was 72.72 (*SD* = 15.85; 95% CI [67.19, 78.25]), indicating acceptable perceived usability in the study context. Appendix 3 provides the complete T1 to T4 and SUS items.

Table 13. Postuse perceptions of FAIRITHM (*N* = 34).

Code	Measure	Mean (*SD*)	Median	HL pseudomedian [95% CI]	Effect size *r*	Holm-adjusted *p*
T1	Perceived appropriateness of bias judgment	5.38 (1.21)	6	5.50 [5.50, 6.00]	.74	<.001
T2	Perceived clarity of revision rationale	5.74 (1.24)	6	6.00 [5.50, 6.50]	.77	<.001
T3	Expected consistency	5.71 (1.03)	6	6.00 [5.50, 6.00]	.84	<.001
T4	Future trust and use intention	3.94 (0.98)	4	4.00 [3.00, 5.00]	.06	.743

Note. T1 to T4 were compared with the 7-point scale midpoint of 4 using the one-sample Wilcoxon signed-rank test. The standardized Wilcoxon effect size was calculated as $r = (|Z|)/\sqrt{N}$.

6.6. Qualitative Follow-up Interview Results

The analysis of the follow-up interviews revealed several critical themes. Many participants described the biased-sentence correction task as "more difficult than expected" and expressed uncertainty about whether their revisions were appropriate. These experiences often led the participants to perceive AI systems more positively, noting that they could save time compared with human revision and were more practical to use.

Notably, some participants' trust in AI shifted after completing the survey. Although some initially believed that human experts would be superior at revising biased sentences, several were surprised to learn that the sentences they had assumed to be expert revisions were AI revisions. Confronted with outcomes that contradicted their expectations, these participants reported a stronger sense of trust in the capabilities of AI.

The interviews also highlighted a recurring tension between fairness and meaning preservation. Some participants emphasized that fairness should take precedence, even if the original meaning was partially compromised, whereas other participants prioritized conveying the intended meaning. These contrasting perspectives imply that the relative strengths of humans and AI may vary depending on the situational demands of sentence revision, indicating the potential for complementary roles between them.

7. Discussion

7.1. Cognitive Demands of Bias Rewriting and the Perceived Value of AI Assistance

The RTLX results suggested that bias rewriting is a cognitively demanding task requiring substantial mental effort. Mental demand (*M* = 5.65, *SD* = 1.15) and effort (*M* = 5.38, *SD* = 1.02) were the highest-rated RTLX dimensions, and

the overall RTLX mean was 4.76 (*SD* = 1.02). These results indicate that bias rewriting is a complex task involving understanding the sociotechnical context and exercising normative judgment. Gillespie (2018) characterized content moderation as a normative and contextual activity rather than a purely technical operation. Similarly, Roberts et al. (2023) reported that reviewing socially sensitive content can be cognitively and emotionally demanding. The median time-on-task was 105 s, whereas the mean was 164.75 s and the observed range was wide (30.13 to 694.59 s), indicating substantial variation in the amount of review and judgment required across participants. Moreover, the mean SEQ rating was 5.38 (*SD* = 1.07), indicating that the participants did not rate the overall ease of performing the task as low despite the substantial mental effort involved.

Given these demands, the participants favorably evaluated the potential value of AI assistance. The post-only item assessing the perceived ease of AI-assisted revision (BP5) reached a mean of 5.76 (*SD* = 1.26), with a Hodges–Lehmann pseudomedian of 6.00, significantly higher than the scale midpoint of 4 ($p < .001$, $r = .77$; Table 9). Moreover, 91.2% of the participants selected values between 5 and 7. These findings suggest that the participants broadly expected AI assistance to make bias rewriting easier than unaided human revision. However, the study did not include a matched condition in which the participants completed the same task using FAIRITHM; therefore, BP5 represents the perceived ease of AI-assisted revision rather than evidence that FAIRITHM reduced cognitive workload or time-on-task.

In contrast, the pre–post ABPS results did not present a clear change in bias-related perceptions. None of the four common items, perceived negative societal influence of biased expressions (BP1), perceived need for bias detection and mitigation (BP2), perceived difficulty of unaided human detection (BP3), and perceived human effort and time required for fair revision (BP4), were significant after Holm correction (all adjusted *p*-values ≥ .250; Table 9). Thus, the study did not find evidence that a brief experience of using FAIRITHM changed the bias-related perceptions of the participants.

Overall, the RTLX and ABPS findings indicate the substantial cognitive demands of manual bias rewriting and the participants' expectations that AI could offer useful support. When considered alongside the interview findings, the initial favorable evaluations of FAIRITHM appear to have been more strongly grounded in its instrumental value for supporting a difficult rewriting task than in measurable attitudinal change. The participants described manual rewriting as more difficult than expected and expressed uncertainty about whether their wording was appropriate. Therefore,

the immediate user value of FAIRITHM appears to lie in generating alternatives for comparison and supporting human review, rather than in producing short-term learning or measurable improvements in bias perception.

7.2. Trade-off Between Fairness and Meaning Preservation

The mixed-effect analyses revealed that the AI-generated and single-expert-authored revisions had different strengths across the evaluated dimensions. Across the sentences representing the six categories, both revision conditions were rated substantially higher on perceived fairness than the original sentences, with the AI revisions receiving an average fairness advant*a*ge of 0.877 points over the single-expert-authored revisions (Table 10). This pattern suggests that the AI-generated revisions were particularly effective at reducing expressions perceived as socially biased.

A smaller difference emerged for sentence completeness. The AI revisions received higher completeness ratings than the single-expert-authored revisions but did not differ significantly from the original sentences (Table 10). Thus, this result does not indicate that AI improved grammatical accuracy or naturalness beyond the original text. Instead, the result suggests that the AI revisions received higher completeness ratings than the particular revision strategy employed by the single expert.

The significant condition × domain interactions for completeness and fairness indicated that these average effects were not uniform across the six evaluated sentences (Table 10). For completeness, the AI revisions received higher ratings for the family and religion category sentences, whereas the single-expert-authored revision received a higher rating for the political sentence. For fairness, the clearest AI advantages were observed for the religion and physical appearance category sentences. No clear Holm-adjusted differences were found for the SES, gender, or political sentences (Appendix 5).

A clearer difference emerged for meaning preservation. Across the six evaluated sentences, the single-expert-authored revisions received an average meaning-preservation advantage of 0.738 points over the AI revisions (Table 11). This advantage was most evident for the family, religion, and political sentences (Appendix 5). The lower meaning-preservation ratings for the AI revisions imply that they reformulated the biased premises more extensively, whereas the single-expert-authored revisions retained more of the original propositions using qualifiers and contextual constraints. Prior work on subjective-bias neutralization has also identified the preservation of bias-independent context as a central challenge, noting that automated neutralization methods can remove critical information unrelated to the bias (Madanagopal & Caverlee, 2023).

However, the higher meaning-preservation ratings should not be interpreted as evidence that the single-expert-authored revision strategy also achieved superior bias mitigation. Perceived bias improvement did not differ significantly between the single-expert-authored and AI revisions, and this pattern did not vary significantly across the evaluated sentences (Table 11). Therefore, the participants regarded the single-expert-authored revisions as preserving meaning more effectively but did not identify a clear advantage for either strategy in the extent to which bias was improved. The principal identified tension concerns how much semantic and structural change is acceptable in the pursuit of fairness, rather than the simple superiority of one strategy for bias reduction.

The direct choices in CABRS Part 3 support this distinction. Among decisive choices, participants were more likely to select the AI revision for bias mitigation and fairness (OR = 2.58) and for appropriateness for use (OR = 2.21), whereas no equally clear preference emerged for overall trustworthiness (Table 12). The participants preferred the AI revisions for fairness and appropriateness for use but did not display an equally clear preference regarding the overall trustworthiness of their content and expression.

The interview findings provide the qualitative context for these quantitative patterns. The participants differed in whether fairness should take precedence over preserving the original meaning, suggesting that the relative importance of these two objectives may depend on the purpose and context of the revision. The findings indicate that AI revisions and the single-expert-authored revisions were evaluated differently in terms of fairness, completeness, and meaning preservation. Overall, these output-level findings support retaining human judgment in the revision process.

The detection-signal validation reinforces this collaborative framing from a complementary direction. As detailed in Section 4.2.2, the PLL-difference signal contained directional information attributable to the unlearning step but showed insufficient detection performance for autonomous classification, supporting its use as an advisory screening signal. Notably, either extreme would undermine the proposed framework. A near-perfect detector would render human review superfluous at the screening stage, whereas a pure-noise signal would make the assessment module unnecessary. The measured compromise is the regime in which a division of labor (machine screening that directs attention, followed by human judgment that decides) is the rational design. Human oversight in FAIRITHM is not a precautionary disclaimer but an evidence-based design requirement.

7.3. Perceived Usability Versus Future Trust and Use Intention

At the system level, the findings indicate that perceived usability and willingness to rely on FAIRITHM are distinct. The mean SUS score was 72.72, indicating acceptable perceived usability. According to Bangor et al. (2008), this score is within an acceptable range and is above the commonly cited average benchmark discussed by Sauro (2018). The participants were generally able to operate and understand FAIRITHM in the study context.

The item-level results provide a more differentiated account. The perceived appropriateness of the bias judgment by the system (T1; $M = 5.38$, $SD = 1.21$), the perceived clarity of the revision rationale (T2; $M = 5.74$, $SD = 1.24$), and the expected consistency of the framework on other biased sentences (T3; $M = 5.71$, $SD = 1.03$) were significantly higher than the midpoint of 4 (all Holm-adjusted p-values < .001; Table 13). In contrast, the item assessing future trust in and intention to use FAIRITHM (T4; $M = 3.94$, $SD = 0.98$) did not differ from the midpoint ($p = .743$). Thus, the participants positively evaluated several immediate system features without expressing a strong future intention to trust and use the system.

The output-level comparison revealed a similar distinction. For overall trustworthiness, AI revisions accounted for 51.0% of the observed responses, and single-expert-authored revisions accounted for 34.3%. However, the conditional OR comparing AI with the single-expert-authored revisions among decisive choices did not meet the Holm-adjusted significance threshold (OR = 1.78, $p = .054$; Table 12). Thus, the preference for AI revisions in terms of fairness and appropriateness for use should not be generalized to comprehensive trust in AI-generated revisions. Evaluating output as fair or practically useful is distinct from being willing to delegate socially consequential judgments to the system. This gap may reflect several interrelated considerations, including the responsibility for socially sensitive decisions, the need to verify system output, the workflow-integration demand, and the possibility of semantic change during fairness-oriented rewriting.

One possible explanation concerns responsibility and overreliance. Bias rewriting requires judgment regarding what constitutes fairness and how much semantic change is acceptable. Ha et al. (2024) found that users were more likely to accept AI suggestions when performing a more difficult task, even when those suggestions were less accurate. Although this finding does not directly explain the intentions of the participants to use FAIRITHM, it illustrates why critical human review is crucial when a demanding task may encourage reliance on imperfect AI guidance. The neutral

T4 response may reflect caution about delegating socially sensitive judgment rather than a simple rejection of the system.

Perceived explanation clarity did not directly translate into future trust and use intention. Participants evaluated the revision rationale of FAIRITHM as clear (T2), whereas T4 remained neutral. In the scope of this study, this difference indicates that understandable explanations can support inspection of system output but alone do not demonstrate that users are prepared to rely on the underlying judgment.

The effort required to integrate FAIRITHM into an existing workflow may also be relevant. Venkatesh et al. (2003) distinguished effort expectancy and facilitating conditions from performance-related benefits and identified them as factors associated with technology acceptance and use. The participants may have recognized the immediate utility of FAIRITHM while anticipating additional comparison, verification, or responsibility during continued use. However, because this study did not directly measure workflow-integration costs, this explanation should be examined in future research rather than be treated as an observed mechanism.

Caution may have also been related to the observed fairness–meaning preservation trade-off. The single-expert-authored revisions were rated 0.738 points higher than the AI revisions on meaning preservation (Table 11), whereas the AI revisions received higher fairness ratings. The neutral T4 response may indicate that the participants distinguished the utility of fairness-oriented AI output from the willingness to rely on such output when semantic or contextual information might be lost. This relationship was not directly tested and should be interpreted as a possible explanation rather than as a causal conclusion.

Overall, FAIRITHM was perceived as usable, and several of its immediate features were evaluated positively. However, the overall pattern reflects a cautious approach to AI-assisted bias rewriting, consistent with the human-oversight design stance of this study.

8. Implications

8.1. Theoretical Implications

This study bridges algorithmic bias mitigation and human–computer interaction and offers three theoretical contributions to responsible AI research. First, this work extends the cultural localization of fairness research beyond the Anglocentric paradigm. Prior bias research has focused on English-centric datasets and norms. Lee et al. (2024)

found that LLMs can reproduce biases rooted in the Korean social context that differ from those in Western benchmarks. By applying and evaluating PCGU in a Korean-language assessment framework, this study demonstrates the importance of treating bias mitigation as a linguistically and socioculturally situated problem rather than assuming that fairness constructs transfer unchanged across languages.

Second, this work extends the methodological scope from stand-alone bias detection to an unlearning-informed assessment-and-rewriting framework. Existing research in this domain has focused on detecting abusive language in classification models or learning fair representations for prediction tasks. This study addresses this gap by demonstrating that weight-level unlearning of an encoder-based assessment model can inform reduced-bias sentence generation while preserving its language-modeling ability (LMS = .7270 before and .7258 after unlearning). Notably, unlearning is applied only to the KcBERT assessment model, whereas generative rewriting is performed using an unmodified GPT-4o. This separation between unlearning-based assessment and unmodified generative rewriting distinguishes this work from detection-centric approaches and positions the modular assessment-and-rewriting framework as a safeguard for LLM-based content creation.

Third, the study empirically characterizes the relationship between fairness and meaning preservation in user evaluations of bias rewriting. Controlled text neutralization treats the removal of subjective bias and preservation of source meaning as joint objectives (Pryzant et al., 2020). The findings extend this formulation by demonstrating that users evaluated these objectives differently across AI-revision and single-expert-authored revision strategies. The AI revisions received higher fairness ratings, whereas the expert-authored revisions received higher meaning-preservation ratings, and neither strategy displayed a clear advantage in perceived bias improvement. This finding supports a theoretical division of labor in which computational systems generate and analyze revision candidates, and human judgment is necessary for resolving the context-dependent tension between fairness, semantic fidelity, and communicative intent.

8.2. Practical Implications

From a system design and application perspective, FAIRITHM offers three practical implications for bias-rewriting tools to support the production and review of public-facing text. First, bias-rewriting interfaces should support comparative and user-controlled revisions rather than present a single corrected sentence as a definitive answer. The AI revisions and single-expert-authored revisions demonstrated different strengths. The AI revisions received higher

fairness ratings, whereas the single-expert-authored revisions better preserved the meaning of the original sentences. Accordingly, systems should present multiple revision candidates and allow users to inspect the rationale for each revision and the extent to which it changes the original meaning. The system should enable them to select, edit, or combine alternatives according to the purpose and context of the communication. This approach shifts the role of the user from passively accepting an automatically generated output to judging which of the transparently presented revision alternatives is most appropriate.

Second, FAIRITHM is best positioned as a first-pass screening and revision-support tool within existing writing and review workflows rather than as a system that autonomously determines whether text is biased. The bias-assessment module can direct the attention of users toward potentially problematic sentences and expressions, whereas the sentence-generation module can provide alternative phrasings for review. In settings that process large volumes of public-facing text, including news organizations, educational institutions, content-review teams, and public agencies, this division of labor can focus human attention on expressions that require closer examination. Experts in the relevant domain should make the final judgment regarding fairness, meaning preservation, and contextual appropriateness. Moreover, the system's bias assessments and revision candidates should be treated as advisory information for review rather than as authoritative determinations.

Third, practical implementation should support calibrated reliance rather than aim to maximize user trust. Participants positively evaluated the usability, perceived appropriateness of bias judgments, clarity of revision rationales, and expected consistency of FAIRITHM, whereas their future trust in and intention to use the system remained neutral. Moreover, the AI revisions did not present a clear advantage over the single-expert-authored revisions in overall trustworthiness. Thus, bias-rewriting systems should allow users to inspect, question, and directly modify the output by providing editable revision suggestions, revision rationales, multiple alternatives, revision histories, and a human-approval step before external use. These features can support the responsible integration of the system into real-world workflows while preserving human responsibility for socially sensitive language decisions.

9. Limitations and Future Work

This study has several limitations that should be addressed in future research. First, the CABRS evaluation of the system output relied on a small and fixed set of six biased sentences. As each of the six predefined bias categories was represented by one selected sentence, the observed differences cannot be assumed to represent the full range of

expressions in each category. The mixed-effect analyses appropriately accounted for repeated ratings by the participants and identified variation across the selected sentences; however, the analyses did not resolve the limited sampling of evaluation material. Accordingly, the CABRS findings should be interpreted as patterns observed for the six evaluated sentences rather than as category-general conclusions about the comparative quality of AI-revision and single-expert-authored revision strategies.

Second, the participant pool was relatively small ($N = 34$) and skewed toward students, engineering majors, and adults under 40 (Section 5.1). Hence, this sample does not fully represent the general population or professional users who routinely assume responsibility for public-facing language, including journalists, editors, educators, public-sector communicators, and content moderators. Moreover, the participants interacted with FAIRITHM during a single study session, and the combined item assessing future trust and use intention was a single self-developed measure. The corresponding result should be understood as an initial postuse perception rather than as evidence of sustained trust, long-term adoption, or actual use in professional settings.

Third, the single-expert-authored revisions included in the main CABRS comparison were produced by a single expert. Although highly qualified, this individual instantiated one professionally grounded revision strategy and does not constitute a gold standard or an exhaustive representation of expert judgment.

The category-level directional analysis in Section 4.2.2 empirically reinforces this point. For symmetric categories (e.g., political orientation and gender identity), the bias-signal direction was significantly inverted, implying that the direction of a ‘fair revision’ is perspective-dependent in these categories and that different experts or community reviewers could produce systematically different references. Consistent with this interpretation, five professionals from different fields independently revised the same six stimuli without AI assistance and produced meaningfully different revision strategies (Appendix 9; Tables A9.1 and A9.2).

Finally, the bias indicators presented to the users were derived from internal model signals and may not precisely align with the bias intensity perceived by humans. The indicator was validated against labeled data and was a weak-to-moderate advisory signal with substantial variation across bias categories. In the annotation sample, 83.3% of flagged sentences were judged to be biased by a majority of annotators, and the top five highlighted words included at least one majority-marked word in 80.5% of the relevant sentences. However, the system still missed many sentences perceived as biased, and alignment with graded bias intensity has yet to be established.

To address these limitations, future research should expand CABRS to include multiple sentences per category and multiple expert- or community-authored revisions. Moreover, future research should examine repeated use with professional users in realistic writing and review settings and should validate sentence- and token-level model signals against human judgment of bias presence and intensity. These broader validation efforts should be accompanied by targeted refinements of the interface and detection strategy.

At the interface level, an ablation study comparing the presentation with the omission of token-level highlights requires a separately controlled user study, which is left for future work. In addition, the binary red/green flag employed in this study withholds weak detection signals that fall below the threshold of 0.2. Lowering the threshold is not a feasible remedy because, at 0.1, the false-positive rate rises from .183 to .505, risking alarm fatigue. Accordingly, future versions of the interface should introduce a three-level display system: red when the aggregated signal reaches the threshold of 0.2, an intermediate caution tier between 0.1 and 0.2, and green below 0.1. A reference implementation of this display is included in the code released with this paper. While retaining the red-flag threshold and detection rule used in the study, the proposed three-level display would highlight sentences with weak signals (e.g., those containing gender stereotypes), where probabilities only slightly shifted during the unlearning process, allowing users to review them. More fundamental remedies for uneven detection performance across categories, including a per-category threshold calibration and category-balanced gradient selection during unlearning, are crucial directions for future work.

10. Conclusion

This study presents FAIRITHM, an interactive Korean bias-rewriting system comprising a KcBERT bias-assessment module updated using PCGU-based unlearning and a GPT-4o-based rewriting module informed by the assessment results. The evaluation examined the performance of the bias-assessment model, its PLL-difference-based assessment signal, user evaluations of the AI and single-expert-authored revisions, the cognitive demands of manual bias rewriting, and the perceived usability of the system. At the model level, PCGU-based unlearning predominantly preserved language-modeling performance, whereas the PLL-difference signal provided statistically reliable directional information but was insufficiently accurate for autonomous classification, supporting its use as an advisory screening signal. The main findings and their implications are discussed below.

First, at the output level, the AI revisions received higher average ratings than the single-expert-authored revisions for perceived fairness and sentence completeness, although these differences varied across the six evaluated sentences. The participants rated the single-expert-authored revisions higher on meaning preservation. However, neither revision strategy revealed a clear advantage in perceived bias improvement relative to the original sentences. In the direct comparative choices, AI revisions were more often selected for bias mitigation and fairness and for appropriateness for use but did not display a clear advantage in overall trustworthiness. These findings indicate complementary strengths between AI-generated and single-expert-authored revisions, supporting the retention of human judgment when resolving context-dependent tensions between fairness and semantic fidelity.

Second, the manual bias-rewriting task elicited high ratings for mental demand and effort, and the participants perceived AI-assisted revision as a promising means of supporting this task. The absence of significant pre–post changes in the four common ABPS items indicates that this immediate practical value was distinct from the short-term changes in bias-related perception.

Third, FAIRITHM achieved acceptable perceived usability (SUS = 72.72), and the participants positively evaluated the perceived appropriateness of its bias judgment, the clarity of its revision rationale, and its expected consistency on other biased sentences. However, future trust in and intention to use the system was rated as neutral. These findings indicate caution rather than an unqualified willingness to rely on FAIRITHM and support calibrated reliance and human-supervised use rather than full automation. Overall, this study demonstrates the potential of unlearning-informed bias assessment for safer content creation while emphasizing that reliability, transparency, and human oversight are necessary for real-world deployment in information systems.

Author Contributions

Yunjin Nam: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing–original draft, Writing–review & editing, Visualization; **Donghui Lim:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing–original draft, Writing–review & editing, Visualization; **Sungkyun Im:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing–original draft, Writing–review & editing, Visualization; **Zuobin Xiong:** Conceptualization, Methodology, Software, Formal analysis, Writing–review & editing, Supervision; **Jeongeun Park:** Conceptualization,

Methodology, Formal analysis, Resources, Writing–original draft, Writing–review & editing, Supervision, Project administration, Funding acquisition

Financial Support

This work was supported by the research fund from Hanyang University (HY-2025-2753).

Declarations of Interest

The authors report no potential competing interests.

Data Availability

Data supporting the study findings are fully accessible through the following link: https://github.com/muse-hy/FAIRITHM.

References

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent Anti-Muslim Bias in Large Language Models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306. https://doi.org/10.1145/3461702.3462624

Albadi, N., Kurdi, M., & Mishra, S. (2018). Are they Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 69–76. https://doi.org/10.1109/ASONAM.2018.8508247

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine bias. In K. Martin (Ed.), *Ethics of data and analytics* (1st ed., pp. 254–265). Auerbach Publications

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI feedback*. arXiv. https://doi.org/10.48550/arXiv.2212.08073

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction*, *24*(6), 574–594. https://doi.org/10.1080/10447310802205776

Blodgett, S. L., Barocas, S., Hernandez, H. D., & Wallach, H. (2020). *Language (technology) is power: A critical survey of "bias" in NLP*. arXiv. https://doi.org/10.48550/arXiv.2005.14050

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., & Papernot, N. (2021). Machine Unlearning. *2021 IEEE Symposium on Security and Privacy (SP)*, 141–159. https://doi.org/10.1109/SP40001.2021.00019

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). Taylor & Francis

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Cao, N. D., Aziz, W., & Titov, I. (2021). *Editing factual knowledge in language models*. arXiv. https://doi.org/10.48550/arXiv.2104.08164

Chai, Y., Xie, H., & Qin, J. S. (2025). *Text data augmentation for large language models: A comprehensive survey of methods, challenges, and opportunities*. arXiv. https://doi.org/10.48550/arXiv.2501.18845

Chisca, A.-V., Rad, A.-C., & Lemnaru, C. (2024). Prompting fairness: Learning prompts for debiasing large language models. In B. R. Chakravarthi, B. B, P. Buitelaar, T. Durairaj, G. Kovács, & M. Á. García Cumbreras (Eds.), *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion* (pp. 52–62). Association for Computational Linguistics. https://aclanthology.org/2024.ltedi-1.6/

Cohausz, L., Tschalzev, A., Bartelt, C., & Stuckenschmidt, H. (2023). Investigating the importance of demographic features for EDM-predictions. In *Proceedings of the 16th International Conference on Educational Data Mining* (pp. 125–136). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.8115647

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, *50*(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Commun. ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622

Gray, M., Milanova, M., & Wu, L. (2024). Enhancing Bias Assessment for Complex Term Groups in Language Embedding Models: Quantitative Comparison of Methods. *JMIR Medical Informatics*, *12*(1), e60272. https://doi.org/10.2196/60272

Guo, W., & Caliskan, A. (2021). Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 122–133. https://doi.org/10.1145/3461702.3462536

Ha, S., Monadjemi, S., & Ottley, A. (2024). Guided By AI: Navigating Trust, Bias, and Data Exploration in AI-Guided Visual Analytics. *Computer Graphics Forum*, *43*(3), e15108. https://doi.org/10.1111/cgf.15108

Hegde, C., & Patil, S. (2020). *Unsupervised paraphrase generation using pre-trained language models*. arXiv. https://doi.org/10.48550/arXiv.2006.05477

Huang, P.-S., Zhang, H., Jiang, R., Stanforth, R., Welbl, J., Rae, J., Maini, V., Yogatama, D., & Kohli, P. (2020). *Reducing sentiment bias in language models via counterfactual evaluation*. arXiv. https://doi.org/10.48550/arXiv.1911.03064

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023). Co-Writing with Opinionated Language Models Affects Users' Views. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3544548.3581196

Jeong, Y., Oh, J., Ahn, J., Lee, J., Moon, J., Park, S., & Oh, A. (2022). *KOLD: Korean offensive language dataset*. arXiv. https://doi.org/10.48550/arXiv.2205.11315

Jin, J., Kim, J., Lee, N., Yoo, H., Oh, A., & Lee, H. (2024). KoBBQ: Korean Bias Benchmark for Question Answering. *Transactions of the Association for Computational Linguistics*, *12*, 507–524. https://doi.org/10.1162/tacl_a_00661

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2021). *The state and fate of linguistic diversity and inclusion in the NLP world*. arXiv. https://doi.org/10.48550/arXiv.2004.09095

Kabir, S., Li, L., & Zhang, T. (2024). STILE: Exploring and Debugging Social Biases in Pre-trained Text Representations. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–20. https://doi.org/10.1145/3613904.3642111

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Kaneko, M., & Bollegala, D. (2021). *Debiasing pre-trained contextualised embeddings*. arXiv. https://doi.org/10.48550/arXiv.2101.09523

Kauf, C., & Ivanova, A. (2023). *A better way to do masked language model scoring*. arXiv. https://doi.org/10.48550/arXiv.2305.10588

Kim, H. (2023). *Study on the selection and grading of basic Korean vocabulary in 2023*. National Institute of Korean Language. https://www.korean.go.kr/front/reportData/reportDataView.do?mn_id=45&searchOrder=&report_seq=1160&pageIndex=5

Kwon, B. C., & Mihindukulasooriya, N. (2022). An empirical study on pseudo-log-likelihood bias measures for masked language models using paraphrased sentences. In A. Verma, Y. Pruksachatkun, K.-W. Chang, A. Galstyan, J. Dhamala, & Y. T. Cao (Eds.), *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)* (pp. 74–79). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.trustnlp-1.7

Lee, J. (2020). *Beomi/KcBERT* [Computer software]. https://github.com/Beomi/KcBERT

Lee, J. K., & Chung, T.-M. (2025). Detecting Bias in Large Language Models: Fine-Tuned KcBERT. In C. Wallraven, C.-L. Liu, & A. Ross (Eds.), *Pattern Recognition and Artificial Intelligence* (pp. 76–90). Springer Nature. https://doi.org/10.1007/978-981-97-8705-0_6

Lee, S., Kim, D., Jung, D., Park, C., & Lim, H. S. (2024, June). Exploring inherent biases in LLMs within Korean social context: A comparative analysis of ChatGPT and GPT-4. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies (Volume 4: Student Research Workshop) (pp. 93-104). https://doi.org/10.18653/v1/2024.naacl-srw.11

Li, B., Hou, Y., & Che, W. (2022). Data augmentation approaches in natural language processing: A survey. *AI Open*, *3*, 71–90. https://doi.org/10.1016/j.aiopen.2022.03.001

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature Selection: A Data Perspective. *ACM Comput. Surv.*, *50*(6), 94:1-94:45. https://doi.org/10.1145/3136625

Liu, Y. (2024). Robust Evaluation Measures for Evaluating Social Biases in Masked Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(17), 18707–18715. https://doi.org/10.1609/aaai.v38i17.29834

Ma, X., Sap, M., Rashkin, H., & Choi, Y. (2020). *PowerTransformer: Unsupervised controllable revision for biased language correction*. arXiv. https://doi.org/10.48550/arXiv.2010.13816

Madanagopal, K., & Caverlee, J. (2023). Bias neutralization in non-parallel texts: A cyclic approach with auxiliary guidance. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14265–14278. Association for Computational Linguistics. http://doi.org/10.18653/v1/2023.emnlp-main.882.

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). *On measuring social biases in sentence encoders*. arXiv. https://doi.org/10.48550/arXiv.1903.10561

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, *54*(6), 115:1-115:35. https://doi.org/10.1145/3457607

Mitchell, E., Lin, C., Bosselut, A., Finn, C., & Manning, C. D. (2022). *Fast model editing at scale*. arXiv. https://doi.org/10.48550/arXiv.2110.11309

Moon, J., Cho, W. I., & Lee, J. (2020). *BEEP! Korean corpus of online news comments for toxic speech detection*. arXiv. https://doi.org/10.48550/arXiv.2005.12503

Nadăș, M., Dioșan, L., & Tomescu, A. (2025). Synthetic data generation using large language models: Advances in text and code. *IEEE Access, 13*, 134615–134633. https://doi.org/10.1109/ACCESS.2025.3589503

Nadeem, M., Bethke, A., & Reddy, S. (2020). *StereoSet: Measuring stereotypical bias in pretrained language models*. arXiv. https://doi.org/10.48550/arXiv.2004.09456

Nguyen, T. T., Huynh, T. T., Ren, Z., Nguyen, P. L., Liew, A. W.-C., Yin, H., & Nguyen, Q. V. H. (2024). *A survey of machine unlearning*. arXiv. https://doi.org/10.48550/arXiv.2209.02299

Nozza, D., Bianchi, F., & Hovy, D. (2021). HONEST: Measuring Hurtful Sentence Completion in Language Models. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2398–2406). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.191

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Padmaja, C. V. R., & Lakshminarayana, S. (2024). Enhancing language models through prompt engineering – A survey. In *2024 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT)* (pp. 117–121). IEEE. https://doi.org/10.1109/ICISSGT58904.2024.00033

Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., & Marco, F. (2020). Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 446–457). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372843

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., & Bowman, S. R. (2022). *BBQ: A hand built bias benchmark for question answering*. arXiv. https://doi.org/10.48550/arXiv.2110.08193

Pastaltzidis, I., Dimitriou, N., Quezada-Tavarez, K., Aidinlis, S., Marquenie, T., Gurzawska, A., & Tzovaras, D. (2022). Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2302–2314. https://doi.org/10.1145/3531146.3534644

Popović, R., Lemmerich, F., & Strohmaier, M. (2020). Joint Multiclass Debiasing of Word Embeddings. In D. Helic, G. Leitner, M. Stettinger, A. Felfernig, & Z. W. Raś (Eds.), *Foundations of Intelligent Systems* (pp. 79–89). Springer International Publishing. https://doi.org/10.1007/978-3-030-59491-6_8

Pryzant, R., Diehl Martinez, R., Dass, N., Kurohashi, S., Jurafsky, D., & Yang, D. (2020). Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(1), 480–489. https://doi.org/10.1609/aaai.v34i01.5385

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Rizwan, H., Shakeel, M. H., & Karim, A. (2020). Hate-Speech and Offensive Language Detection in Roman Urdu. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2512–2522). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.197

Roberts, S. T., Wood, S. E., & Eadon, Y. (2023). "We care about the internet; we care about everything": Understanding social media content moderators' mental models and support needs. In *Proceedings of the 56th Hawaii International Conference on System Sciences*. University of Hawaii at Mānoa. https://doi.org/10.24251/HICSS.2023.252

Salazar, J., Liang, D., Nguyen, T. Q., & Kirchhoff, K. (2020). Masked Language Model Scoring. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2699–2712. https://doi.org/10.18653/v1/2020.acl-main.240

Sauro, J. (2018, 09 19). 5 ways to interpret a SUS score. *MeasuringU*. https://measuringu.com/interpret-sus-score/

Schick, T., Udupa, S., & Schütze, H. (2021). Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, *9*, 1408–1424. https://doi.org/10.1162/tacl_a_00434

Sharma, N., Liao, Q. V., & Xiao, Z. (2024). Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3613904.3642459

Sharma, S., Zhang, Y., Ríos Aliaga, J. M., Bouneffouf, D., Muthusamy, V., & Varshney, K. R. (2020). Data Augmentation for Discrimination Prevention and Bias Disambiguation. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 358–364. https://doi.org/10.1145/3375627.3375865

Shrestha, R., Kafle, K., & Kanan, C. (2022). An investigation of critical issues in bias mitigation techniques. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 1943–1954). IEEE/CVF.

Spinde, T., Rudnitckaia, L., Mitrović, J., Hamborg, F., Granitzer, M., Gipp, B., & Donnay, K. (2021). Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management*, *58*(3), 102505. https://doi.org/10.1016/j.ipm.2021.102505

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research, 2023*(5), 1–95. https://openreview.net/pdf?id=uyTL5Bvosj

Streamlit Developers. (2021, January 14). *Streamlit* [Computer software]. https://streamlit.io/

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *54*(2), 119–135. https://doi.org/10.1016/j.erap.2003.12.004

Vejsbjerg, I., Daly, E. M., Nair, R., & Nizhnichenkov, S. (2024). Interactive Human-Centric Bias Mitigation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(21), 23838–23840. https://doi.org/10.1609/aaai.v38i21.30582

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, *27*(3), 425–478. JSTOR. https://doi.org/10.2307/30036540

Wang, Z., Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., & Russakovsky, O. (2020). Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)* (pp. 8919–8928). IEEE/CVF. https://doi.org/10.1109/CVPR42600.2020.00894

Wei, J., & Zou, K. (2019). *EDA: Easy data augmentation techniques for boosting performance on text classification tasks*. arXiv. https://doi.org/10.48550/arXiv.1901.11196

Xiao, Y., Liu, A., Liang, S., Liu, X., & Tao, D. (2025). Fairness Mediator: Neutralize Stereotype Associations to Mitigate Bias in Large Language Models. *Proc. ACM Softw. Eng.*, *2*(ISSTA), ISSTA012:250-ISSTA012:273. https://doi.org/10.1145/3728881

Yang, K., Yu, C., Fung, Y. R., Li, M., & Ji, H. (2023). ADEPT: A DEbiasing PrompT Framework. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(9), 10780–10788. https://doi.org/10.1609/aaai.v37i9.26279

Yao, J., Chien, E., Du, M., Niu, X., Wang, T., Cheng, Z., & Yue, X. (2024, August). Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 8403-8419). https://doi.org/ 10.18653/v1/2024.acl-long.457

Yu, C., Jeoung, S., Kasi, A., Yu, P., & Ji, H. (2023). Unlearning bias in language models by partitioning gradients. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 6032–6048). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.375

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340. https://doi.org/10.1145/3278721.3278779

Zhang, Y., Li, B., Ling, Z., & Zhou, F. (2024). Mitigating label bias in machine learning: Fairness through confident learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(15), 16917–16925. https://doi.org/10.1609/aaai.v38i15.29634

Zhang, Z., Ren, S., Liu, S., Wang, J., Chen, P., Li, M., Zhou, M., & Chen, E. (2018). *Style transfer as unsupervised machine translation*. arXiv. https://doi.org/10.48550/arXiv.1808.07894

Appendices

Appendix 1. GPT Prompt Guidelines.

GPT Prompt Guidelines
[Role] You are a bias mitigation expert. Revise the input sentence neutrally according to the selected debiasing level (Standard/High). **[Input]** • Original: {sentence} • Bias analysis results: {analysis_json} • Debiasing Level: {Standard|High} • Alternative Generation: {true|false} (if true, also produce an alternative version with different neutral phrasing while following the same rules) **[Common Rules]** 1. Preserve original meaning 2. Remove exaggeration/generalization/stereotypes/derogatory terms 3. Clarify ambiguity with context 4. Do not invent new facts/statistics 5. Respect sensitive attributes (race, gender, age, etc.) 6. No victim-blaming **[Debiasing Level Guidelines]** • Standard: preserve structure, minimal edits (alternative = small lexical/nuance changes) • High: prioritize bias removal, allow structural changes (alternative = different neutral structure/wording) **[Tone Guidelines]** • Use hedged/probabilistic language • Avoid categorical statements • Keep original style and language (formal/casual, Korean/English)

Appendix 2. Pre-Post Survey Items on AI Bias Perception Scale.

Code	Items	Scale	Measurement point
EXP	I have personally encountered biased/unfair expressions while using generative AI services.	Binary (Yes/No)	Pre only
BP1	Generative AI's biased expressions can have negative impacts on society.	7-point Likert	Pre & Post
BP2	A function to detect and mitigate bias in AI-generated text is necessary.	7-point Likert	Pre & Post
BP3	It is difficult for humans alone to identify all biases in AI-generated text.	7-point Likert	Pre & Post
BP4	Humans require considerable effort and time to revise biased text into a fair and neutral form.	7-point Likert	Pre & Post
BP5	AI can revise biased text more easily (compared to humans).	7-point Likert	Post only

Appendix 3. Survey Items for Usability Evaluation.

Category	Code	Statement	Scale
Post-Use Perception Items	T1	FAIRITHM provides an appropriate bias judgment (bias score).	7-point Likert
	T2	FAIRITHM offers clear reasons and explanations for sentence revision.	7-point Likert
	T3	FAIRITHM will consistently improve other biased sentences in the future.	7-point Likert
	T4	I am willing to trust and use FAIRITHM in the future.	7-point Likert
System Usability Scale (SUS)	S1	I think I would like to use FAIRITHM frequently.	5-point Likert
	S2	I found FAIRITHM unnecessarily complex.	5-point Likert
	S3	I thought FAIRITHM was easy to use.	5-point Likert
	S4	I believe I would need the assistance of an expert to utilize FAIRITHM effectively.	5-point Likert
	S5	I found that the functions of FAIRITHM were well-integrated.	5-point Likert
	S6	I thought there were too many inconsistencies in FAIRITHM.	5-point Likert
	S7	I would imagine that most people would learn to use FAIRITHM very quickly.	5-point Likert
	S8	I found FAIRITHM very cumbersome to use.	5-point Likert
	S9	I felt very confident using FAIRITHM.	5-point Likert
	S10	I needed to learn many things before I could get going with FAIRITHM.	5-point Likert

Appendix 4A. Descriptive Statistics for Comparative Assessment of Bias-Reduced Sentences Composite Outcomes by Sentence Condition or Revision Source and Category-Specific Sentence., *M* (*SD*).

Outcome	Condition	Overall	Family	Religion	SES	Physical	Gender	Political
Completeness	Original	6.02 (0.99)	5.93 (1.19)	6.00 (1.06)	5.99 (0.95)	5.99 (0.99)	5.97 (1.01)	6.24 (0.70)
	Single expert	5.72 (1.24)	4.82 (1.77)	5.38 (1.30)	6.10 (0.78)	6.12 (0.89)	5.65 (1.14)	6.26 (0.65)
	AI	6.00 (0.93)	5.71 (1.19)	6.18 (0.65)	6.16 (0.65)	6.18 (0.92)	6.16 (0.68)	5.62 (1.19)
Fairness	Original	1.91 (1.12)	1.84 (1.01)	1.55 (0.76)	2.65 (1.48)	1.41 (0.61)	1.94 (1.16)	2.05 (1.14)
	Single expert	4.58 (1.60)	4.33 (1.52)	4.62 (1.48)	4.87 (1.32)	4.10 (1.88)	4.74 (1.67)	4.81 (1.64)
	AI	5.46 (1.36)	5.12 (1.43)	5.95 (1.06)	5.10 (1.56)	5.71 (1.42)	5.40 (1.29)	5.46 (1.23)
Meaning preservation	Single expert	4.75 (1.72)	5.75 (0.96)	4.85 (1.63)	5.21 (1.38)	4.15 (1.96)	3.75 (1.84)	4.78 (1.65)
	AI	4.01 (1.96)	4.29 (1.82)	3.19 (1.90)	5.13 (1.64)	3.53 (2.05)	4.26 (1.88)	3.65 (1.94)
Perceived bias improvement	Single expert	4.84 (1.56)	4.82 (1.40)	4.94 (1.76)	5.00 (1.46)	4.47 (1.60)	4.79 (1.53)	5.00 (1.61)
	AI	5.06 (1.63)	5.26 (1.29)	5.35 (1.54)	4.79 (1.92)	4.85 (1.69)	5.50 (1.54)	4.62 (1.69)

Note. Values are raw descriptive statistics presented as *M* (*SD*). The Overall column pools ratings across the six category-specific sentences.

Appendix 4B. Item-Level Descriptive Statistics for Parts 1 and 2 of the Comparative Assessment of Bias-Reduced Sentences by Sentence Condition or Revision Source and Category-Specific Sentence, *M* (*SD*).

Part	Code	Overall	Family	Religion	SES	Physical	Gender	Political
Part 1	O1	6.09 (1.02)	5.94 (1.25)	5.94 (1.32)	6.15 (0.86)	6.12 (0.95)	6.06 (0.98)	6.32 (0.64)
	O2	5.95 (1.12)	5.91 (1.31)	6.06 (1.07)	5.82 (1.17)	5.85 (1.18)	5.88 (1.15)	6.15 (0.82)
	O3	1.88 (1.27)	1.74 (1.05)	1.41 (0.66)	2.76 (1.79)	1.41 (0.66)	2.03 (1.42)	1.91 (1.22)
	O4	1.77 (1.29)	1.91 (1.85)	1.41 (1.08)	2.38 (1.50)	1.35 (0.54)	1.76 (1.10)	1.82 (1.11)
	O5	2.07 (1.41)	1.88 (1.15)	1.82 (1.36)	2.79 (1.70)	1.47 (0.75)	2.03 (1.55)	2.41 (1.42)
Part 2 (Single expert)	R1	5.86 (1.20)	5.06 (1.79)	5.62 (1.23)	6.12 (0.84)	6.18 (0.83)	5.88 (1.09)	6.29 (0.63)
	R2	5.59 (1.48)	4.59 (1.97)	5.15 (1.69)	6.09 (0.79)	6.06 (1.01)	5.41 (1.50)	6.24 (0.78)
	R3	4.71 (1.72)	4.53 (1.64)	4.74 (1.66)	4.94 (1.48)	4.09 (2.01)	4.97 (1.68)	5.00 (1.76)
	R4	4.32 (1.81)	4.00 (1.65)	4.47 (1.71)	4.65 (1.55)	4.03 (2.11)	4.59 (1.83)	4.21 (1.95)
	R5	4.70 (1.69)	4.47 (1.69)	4.65 (1.63)	5.03 (1.42)	4.18 (1.80)	4.65 (1.87)	5.24 (1.62)
	R6	4.75 (1.80)	5.82 (0.94)	4.68 (1.82)	5.21 (1.43)	4.18 (2.02)	3.74 (2.02)	4.91 (1.64)
	R7	4.74 (1.79)	5.68 (1.17)	5.03 (1.59)	5.21 (1.51)	4.12 (2.07)	3.76 (1.88)	4.65 (1.79)
	R8	4.84 (1.56)	4.82 (1.40)	4.94 (1.76)	5.00 (1.46)	4.47 (1.60)	4.79 (1.53)	5.00 (1.61)
Part 2 (AI)	R1	6.12 (0.89)	5.85 (1.23)	6.26 (0.67)	6.29 (0.52)	6.24 (0.78)	6.29 (0.63)	5.79 (1.15)
	R2	5.88 (1.15)	5.56 (1.33)	6.09 (0.83)	6.03 (0.97)	6.12 (1.09)	6.03 (1.03)	5.44 (1.42)
	R3	5.58 (1.42)	5.26 (1.66)	5.94 (1.23)	5.24 (1.65)	5.71 (1.43)	5.50 (1.33)	5.82 (1.11)
	R4	5.26 (1.57)	4.71 (1.71)	5.82 (1.24)	4.85 (1.71)	5.68 (1.47)	5.41 (1.35)	5.09 (1.62)
	R5	5.53 (1.39)	5.38 (1.44)	6.09 (1.06)	5.21 (1.53)	5.74 (1.40)	5.29 (1.51)	5.47 (1.26)
	R6	4.01 (2.07)	4.35 (1.97)	3.12 (2.04)	5.15 (1.69)	3.59 (2.18)	4.29 (1.93)	3.56 (2.03)
	R7	4.01 (2.00)	4.24 (1.79)	3.26 (1.99)	5.12 (1.70)	3.47 (2.11)	4.24 (1.91)	3.74 (2.03)
	R8	5.06 (1.63)	5.26 (1.29)	5.35 (1.54)	4.79 (1.92)	4.85 (1.69)	5.50 (1.54)	4.62 (1.69)

Appendix 5. Category-Sentence Simple Contrasts from the Comparative Assessment of Bias-Reduced Sentences Mixed-Effects Models.

Appendix 5A. Sentence Completeness.

Category sentence	Contrast	Estimate	Unadjusted 95% CI	Holm-adjusted *p*
Family	AI − Single expert	0.882	[0.464, 1.300]	< .001
	AI − Original	-0.221	[-0.639, 0.198]	1.000
	Single expert − Original	-1.103	[-1.521, -0.685]	< .001
Religion	AI − Single expert	0.794	[0.376, 1.212]	0.001
	AI − Original	0.176	[-0.242, 0.595]	1.000
	Single expert − Original	-0.618	[-1.036, -0.200]	0.019
SES	AI − Single expert	0.059	[-0.359, 0.477]	1.000
	AI − Original	0.176	[-0.242, 0.595]	1.000
	Single expert − Original	0.118	[-0.300, 0.536]	1.000

Physical	AI − Single expert	0.059	[-0.359, 0.477]	1.000
	AI − Original	0.191	[-0.227, 0.609]	1.000
	Single expert − Original	0.132	[-0.286, 0.550]	1.000
Gender	AI − Single expert	0.515	[0.097, 0.933]	0.048
	AI − Original	0.191	[-0.227, 0.609]	1.000
	Single expert − Original	-0.324	[-0.742, 0.095]	0.516
Political	AI − Single expert	-0.647	[-1.065, -0.229]	0.010
	AI − Original	-0.618	[-1.036, -0.200]	0.023
	Single expert − Original	0.029	[-0.389, 0.448]	1.000

Appendix 5B. Sentence Fairness.

Category sentence	Contrast	Estimate	Unadjusted 95% CI	Holm-adjusted *p*
Family	AI − Single expert	0.784	[0.176, 1.393]	0.047
	AI − Original	3.275	[2.666, 3.883]	< .001
	Single expert − Original	2.49	[1.882, 3.099]	< .001
Religion	AI − Single expert	1.333	[0.725, 1.942]	< .001
	AI − Original	4.402	[3.793, 5.010]	< .001
	Single expert − Original	3.069	[2.460, 3.677]	< .001
SES	AI − Single expert	0.225	[-0.383, 0.834]	0.467
	AI − Original	2.451	[1.842, 3.059]	< .001
	Single expert − Original	2.225	[1.617, 2.834]	< .001
Physical	AI − Single expert	1.608	[0.999, 2.216]	< .001
	AI − Original	4.294	[3.686, 4.903]	< .001
	Single expert − Original	2.686	[2.078, 3.295]	< .001
Gender	AI − Single expert	0.667	[0.058, 1.275]	0.095
	AI − Original	3.461	[2.852, 4.069]	< .001
	Single expert − Original	2.794	[2.186, 3.403]	< .001
Political	AI − Single expert	0.647	[0.039, 1.256]	0.095
	AI − Original	3.412	[2.803, 4.020]	< .001
	Single expert − Original	2.765	[2.156, 3.373]	< .001

Appendix 5C. Meaning Preservation.

Category sentence	Contrast	Estimate	Unadjusted 95% CI	Holm-adjusted *p*
Family	Single expert − AI	1.456	[0.804, 2.107]	< .001
Religion	Single expert − AI	1.662	[1.010, 2.313]	< .001
SES	Single expert − AI	0.074	[-0.578, 0.725]	0.824
Physical	Single expert − AI	0.618	[-0.034, 1.269]	0.189
Gender	Single expert − AI	-0.515	[-1.166, 0.137]	0.242
Political	Single expert − AI	1.132	[0.481, 1.784]	0.003

Appendix 6. Observed Part 3 Choices in the Comparative Assessment of Bias-Reduced Sentences by Category-Specific Sentence and Evaluation Criterion, *n* (%).

Category sentence	Criterion	AI n (%)	No difference n (%)	Single expert n (%)
Family	Fairness	21 (61.8)	1 (2.9)	12 (35.3)
	Appropriateness	19 (55.9)	3 (8.8)	12 (35.3)
	Trustworthiness	16 (47.1)	3 (8.8)	15 (44.1)
Religion	Fairness	20 (58.8)	5 (14.7)	9 (26.5)
	Appropriateness	16 (47.1)	8 (23.5)	10 (29.4)
	Trustworthiness	14 (41.2)	4 (11.8)	16 (47.1)
SES	Fairness	18 (52.9)	5 (14.7)	11 (32.4)
	Appropriateness	18 (52.9)	6 (17.6)	10 (29.4)
	Trustworthiness	18 (52.9)	5 (14.7)	11 (32.4)
Physical	Fairness	22 (64.7)	3 (8.8)	9 (26.5)
	Appropriateness	23 (67.6)	4 (11.8)	7 (20.6)
	Trustworthiness	19 (55.9)	6 (17.6)	9 (26.5)
Gender	Fairness	21 (61.8)	6 (17.6)	7 (20.6)
	Appropriateness	20 (58.8)	7 (20.6)	7 (20.6)
	Trustworthiness	20 (58.8)	8 (23.5)	6 (17.6)
Political	Fairness	19 (55.9)	3 (8.8)	12 (35.3)
	Appropriateness	14 (41.2)	5 (14.7)	15 (44.1)
	Trustworthiness	17 (50.0)	4 (11.8)	13 (38.2)

Appendix 7. Participant-Level Sensitivity Analyses for the Comparative Assessment of Bias-Reduced Sentences Results.

Appendix 7A. Participant-Level Sensitivity Analyses for Overall Common-Item Contrasts.

Outcome	Contrast	HL shift	Holm-adjusted *p*
Completeness	AI − Single expert	0.333	0.006
	AI − Original	−0.051	0.674
	Single expert − Original	−0.375	0.034
Fairness	AI − Single expert	0.889	< .001
	AI − Original	3.583	< .001
	Single expert − Original	2.667	< .001

Appendix 7B. Participant-Level Preference-Index Sensitivity Analysis for Part 3 of the Comparative Assessment of Bias-Reduced Sentences.

Criterion	HL pseudomedian [95% CI]	Effect size *r*	Holm-adjusted *p*
Fairness	0.333 [0.167, 0.583]	0.567	0.003
Appropriateness	0.250 [0.000, 0.500]	0.398	0.041
Trustworthiness	0.167 [0.000, 0.333]	0.321	0.062

Note. Positive values indicate preference for the AI revision. Confidence intervals are unadjusted 95% Hodges–Lehmann intervals; *p*-values were Holm-adjusted across the three criteria.

Appendix 8. Implementation and Reproducibility Details

Dataset composition. The 12,720 sentences comprise 4,240 triplets, each consisting of one biased, one counter-stereotypical, and one unrelated sentence. The bundle-level split preserves an 8:2 ratio with fully disjoint templates between splits: 3,392 triplets (10,176 sentences) for training and 848 triplets (2,544 sentences) for testing. Equivalently, the respective splits contain 3,392 and 848 sentences per role. The per-category totals are unchanged from Table 1.

Bias-assessment models. The baseline model is KcBERT-base (~110 M parameters; beomi/kcbert-base). The retrained model applies PCGU to the baseline in the following configuration: batch size = 16, learning rate = 1e-5, $k = 10{,}000$ updated gradient dimensions with dynamic-k selection, and symmetric pairwise gradients. Training employed the 10,176-sentence training split described in Section 3.2.

Bias-assessment and flagging rule. For an input sentence, each token is masked, and the log-probability of the original token is computed under both models (pseudo-log-likelihood [PLL] scoring). The per-token signal is s_i = max(logP_baseline − logP_retrained, 0). Subword scores are merged into words by summation to produce the top five identified words. The sentence-level signal is the mean of the top-k token signals, with k = max(3, ⌈0.2 × number of positive-signal tokens⌉). Sentences whose aggregated signal reaches the threshold of 0.2 receive a red flag; otherwise, a green flag is presented. The 0 to 10 score displayed in the interface is min(signal/0.3, 1) × 10. The display constant 0.3 corresponds to about the 95th percentile of the aggregated signal on neutral sentences and does not affect the flag decision.

Rewriting module. Bias-reduced revision candidates were generated using GPT-4o via the OpenAI API with temperature = 0.3, a maximum output of 500 tokens, and JSON-constrained responses, using the prompt template in Appendix 1. The prompt receives the original sentence, bias-assessment result (flag and top highlighted words), user-selected debiasing level (standard/high), and alternative options.

Example (Korean input with English translation). Input: "어린이는 공공장소 예절이 부족하다." ("Children lack manners in public places.") Output at the standard level: "어린이는 공공장소에서 예절을 배워가는 과정에 있을 수 있다." ("Children may be in the process of learning manners in public places.")

Software environment. The system was developed and evaluated using Torch (v. 1.7.1, CUDA 11.0) and transformers (v. 4.15.0). The additional validation analyses were executed using Torch (v. 2.3.1, CUDA 12.1) and transformers (v. 4.55.4). The StereoSet-style evaluation used *kiwipiepy* for morpheme-based masking. The stereotype score (SS), language modeling score (LMS), and idealized context association test (ICAT) values vary by about ±0.005 across these library versions. Exact package versions are provided in the released requirement file.

Availability. The reconstructed dataset (training/testing splits with bundle identifiers), prompt templates, PCGU training configuration, and all evaluation materials are available at https://github.com/muse-hy/FAIRITHM. The repository provides executable implementations of all procedures specified in this appendix.

Appendix 9. Supplementary Expert-Revision Exercise

To examine the variation in expert revisions of the same material, five professionals from diverse fields (educational psychology, Korean language education, social policy, architecture/education, and English literature/gender discourse; anonymized as E1 to E5, respectively) independently revised the six user-study stimuli. The professionals worked without AI assistance, under the instructions to revise each sentence toward neutrality without substantially altering its form while preserving its core meaning.

The revision strategies diverged into five distinct types (Tables A9.1 and A9.2): negation rebuttal (quoting the claim and denying it), subject generalization (removing the group mention), quotative distancing (recasting the claim as reported speech), quantifier insertion (restricting scope with "some"), and reframing (recasting the proposition in a different causal structure). Surface overlaps with the original sentence (character-bigram Jaccard with whitespace removed) ranged from .436 to .848 across revisers, a nearly two-fold difference in revision depth. The mean pairwise overlap between revisers was .445 (range .278–.691), and the overlap with the rewrites by the system averaged .319. Two revisers judged one stimulus (a meta-statement that a stereotype "is widespread") to require no revision. These figures describe the surface form only, not the agreement on bias judgment. The results indicate that professional revisions can differ substantially in strategy and revision depth even when applied to the same source material.

Table A9.1. Independent revisions of the six user-study stimuli by five professionals: Strategy typology and surface similarity.

Expert (field)	Dominant strategy	Overlap w/ original	Overlap w/ other experts (mean pairwise)	Overlap w/ system rewrite
E1 Educational psychology	Negation rebuttal ("it cannot be said that …")	.550	.408	.326
E2 Korean language education	Subject generalization (group mention removed)	.525	.432	.265
E3 Social policy	Quotative distancing ("there is a claim that …")	.848	.522	.398
E4 Architecture/education	Quantifier insertion ("some …")	.780	.519	.316
E5 English literature/gender discourse	Reframing (different causal structure)	.436	.343	.291

Note. Surface similarity was calculated using character-bigram Jaccard similarity with whitespace removed and averaged across the six stimuli. Higher overlap with the original indicates a shallower revision. Overall inter-expert pairwise overlap = .445 (range .278 to .691); experts-vs.-system overall mean = .319. E4 and E5 each judged one stimulus (a meta-statement that a stereotype "is widespread") to require no revision. Similarity figures describe the surface form only, not the agreement on bias judgment.

Table A9.2. Excerpt of the five professionals' independent revisions for one user-study stimulus.

Reviser (strategy)	Sentence
Original (biased)	"레즈비언 부부는 엄마와 아빠가 함께 있는 가정이 아니어서 아이의 성장과 발달에 심각한 결핍을 초래한다." ("Lesbian couples do not provide a home with both mother and father, causing serious deficiencies in children's growth and development.")
E1: Negation rebuttal	"레즈비언 부부, 엄마와 아빠가 함께 있는 가정이 아니라고 하여 아이의 성장과 발달에 심각한 정서적 결핍을 초래한다고 할 수 없다." ("It cannot be said that lesbian couples, a home without both a mother and a father, cause serious emotional deficiencies in children's growth and development.")
E2: Subject generalization	"엄마와 아빠의 가정 내 역할이 균형을 잃으면 아이의 성장과 발달에 심각한 결핍을 초래할 수 있다." ("When the roles of mother and father within a family lose balance, serious deficiencies in children's growth and development may result.")
E3: Quotative distancing	"레즈비언 부부는 엄마와 아빠가 함께 있는 가정이 아니므로 아이의 성장과 발달에 심각한 결핍을 초래한다는 주장이 있다." ("There is a claim that lesbian couples cause serious deficiencies in children's growth and development because theirs is not a home with both a mother and a father.")
E4: Qualification/hedging	"레즈비언 부부는 엄마와 아빠가 함께 있는 가정이 아니어서 아이의 성장과 발달에 심각한 결핍을 초래한다는 오해를 받는다." ("Lesbian couples face the misconception that they cause serious deficiencies in children's growth and development because theirs is not a home with both a mother and a father.")
E5: Reframing	"가정 내에 건강한 양육자의 역할이 부재할 때, 아이의 성장과 발달에 심각한 결핍이 발생할 가능성이 높다." ("When the role of a healthy caregiver is absent in a family, serious deficiencies in children's growth and development are likely to arise.")

Note. The excerpt presents the lesbian-couple stimulus. Responses are reproduced verbatim, including typographical errors, and English glosses are provided in parentheses. E2 and E5 removed the group mention but preserved the underlying normative frame (parental-role balance/absence of a "healthy caregiver"), illustrating how bias can be reintroduced in an altered form.