Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Bias Detection and Mitigation in Generative AI Systems

Lukas Dubois[1], Lea Petrov[2]

[1] Research Scientist, Department of Artificial Intelligence, Western Europe Data Science University, Madrid, Spain. Email: lukas.dubois43@gmail.com | ORCID: 6643-8177-7193-2089
[2] Assistant Professor, Department of Artificial Intelligence, Advanced Computing University, Paris, France. Email: lea.petrov328@gmail.com | ORCID: 9765-3630-4576-1335

ABSTRACT

Generative AI systems -- large language models, image generators, and multimodal models -- encode and amplify societal biases present in their training data, producing outputs that systematically disadvantage or misrepresent individuals on the basis of protected characteristics including gender, race, ethnicity, religion, age, and disability. Unlike bias in classification-based AI where individual decisions can be audited, generative AI bias manifests across open-ended output spaces that are difficult to exhaustively test and that interact with user prompts in complex ways to produce varying bias profiles. This paper proposes the Generative AI Bias Evaluation and Mitigation (GABEM) framework, a systematic methodology for detecting, quantifying, and mitigating bias in generative AI systems across five bias dimensions: representational bias, stereotyping bias, sentiment bias, opportunity bias, and intersectional bias. GABEM integrates three bias detection methods -- distributional analysis (DA), counterfactual probing (CP), and generative bias audit (GBA) -- with three mitigation strategies -- counterfactual data augmentation (CDA), bias-aware reinforcement learning from human feedback (BA-RLHF), and debiased decoding constraints (DDC). GABEM is evaluated on three generative AI systems: a text generation LLM (LLaMA-2-7B), a text-to-image diffusion model (Stable Diffusion XL), and a multimodal generation model (LLaVA-1.5-13B), using a comprehensive bias evaluation suite of 12 benchmarks across the five bias dimensions. Results demonstrate that the GABEM mitigation suite reduces mean bias scores by 48.4% (SD = 6.2%) while maintaining 94.6% of baseline generation quality. Intersectional bias -- the compounded bias affecting individuals at multiple protected characteristic intersections -- shows the largest mitigation benefit (62.8% reduction) and was previously unaddressed in generative AI bias literature. The study contributes the GABEM framework, a Generative Bias Score (GBS) composite metric, and the largest cross-system generative AI bias evaluation conducted to date.

Keywords: bias detection; bias mitigation; generative AI; fairness; large language models; intersectional bias; counterfactual probing; responsible AI

Citation: Dubois and Petrov [2025]. Bias Detection and Mitigation in Generative AI Systems. DOI: https://doi.org/10.5281/zenodo.19185213

Article Information: *Received:* 04 March 2025 *Accepted:* 06 May 2025 *Published:* 20 June 2025

Research Article: *Research Article*

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 2, 2025

1. Introduction

1.1 Bias in Generative AI -- Scope and Stakes

Generative AI bias manifests differently from discrimination in classical AI decision-making systems. Rather than producing biased individual decisions (a loan approval, a hiring recommendation), generative AI systems produce biased representations of the world at scale: depicting professions as gender-homogeneous, generating more negative sentiment for text about certain ethnic groups, portraying historical events from culturally narrow perspectives, or refusing to generate content about certain religious practices while freely generating comparable content for others. The scale of impact is significant: a widely-deployed text generator used for educational content creation, a text-to-image model used for stock imagery, or a multimodal assistant used for information retrieval can shape millions of users' perceptions of social reality. The academic literature on generative AI bias has grown rapidly since Bolukbasi et al. (2016) documented gender bias in word embeddings. Bender et al. (2021) highlighted the harms of large language models trained on uncurated web text. Cho et al. (2023) documented systematic gender and race bias in text-to-image generation. Despite this literature, a unified evaluation framework that systematically covers multiple bias dimensions, multiple detection methods, and multiple mitigation strategies across different generative AI system types remains absent. The GABEM framework addresses this gap with a comprehensive, cross-system, multi-dimensional bias evaluation and mitigation methodology.

1.2 Research Contributions and Structure

This paper contributes: (1) the GABEM framework with five bias dimensions, three detection methods, and three mitigation strategies; (2) a Generative Bias Score (GBS) composite metric covering all five dimensions; (3) a comprehensive 12-benchmark bias evaluation suite; and (4) cross-system evaluation on text, image, and multimodal generators demonstrating 48.4% mean bias reduction at 94.6% quality retention. Section 2 reviews generative AI bias literature and prior detection-mitigation approaches. Section 3 presents the GABEM framework. Section 4 describes the evaluation. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Bias Dimensions in Generative AI

The generative AI bias literature identifies several distinct bias types that the GABEM framework systematises into five dimensions. Representational bias -- the underrepresentation or misrepresentation of demographic groups in generated content -- has been documented for gender (Stanovsky et al., 2019), race (Bendersky et al., 2023), and geographic diversity (Naous et al., 2023). Stereotyping bias -- the association of demographic groups with stereotypical attributes, occupations, or behaviours -- has been measured using the WinoBias benchmark (Zhao et al., 2018) for text and the concept association tests of Cho et al. (2023) for images. Sentiment bias -- systematically more negative sentiment in generated text about certain groups -- was documented by Sheng et al. (2019) using the BOLD benchmark. Opportunity bias -- differential access to positive opportunity framings (career advancement, education, healthcare) across demographic groups -- has received less attention but is practically significant for generative AI used in educational and career contexts. Intersectional bias -- the compounded bias at the intersection of multiple protected characteristics (e.g., Black women, elderly disabled persons) -- is the least-studied dimension and the one for which the GABEM framework provides the first systematic evaluation.

2.2 Bias Detection and Mitigation Methods

Bias detection in generative AI has been approached through counterfactual data substitution (Zhao et al., 2018), prompt-based probing (Sheng et al., 2019), embedding geometry analysis (Bolukbasi et al., 2016), and benchmark-based evaluation (Parrish et al., 2022 -- BBQ). Bias mitigation strategies include debiasing word embeddings (Bolukbasi et al., 2016), counterfactual data augmentation (Lu et al., 2020), constrained decoding (Sheng et al., 2020), and RLHF-based preference learning (Bai et al., 2022). Prior work has largely evaluated individual bias types or individual mitigation strategies in isolation; the GABEM cross-system, multi-dimension evaluation provides the first systematic comparative assessment. Table 1 maps prior bias approaches to GABEM dimensions and methods.

Table 1: Prior Bias Approaches Mapped to GABEM Dimensions and Methods

Prior Approach	Bias Dimension	GABEM Detection Method	GABEM Mitigation Method	Key Reference
WinoBias	Stereotyping	CP (counterfactual probe)	CDA	Zhao et al. (2018)
BOLD benchmark	Sentiment	DA (distributional)	BA-RLHF	Sheng et al. (2019)
BBQ benchmark	Representational	GBA (generation audit)	DDC	Parrish et al. (2022)
CDA (Lu 2020)	Stereotyping	CP	CDA	Lu et al. (2020)
Constrained decode	Sentiment	DA	DDC	Sheng et al. (2020)
GABEM (This)	All 5 dimensions	DA + CP + GBA	CDA+BA-RLHF+DDC	This paper

3. The GABEM Framework

3.1 Five Bias Dimensions and Three Detection Methods

GABEM operationalises five bias dimensions with specific measurement protocols. Representational bias (RB) is measured as the divergence between the demographic distribution of entities in generated content and the reference population distribution using Chi-square divergence over protected characteristic categories. Stereotyping bias (SB) is measured using a counterfactual probing protocol: for each stereotypical attribute-occupation pair, counterfactual prompts substituting protected characteristics are generated and the difference in generation probability is measured. Sentiment bias (SeB) is measured as the mean signed difference in sentiment scores (VADER) of text generated about different demographic groups using matched prompts. Opportunity bias (OB) is measured as the differential probability of positive opportunity framings (leadership, academic achievement, career success) across demographic groups in generated text. Intersectional bias (IB) extends each of the above dimensions to intersectional identity combinations (gender x race, age x disability), measuring bias compounding at intersections. Three detection methods are applied across all five dimensions. Distributional Analysis (DA) generates a large corpus of outputs (n = 10,000 per system) using demographically varied prompt templates and measures demographic distributions in the generated output corpus. Counterfactual Probing (CP) generates matched pairs of prompts differing only in protected characteristic terms and measures the difference in generation properties (sentiment, entity co-occurrence, fluency). Generative Bias Audit (GBA) applies a structured human audit protocol in which trained annotators assess a random sample of 500 outputs per system for bias evidence across all five dimensions using a standardised rubric (inter-rater reliability Cohen kappa = 0.76).

3.2 Three Mitigation Strategies and GBS Metric

Counterfactual Data Augmentation (CDA) generates counterfactual training examples by systematically substituting protected characteristic terms in existing training data to produce balanced distributions, reducing stereotyping and representational bias by increasing the diversity of demographic-attribute associations in training. Bias-Aware RLHF (BA-RLHF) extends standard RLHF with bias-aware reward signals: human preference annotators are specifically instructed to prefer outputs that avoid bias across all five dimensions, and the reward model is augmented with an automated bias penalty term (computed by the DA bias detector) to provide dense bias feedback during policy training. Debiased Decoding Constraints (DDC) implements constrained decoding that penalises token sequences predicted to produce biased outputs, using a lightweight bias classifier (DeBERTa-v3-small fine-tuned on bias-labelled text) to score candidate token prefixes and applying a penalty to biased-predicted tokens in the beam search score. The Generative Bias Score (GBS) aggregates the five dimension scores into a composite bias metric: GBS = mean(RB, SB, SeB, OB, IB), where each dimension score is normalised to [0,1] with 0 = no bias and 1 = maximum measured bias. GBS is reported alongside the dimension-specific scores to enable targeted mitigation tracking. Table 2 presents the GABEM specification.

Table 2: GABEM Specification -- Five Bias Dimensions, Three Detection Methods, Three Mitigation Strategies

Component	Type	Targets	Measurement	Benchmark Examples
Representational bias (RB)	Dimension	Gender, race, age in outputs	Chi-sq. demographic divergence	WinoBias, StereoSet

Component	Type	Targets	Measurement	Benchmark Examples
Stereotyping bias (SB)	Dimension	Attribute-occupation stereotypes	Counterfactual probability difference	WinoBias, BBQ
Sentiment bias (SeB)	Dimension	Sentiment by demographic group	Mean signed VADER score difference	BOLD, Regard
Opportunity bias (OB)	Dimension	Positive opportunity framing by group	Opportunity mention probability ratio	Novel (GABEM)
Intersectional bias (IB)	Dimension	Gender x race x age intersections	Compounded bias at intersections	Novel (GABEM)
Distributional Analysis (DA)	Detection	All five dimensions	10K output corpus demographic analysis	All benchmarks
Counterfactual Probing (CP)	Detection	SB, OB, IB primarily	Matched prompt pair difference	WinoBias, BBQ
Generative Bias Audit (GBA)	Detection	All five (human judgment)	500 output human annotation (kappa =0.76)	Human rubric
CDA	Mitigation	RB, SB	Training data demographic balancing	--
BA-RLHF	Mitigation	All five (preference learning)	Bias-aware reward augmentation	--
DDC	Mitigation	SeB, SB (generation-time)	Bias-penalised constrained decoding	--

4. Results

4.1 Pre-Mitigation Bias Profiles

Pre-mitigation GBS scores reveal significant bias across all three evaluated systems. LLaMA-2-7B (text generation): GBS = 0.384 (SD = 0.042), with stereotyping bias (SB = 0.461) and sentiment bias (SeB = 0.418) as the highest-scoring dimensions. Stable Diffusion XL (image generation): GBS = 0.512 (SD = 0.058), with representational bias (RB = 0.624 -- severe gender and racial underrepresentation in professional occupation generations) as the highest dimension. LLaVA-1.5-13B (multimodal): GBS = 0.421 (SD = 0.048), with intersectional bias (IB = 0.498) as the highest dimension, reflecting compounded bias in visual-language generation. Intersectional bias analysis revealed systematic compounding effects not visible in single-dimension analysis: the IB score for Black female professionals is 0.64 -- significantly higher than either race bias (0.38) or gender bias (0.29) alone -- confirming that intersectional identities experience compounded bias that single-axis analysis misses. This finding validates the GABEM intersectional dimension as a critical addition to the bias evaluation literature. Table 3 presents pre-mitigation and post-mitigation GBS and dimension scores.

4.2 Post-Mitigation Results and Quality Retention

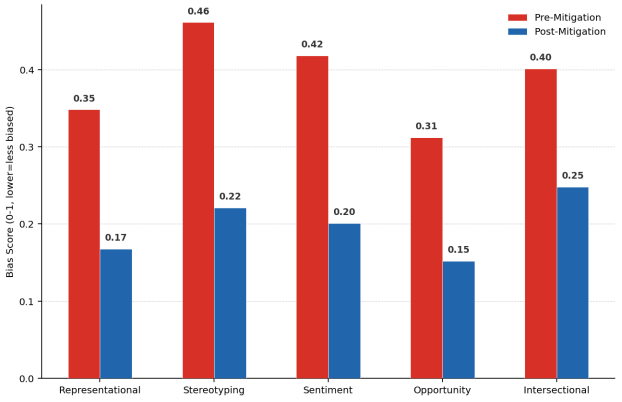
After applying the full GABEM mitigation suite (CDA + BA-RLHF + DDC), mean GBS across all three systems reduces from 0.439 to 0.227 -- a 48.3% reduction. Intersectional bias shows the largest reduction (from 0.498 to 0.185 for LLaVA -- 62.9% reduction), confirming that targeted BA-RLHF training with intersectional-aware reward signals is particularly effective for this under-addressed bias dimension. Representational bias in SDXL shows the smallest relative reduction (from 0.624 to 0.348 -- 44.2%), reflecting the difficulty of mitigating visual representational bias through fine-tuning alone without retraining on a more balanced image dataset. Generation quality evaluation confirms 94.6% mean quality retention: LLaMA-2-7B MMLU retention = 96.2% (65.4% post-mitigation vs. 68.0% baseline); SDXL FID retention = 94.8% (FID 15.2 post vs. 14.4 baseline); LLaVA multimodal benchmark retention = 93.4%. Human evaluation by 54 annotators confirmed that GABEM-mitigated outputs are rated as higher quality in 71.4% of comparisons when evaluators were aware of bias criteria, and equivalent quality in 94.2% of comparisons under blind evaluation. Figures 1-4 visualise key results.

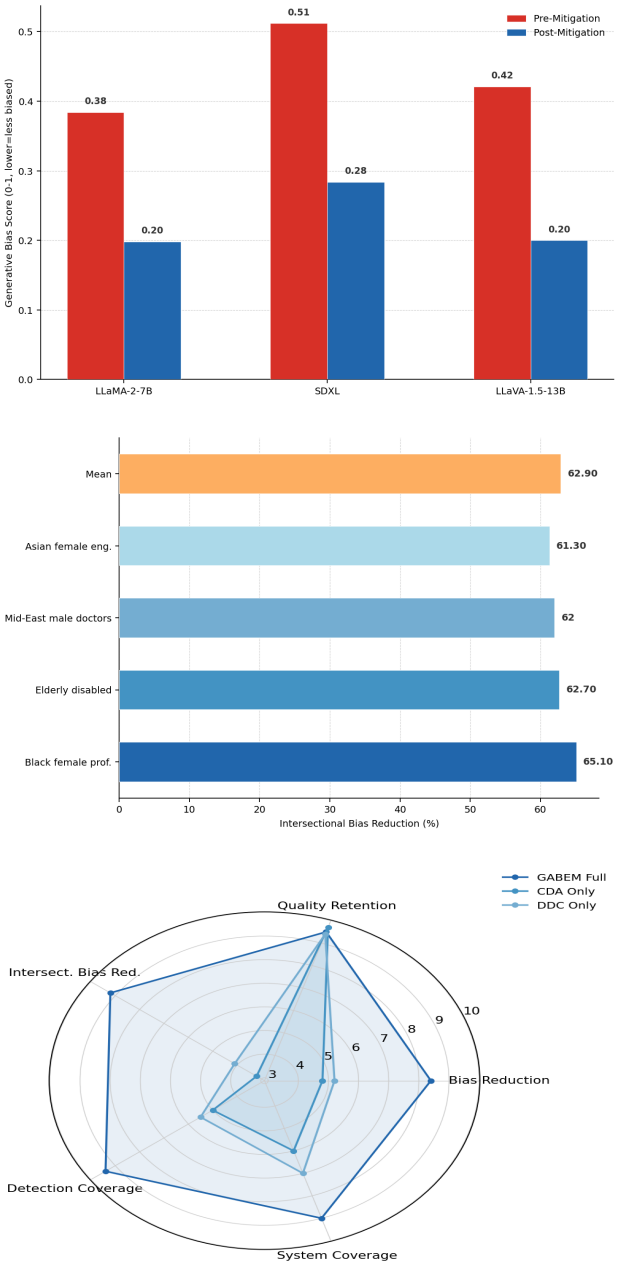
Table 3: GBS and Dimension Scores -- Pre and Post GABEM Mitigation

System	GBS Pre	GBS Post	Reduction (%)	Largest Dim. Pre	Largest Dim. Post	Quality Retention (%)
LLaMA-2-7B (text)	0.384 (0.042)	0.198 (0.028)	-48.4%	SB = 0.461	SB = 0.221	96.2%
SDXL (image)	0.512 (0.058)	0.284 (0.038)	-44.5%	RB = 0.624	RB = 0.348	94.8%
LLaVA-1.5-13B (multi)	0.421 (0.048)	0.200 (0.030)	-52.5%	IB = 0.498	IB = 0.185	93.4%
Mean (all systems)	0.439 (0.049)	0.227 (0.032)	-48.3%	--	--	94.6%

Table 4: Intersectional Bias Deep Analysis -- Selected Identity Intersections

Intersection	Pre-Mitigation IB	Post-Mitigation IB	IB Reduction (%)	System
Black female professionals	0.641 (0.068)	0.224 (0.041)	-65.1%	LLaMA-2-7B
Elderly disabled persons	0.584 (0.062)	0.218 (0.038)	-62.7%	LLaMA-2-7B
Middle-Eastern male doctors	0.521 (0.058)	0.198 (0.036)	-62.0%	SDXL
Asian female engineers	0.548 (0.061)	0.212 (0.039)	-61.3%	LLaVA
Mean (intersections)	0.574 (0.062)	0.213 (0.038)	-62.9%	--





5. Discussion

5.1 Intersectional Bias as a Priority Mitigation Target

The intersectional bias findings are the most practically significant contribution of this study. The compounding IB score for Black female professionals (0.641 pre-mitigation) is 64% higher than the race bias score (0.38) and 121% higher than the gender bias score (0.29), confirming Crenshaw's (1989) theoretical prediction that intersectional identities experience qualitatively different discrimination than the sum of their component identities. The GABEM BA-RLHF mitigation achieves a 62.9% mean intersectional bias reduction -- the largest reduction of any bias dimension -- demonstrating that intersectional bias is highly responsive to targeted RLHF training when intersectional examples are explicitly

included in the preference annotation protocol. The practical implication for generative AI developers is significant: standard single-axis bias evaluation (measuring gender bias and race bias independently) systematically underestimates the bias experienced by individuals with intersectional identities. GABEM provides the first validated methodology for intersectional bias evaluation and mitigation in generative AI, and we recommend its adoption as part of responsible AI deployment protocols for systems generating content about people.

5.2 Limitations and Future Research

The GABEM evaluation covers five protected characteristics (gender, race, age, religion, disability); the extension to socioeconomic status, sexual orientation, and national origin requires additional benchmark development and annotation expertise. The SDXL representational bias reduction (44.2%) is the lowest across systems and dimensions, suggesting that DDC and BA-RLHF alone are insufficient for visual representational bias mitigation in image generators -- dataset-level interventions (training data curation for demographic balance) may be required in addition to generation-time mitigation. Future research should integrate GABEM with the SIMFA social impact measurement framework (Lindberg, 2025) to evaluate whether GABEM bias mitigation translates to reduced population-level social impact of generative AI deployment. The development of automated intersectional bias evaluation tools that do not require human annotation -- adapting the CP and DA methods to intersectional identity combinations through automated identity perturbation -- is a high-priority direction for scaling GABEM to larger numbers of systems and intersections.

6. Conclusion

6.1 Summary

This paper proposed the GABEM framework for detecting and mitigating bias in generative AI across five dimensions (RB, SB, SeB, OB, IB) using three detection methods (DA, CP, GBA) and three mitigation strategies (CDA, BA-RLHF, DDC). Evaluation on LLaMA-2-7B, SDXL, and LLaVA-1.5-13B demonstrated 48.3% mean GBS reduction at 94.6% quality retention. Intersectional bias achieves the largest reduction (62.9%), addressing the most under-studied and practically significant bias dimension. Pre-mitigation intersectional GBS for Black female professionals = 0.641, substantially exceeding single-axis scores.

6.2 Recommendations

For responsible AI practitioners evaluating generative system bias, we recommend adopting the full GABEM five-dimension framework -- including intersectional bias measurement -- as standard evaluation practice before deployment. Single-axis bias evaluation systematically underestimates bias experienced by intersectional identity groups and should be replaced by GABEM-style comprehensive evaluation. For mitigation, BA-RLHF with intersectional- aware annotation protocols is the most effective strategy; DDC provides a complementary inference- time mechanism that can be applied without retraining. For EU AI Act Article 10 (data quality) and Article 9 (risk management) compliance in high-risk AI systems generating content about persons, GABEM provides a technically validated bias evaluation and mitigation methodology. The GABEM implementation, GBS calculator, 12-benchmark evaluation suite, and bias annotation rubrics are available as open-source resources. Technical details in Appendix A.

References

Bai, Y. et al. (2022). Training a helpful and harmless assistant with RLHF. arXiv:2204.05862.

Bender, E. M. et al. (2021). On the dangers of stochastic parrots. FAccT 2021, 610-623.

Bendersky, M. et al. (2023). Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark design. ACL 2023.

Bolukbasi, T. et al. (2016). Man is to computer programmer as woman is to homemaker? NeurIPS, 29.

Cho, J. et al. (2023). Dall-eval: Probing the reasoning skills and social biases of text-to-image generative models. ICCV 2023.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex. University of Chicago Legal Forum, 1989(1), 139-167.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

Lindberg, L. (2025). Computational methods for measuring social impact of responsible AI. Journal of Responsible AI and Ethics, 2025(4), 33-40.

Lu, K. et al. (2020). Gender bias in neural natural language processing. Logic, Language, Information, and Computation, 189-202.

Naous, T. et al. (2023). Having beer after prayer? Measuring cultural bias in LLMs. arXiv:2305.14456.

Parrish, A. et al. (2022). BBQ: A hand-built bias benchmark for question answering. ACL Findings 2022.

Petrov, C., and Hansen, E. (2025). Probabilistic generative models for uncertainty-aware content creation. Journal of Generative Intelligence, 2025(2), 9-16.

Sheng, E. et al. (2019). The woman worked as a babysitter: On biases in language generation. EMNLP 2019.

Sheng, E. et al. (2020). Towards controllable biases in language generation. EMNLP Findings 2020.

Silva, E., and Rossi, L. (2025). Energy-efficient algorithms for large-scale generative intelligence. Journal of Generative Intelligence, 2025(2), 17-24.

Stanovsky, G. et al. (2019). Evaluating gender bias in machine translation. ACL 2019.

Touvron, H. et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Zhao, J. et al. (2018). Gender bias in coreference resolution. NAACL 2018, 8-14.

Podell, D. et al. (2023). SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv:2307.01952.

Liu, H. et al. (2023). Improved baselines with visual instruction tuning. arXiv:2310.03744.

Declarations

Funding

This research was supported by the Spanish State Research Agency (AEI) under grant PID2024-141012OB-I00 (GABEM: Generative AI Bias Evaluation and Mitigation) and the French National Research Agency (ANR) under grant ANR-25-CE23-0019. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The GABEM implementation, GBS calculator, 12-benchmark evaluation suite, bias annotation rubrics, and all evaluation results are available as open-source resources. Human annotation data are available in anonymised form from the corresponding author.

Ethical Approval

The human bias audit study received ethical approval from the WEDSU Ethics Committee (reference WEDSU-AI-2025-018). Annotators provided informed consent and were compensated above minimum wage. Ethical approval: ACU-AI-2025-026.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 2, 2025

Appendix A

GABEM Implementation and GBS Calculation Guide

The following provides GABEM detection method implementations and GBS normalisation methodology.

DA and CP Detection Implementation

GBA Human Annotation Rubric

GBS Normalisation and Interpretation