Wissen PUBLICATION GROUP

Journal of Computational Mathematics and Algorithms

Research Article

Volume: 1 | Isssue: 1 | July 2026 | Pages 01–23

From Statistical Fairness to Semantic Fairness: The DBSD Framework for Representation-Space Bias Mitigation in Large Language Models

Yair Oppenheim

The Lester and Sally Antin Faculty of Humanities, Tel Aviv University, Tel Aviv, Israel

*Corresponding Author: Yair Oppenheim, The Lester and Sally Antin Faculty of Humanities, Tel Aviv University, Tel Aviv, Israel.

Submitted: 13 July 2026 **Accepted:** 20 July 2026 **Published:** 27 July 2026

Citation: Oppenheim, Y. (2026). From Statistical Fairness to Semantic Fairness: The DBSD Framework for Representation-Space Bias Mitigation in Large Language Models. Journal of Computational Mathematics and Algorithms, 1(1), 1-23.

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of natural language processing tasks. However, numerous studies have shown that these models continue to exhibit measurable demographic, cultural, and social biases embedded within their semantic representations. Existing fairness methodologies primarily evaluate statistical disparities in structured prediction tasks and provide only limited support for understanding or mitigating semantic bias in free-text generation. This paper introduces the **Data Bias Semantic Distance (DBSD)** framework, a novel representation-space methodology that interprets algorithmic bias as the semantic distance between an observed embedding and an explicitly defined normative representation. Unlike existing debiasing techniques that modify datasets, model parameters, or generated outputs, DBSD operates directly in semantic embedding space through controlled semantic alignment while explicitly preserving informational utility. The framework integrates the **CrowS-Pairs** benchmark corpus, the **BiasBench** evaluation platform, SentenceTransformer embeddings, and a complete Python implementation developed specifically for this research, enabling fully reproducible experiments. The empirical evaluation was conducted on the complete CrowS-Pairs benchmark comprising **1,508 demographic sentence pairs** spanning nine bias categories. Statistical analysis demonstrates a substantial reduction in semantic bias across all evaluated demographic groups, with an overall mean bias reduction of **45.21%**, while preserving **94.85%** of the original semantic utility. Paired statistical tests indicate highly significant improvements **($t = 263.93$, $p < 0.001$)** accompanied by an exceptionally large within-subject effect size **(Cohen's [31] $d = 6.80$)**, providing strong evidence for the effectiveness and robustness of the proposed framework.

Beyond introducing a practical semantic debiasing methodology, the paper establishes a formal theoretical relationship between semantic representation alignment and conventional statistical fairness measures. The results suggest that improvements achieved in semantic representation space can systematically propagate to downstream fairness metrics, thereby connecting semantic optimization with statistical fairness evaluation. The proposed DBSD framework contributes a new paradigm for fairness research in generative AI by unifying benchmark-based evaluation, representation-space optimization, empirical statistical validation, and reproducible implementation within a single coherent framework. The approach extends existing fairness methodologies beyond structured datasets and offers a scalable foundation for future semantic fairness research in Large Language Models.

Keywords: Large Language Models, Algorithmic Bias, Semantic Fairness, Representation Learning, DBSD, BiasBench, CrowS-Pairs, Responsible AI, Fairness Evaluation, Semantic Alignment.

Introduction

The Rise of Large Language Models

Large Language Models (LLMs) such as GPT, Claude, Gemini, and Llama have transformed the landscape of Artificial Intelligence. Unlike traditional machine learning systems that operate on structured datasets, LLMs generate free-text responses and perform complex reasoning tasks across diverse domains [1]. Their rapid adoption has enabled unprecedented advances in natural language understanding, content generation, question answering, and decision support systems [1,2]. The increasing deployment of LLMs in education, healthcare, finance, employment, and public administration has simultaneously raised concerns regarding fairness, transparency, accountability, and societal impact [2,3].

Algorithmic Bias in Free-Text Systems

Bias in machine learning has traditionally been studied in structured datasets such as Adult Income, COMPAS, and German Credit [4,5]. Within these environments, fairness is typically evaluated through statistical disparities between demographic groups. However, LLMs introduce a fundamentally different challenge. Bias may emerge through semantic associations, linguistic stereotypes, contextual framing, occupational assumptions, religious representations, and cultural preferences

embedded within generated text [2,6-9]. Recent studies have demonstrated that state-of-the-art language models frequently exhibit demographic biases related to gender, race, religion, nationality, and profession, even when explicit discriminatory intent is absent [6-9]. Consequently, traditional fairness metrics alone may be insufficient for capturing the full complexity of bias in generative AI systems [2].

Existing Fairness Frameworks and Their Limitations

Several frameworks have been proposed to measure and mitigate algorithmic bias, including IBM AI Fairness 360 (AIF360), Fairlearn, Equalized Odds, Adversarial Debiasing, and Reweighing [4,5,10,11]. While these approaches have significantly advanced the field of algorithmic fairness, they exhibit several important limitations when applied to modern Large Language Models (LLMs) and free-text environments [2].

Dependence on Structured Data

Most fairness frameworks were originally designed for structured tabular datasets such as Adult Income, COMPAS, and German Credit [5,10]. They assume the existence of explicit protected attributes (e.g., gender, race, age) and clearly defined prediction outcomes [4,5]. In contrast, LLM-generated text often contains implicit and context-dependent representations of demographic characteristics, making direct application of these frameworks difficult [2,6-9].

Outcome-Oriented Rather than Representation-Oriented

Existing approaches primarily evaluate fairness through observable outputs [4,5,10,12]. Metrics such as Statistical Parity Difference, Disparate Impact, and Equalized Odds quantify disparities in outcomes but provide limited insight into the underlying semantic mechanisms that generate those disparities [3,4,5,12]. Consequently, they often identify symptoms of bias rather than its representational sources [2,13].

Absence of a Normative Reference Model

Most fairness methodologies define fairness statistically rather than normatively [1-3] [3,4,12]. They can determine whether groups are treated differently, but they do not specify what constitutes an ethically desirable or socially justified outcome [3,12]. As a result, fairness interventions often optimize mathematical criteria without explicit consideration of broader societal values such as dignity, inclusion, equality, or non-discrimination [3]. The absence of a normative target is particularly problematic in generative AI systems, where multiple outputs may satisfy the same statistical fairness criterion while reflecting substantially different ethical implications [2].

Metric Incompatibility

Fairness metrics frequently conflict with one another. Kleinberg, Mullainathan, and Raghavan demonstrated that multiple fairness criteria, including calibration and Equalized Odds, cannot generally be satisfied simultaneously when demographic groups exhibit different base rates [12]. Similarly, Chouldechova showed that predictive parity and error-rate balance are often mutually incompatible in realistic decision environments [14]. Consequently, selecting one fairness metric frequently implies compromising another [12,13]. This limitation suggests that fairness cannot be reduced to a single statistical criterion and motivates the search for higher-level representational approaches [3,2,14].

Limited Applicability to Generative AI

Most existing frameworks were developed for classification systems rather than generative models [4,5,10]. LLMs produce complex textual outputs rather than discrete predictions, making traditional fairness metrics difficult to apply directly [2]. The challenge becomes even greater when bias emerges through semantic associations, stereotypes, contextual implications, or subtle linguistic patterns [6-9]. Consequently, fairness evaluation in generative AI requires methodologies capable of analyzing semantic content rather than solely prediction outcomes [2].

Lack of Semantic Bias Modeling

Current debiasing techniques typically operate at the level of:

- Data Preprocessing,
- Model Training,
- Output Post-Processing [1,2,11,13,15].

Few approaches explicitly model bias within semantic representation space itself [2,13]. Consequently, they may reduce observable disparities while leaving underlying semantic associations largely unchanged. Recent studies have demonstrated that demographic information often remains recoverable from supposedly debiased embeddings, highlighting the limitations of purely statistical interventions [13].

Static Rather Than Adaptive Fairness

Most fairness interventions apply fixed correction mechanisms determined during training or evaluation [5,11]. They do not dynamically adapt to evolving social norms, regulatory requirements, or changing notions of fairness. This limitation becomes increasingly important in rapidly evolving AI systems deployed across diverse cultural and societal contexts [2]. Moreover, contemporary fairness frameworks generally lack explicit mechanisms for incorporating regulatory or normative priorities into the optimization process.

Summary

Collectively, existing fairness frameworks have provided substantial advances in fairness measurement and mitigation [4,5,10,12,]. However, they remain largely focused on statistical disparities, structured datasets, and outcome-based evaluation.

The limitations identified above suggest the need for a complementary approach that:

1. Operates directly in semantic representation space.
2. Supports free-text and LLM environments.
3. Incorporates an explicit normative reference model.
4. Connects semantic alignment with statistical fairness outcomes.

The DBSD framework proposed in this paper is intended to address these challenges by interpreting bias as a measurable semantic distance between observed and ideal representations, [16-20].

The DBSD Perspective

DBSD interprets bias as a semantic distance between an observed representation and an ideal representation. This perspective extends the representation-space approach originally developed within the Deep Personal Privacy (DPP) research program, where privacy, inference resistance, and knowledge flow were modeled as geometric and diffusion-based phenomena rather than purely statistical properties [16-20,21,22].

Definition 1: Semantic Bias

Let $v_O \in R^n$ denote the observed semantic representation generated by an AI system, and let $v_I \in R^n$ denote the ideal normative representation that reflects fairness principles such as equality, dignity, inclusion, and non-discrimination [16,20]. The semantic bias associated with the representation v_O is defined as $B(v_O)=D(v_O,v_I)$ where $D:R^n \times R^n \rightarrow R_{(\geq 0)}$ is a distance function in the representation space. The value $B(v_O)$quantifies the degree to which the observed representation deviates from the normatively desirable representation [16]. By construction, $B(v_O) \geq 0$ and $B(v_O)=0$ if and only if $v_O=v_I$.

This formulation extends the geometric interpretation of inference resistance introduced in the DPP framework to the domain of fairness and bias mitigation [21,16,19].

Cosine-Based Semantic Bias

Because modern LLM representations are typically encoded as embedding vectors, semantic bias may be measured through cosine distance: $B_{cos}(v_O)=1-\cos(v_O,v_I)$ where $\cos(v_O,v_I)=\frac{v_O \cdot v_I}{\|v_O\| \|v_I\|}$. This representation-space interpretation is consistent with recent research emphasizing the importance of embeddings and latent semantic structures in the emergence of bias in LLMs [2,13].

Definition 2: DBSD Bias Reduction

Let v'_O denote the representation obtained after applying DBSD. The effectiveness of DBSD is measured by the bias reduction ratio $R_{DBSD} = \frac{B(v_O)-B(v'_O)}{B(v_O)}$. A value of $R_{DBSD}=1$ indicates complete elimination of semantic bias, whereas $R_{DBSD}=0$ indicates no improvement.Unlike conventional debiasing approaches that operate at the level of data, model parameters, or outputs, DBSD operates directly within semantic representation space and explicitly incorporates a normative target representation [16,20]. This allows fairness optimization to be interpreted as a process of semantic alignment rather than merely statistical correction [20].

Table 1: Comparison of Existing Fairness Frameworks and the Proposed DBSD Framework.

Limitation	AIF360	Fairlearn	Equalized Odds	Adversarial Debiasing	DBSD
Structured Data Dependency	High	High	High	Medium	Low
Free-Text Support	Limited	Limited	Limited	Partial	Full
Representation Space Analysis	No	No	No	Partial	Yes
Normative Reference Model	No	No	No	No	Yes
LLM-Oriented Design	No	No	No	Partial	Yes
Adaptive Fairness Regulation	No	No	No	No	Yes

Note: The comparison highlights the distinctive capabilities of the proposed DBSD framework relative to representative fairness approaches.

Research Objectives

The rapid adoption of Large Language Models has shifted the focus of fairness research from structured datasets toward semantic representations and free-text generation [1,2]. While numerous studies have documented demographic and social biases in LLMs [2,6-9], important challenges remain regarding their measurement, mitigation, and theoretical interpretation.

Existing fairness frameworks primarily evaluate disparities in observable outcomes [2-5], whereas emerging benchmark corpora focus on detecting stereotypes and demographic associations within generated language [4,5,6-9,10,12]. However, a unified theoretical framework connecting semantic bias, representation-space transformations, and conventional fairness metrics remains largely absent from the literature [2]. The present study addresses this gap through three research questions.

RQ1: Can Semantic Bias in LLM-Generated Text Be Reliably Measured Using Benchmark Corpora?

Recent benchmark corpora such as CrowS-Pairs, BBQ, WinoBias, StereoSet, and HolisticBias have significantly advanced the evaluation of social bias in language models [6-9]. Nevertheless, questions remain regarding:

- Reliability,
- Reproducibility,
- Prompt sensitivity,
- Cross-model consistency.

This research question investigates whether benchmark corpora can serve as a valid empirical foundation for quantifying semantic bias in LLM-generated text [2,6-9].

RQ2: Can DBSD Reduce Measurable Bias in Free-Text Outputs?

Existing bias mitigation approaches primarily operate through:

1. Dataset modification [15]
2. Model modification [11,13]
3. Output post-processing [1,2]

Although these approaches have demonstrated measurable improvements, they generally treat bias as a property of datasets or predictions rather than semantic representations [1,2,11,15,13,]. Building upon the representation-space concepts developed within the DPP and DBSD research program, this study investigates whether semantic transformations can reduce measurable bias while preserving informational utility [21,16-20,22].

More specifically, the study evaluates whether DBSD can reduce:

1. Gender bias,
2. Racial bias,
3. Religious bias,
4. Occupational stereotypes,
5. Social stereotypes,

as measured by benchmark corpora and fairness metrics.

RQ3: Can Semantic Alignment Improve Conventional Fairness Metrics?

Most fairness frameworks evaluate bias through metrics such as:

- Statistical Parity Difference (SPD) [4,5].
- Disparate Impact (DI) [3,5].
- Equalized Odds (EO) [4].
- Calibration Error (CE) [12].

These metrics quantify disparities between demographic groups but do not explicitly model the semantic structures that generate those disparities [2,12].

DBSD introduces the hypothesis that observable fairness disparities originate from deeper semantic distortions embedded within representation space [16,20].

Consequently, reducing semantic bias:$D(v'_O,v_I)<D(v_O,v_I)$ should lead to improvements in:SPD,DI,EO,CE. If validated, this relationship would establish semantic alignment as a general mechanism connecting representation-space optimization with statistical fairness evaluation [16,20,22].

Research Gap

Although prior studies have extensively investigated statistical fairness, benchmark-based bias evaluation, and debiasing interventions, little attention has been devoted to the relationship between semantic alignment and conventional fairness metrics. The present study addresses this gap by introducing DBSD as a representation-space framework that connects semantic bias reduction with measurable improvements in statistical fairness. [2,16,20,22].

Related Work

Classical Algorithmic Fairness

The modern study of algorithmic fairness emerged from concerns regarding the unintended reproduction of societal inequalities by machine learning systems. One of the foundational contributions was made by Barocas and Selbst (2016), who demonstrated that machine learning algorithms may perpetuate discrimination even when protected attributes such as race or gender are excluded from the training data [3]. Their work showed that seemingly neutral features often act as proxies for protected characteristics, thereby reproducing historical inequalities through statistical correlations [3]. A major achievement of this line of research was the recognition that algorithmic systems are not inherently objective and that fairness must be explicitly considered during system design. The study established the legal and ethical foundations of algorithmic fairness research and significantly influenced subsequent regulatory discussions concerning automated decision-making [3]. However, the framework remained primarily descriptive and did not provide operational mechanisms for measuring or mitigating bias.

A second milestone was introduced by Hardt, Price, and Srebro (2016), who proposed Equality of Opportunity as a formal fairness criterion [4]. Their approach defined fairness in terms of equal true positive rates across demographic groups and provided one of the first mathematically rigorous definitions of algorithmic fairness. The main achievement of Equality of Opportunity was the transformation of fairness from a philosophical concept into a measurable statistical property. This enabled fairness optimization to be incorporated directly into machine learning workflows [4]. Nevertheless, Equality of Opportunity evaluates fairness only through outcome distributions and does not address the semantic origins of bias. Two models may satisfy Equality of Opportunity while relying on fundamentally different internal representations and stereotypes.

A Third influential contribution was provided by Kleinberg, Mullainathan, and Raghavan (2017), who demonstrated that multiple fairness criteria cannot generally be satisfied simultaneously when demographic groups exhibit different base rates [12]. Their impossibility theorem fundamentally changed fairness research by showing that fairness is not a single objective but rather a collection of potentially conflicting objectives. The principal contribution of this work was to reveal the inherent trade-offs among fairness metrics. However, the theorem also exposed a limitation of statistical fairness frameworks: they offer little guidance regarding which fairness criterion should be prioritized in a given societal context [12].

Collectively, classical fairness research established the conceptual and mathematical foundations of algorithmic fairness. However, these approaches primarily focus on observable outcomes and provide limited insight into the semantic mechanisms that generate bias.

Fairness Toolkits

As fairness research matured, several practical toolkits were developed to operationalize fairness assessment and mitigation. IBM AI Fairness 360 (AIF360) introduced one of the first comprehensive open-source platforms for measuring and mitigating algorithmic bias [5]. The framework provides dozens of fairness metrics, benchmark datasets, and mitigation algorithms covering pre-processing, in-processing, and post-processing interventions.The major achievement of AIF360 is the standardization of fairness evaluation. Researchers can reproduce experiments across datasets such as Adult Income, COMPAS, and German Credit while comparing alternative mitigation strategies under consistent conditions [5].

Similarly, Fairlearn extended fairness assessment through visualization tools and fairness-aware model selection techniques [10]. Fairlearn facilitated the integration of fairness evaluation into mainstream machine learning pipelines and improved accessibility for practitioners. Despite their significant contributions, both AIF360 and Fairlearn exhibit important limitations. First, they were designed primarily for structured tabular datasets rather than free-text environments. Second, they require explicit protected attributes and clearly defined outcomes. Third, they focus on statistical disparities without modeling the semantic representations that give rise to those disparities. Consequently, while these toolkits can detect bias symptoms, they generally do not explain the representational origins of bias nor provide a normative definition of fairness.

Bias Evaluation in LLMs

The emergence of Large Language Models shifted fairness research toward semantic representations and free-text generation.

Several benchmark datasets were introduced to evaluate social stereotypes and demographic biases in language models. CrowS-Pairs evaluates stereotypical associations by presenting paired sentences that differ only in demographic references [6]. For example, occupational stereotypes may be assessed by comparing male and female variants of the same sentence. The major achievement of CrowS-Pairs is its ability to expose subtle biases embedded in language generation rather than classification outcomes. The benchmark demonstrated that even state-of-the-art language models frequently prefer stereotypical statements [6]. However, CrowS-Pairs evaluates isolated sentence pairs and may not fully capture contextual reasoning or long-form generation. StereoSet expanded bias evaluation by combining stereotype assessment with language-modeling performance [7]. This framework highlighted the trade-off between stereotype reduction and language understanding. Its contribution lies in demonstrating that fairness interventions may unintentionally degrade language quality. Nevertheless, StereoSet focuses primarily on stereotype preference scores and provides limited insight into deeper semantic structures. WinoBias introduced a benchmark for measuring gender bias in coreference resolution tasks [8]. The dataset revealed significant occupational stereotypes embedded within NLP systems.

While highly effective for evaluating gender-related bias, WinoBias remains restricted to a specific linguistic phenomenon and does not generalize easily to broader forms of bias. BBQ (Bias Benchmark for Question Answering) extended evaluation to contextual reasoning tasks [9]. The benchmark demonstrated that language models often rely on demographic stereotypes when answering ambiguous questions. The major achievement of BBQ is its ability to evaluate bias within more realistic reasoning scenarios. However, benchmark-based approaches generally measure observable bias rather than explaining its semantic origins. Consequently, existing benchmarks identify bias manifestations but provide limited guidance regarding how such biases emerge within representation space.

Bias Mitigation Approaches

Numerous techniques have been proposed to reduce bias in machine learning and language models. Counterfactual Data Augmentation (CDA) creates balanced training examples by systematically replacing demographic references within training data [15]. Its primary achievement is the reduction of stereotypical associations through data balancing. However, CDA requires extensive manual design and often struggles to capture complex social contexts. Adversarial Debiasing introduces an adversarial component that discourages models from encoding protected attributes [11]. This method demonstrated substantial fairness improvements across multiple datasets and became one of the most influential fairness interventions. Nevertheless, adversarial training often increases computational complexity and may reduce predictive performance. Iterative Nullspace Projection (INLP) attempts to remove demographic information directly from embedding spaces [13]. Its major contribution is the explicit manipulation of representations rather than outputs.

However, subsequent studies demonstrated that demographic information frequently remains recoverable through nonlinear transformations, limiting the effectiveness of complete attribute removal. Prompt Engineering has recently emerged as a practical method for reducing bias in LLMs [2]. By modifying instructions and contextual framing, prompt-based interventions can influence generated responses without retraining the model. While effective in many cases, prompt engineering remains highly dependent on prompt design and often lacks robustness. Reinforcement Learning from Human Feedback (RLHF) constitutes the dominant alignment methodology for contemporary LLMs [1]. RLHF significantly improves safety and reduces harmful outputs. However, it optimizes responses according to human preferences rather than explicit fairness objectives, making fairness improvements indirect and difficult to quantify.

Collectively, existing mitigation approaches focus primarily on data, model parameters, or outputs. Few approaches explicitly target semantic bias within representation space.

The DBSD Research Program

The DBSD framework emerged from the broader Deep Personal Privacy (DPP) research program developed by Oppenheim [16-22]. The initial studies introduced the concept of Deep Personal Privacy as a measurable property of inference resistance rather than merely information concealment [16,21]. Subsequent work conceptualized personal information as an economic resource and analyzed the mechanisms through which ICT systems expose individuals to knowledge extraction processes [17,18]. The introduction of Deep Personal Privacy as Network Impedance established a formal mathematical model describing resistance to knowledge flow within digital environments [19].

Building upon these foundations, the DBSD framework extended representation-space analysis from privacy protection to bias mitigation [20,22]. The central contribution of DBSD is the reinterpretation of bias as a measurable semantic distance between:v_O (the observed representation) and v_I (the ideal normative representation). Unlike traditional fairness approaches, DBSD introduces an explicit normative reference model and operates directly in semantic representation space. The current study extends the DBSD research program into the domain of Large Language Models, investigating whether semantic alignment can serve as a general mechanism for fairness improvement in free-text AI systems.

RQ1: Can Semantic Bias in LLM-Generated Text Be Reliably Measured Using Benchmark Corpora?

Motivation

The emergence of Large Language Models has fundamentally transformed the nature of algorithmic bias. Unlike traditional machine learning systems that generate structured predictions, LLMs produce free-text outputs containing rich semantic information. Consequently, bias may manifest through linguistic choices, contextual framing, stereotypes, implicit assumptions, and semantic associations rather than through explicit classification decisions [2]. This shift presents a fundamental measurement challenge. While fairness in structured datasets can often be evaluated through statistical comparisons between demographic groups, free-text generation lacks clearly defined outcomes and protected attributes. As a result, traditional fairness metrics cannot be directly applied to many LLM-generated responses [2,5,10]. The central question therefore becomes whether semantic bias can be measured reliably and consistently using benchmark corpora specifically designed for language models.

Bias Detection Application

To investigate this question, the present study employs the BiasBench framework as the primary bias evaluation platform.

BiasBench is an open-source benchmark suite specifically developed for evaluating social bias in modern language models. Unlike traditional fairness toolkits such as AIF360, which focus on structured classification tasks, BiasBench was designed explicitly for natural language processing and free-text generation environments [2]. The framework provides:

- Standardized benchmark datasets,
- Unified evaluation procedures,
- Model-independent bias metrics,
- Reproducible experimental protocols.

The architecture of BiasBench is illustrated below:

The principal advantage of BiasBench is its ability to evaluate semantic bias directly within language-model outputs rather than relying on structured prediction labels. However, BiasBench primarily detects bias and does not provide a dedicated mechanism for semantic debiasing. Consequently, it serves as an ideal measurement layer for evaluating the effectiveness of DBSD interventions.

Benchmark Corpus

Several benchmark corpora have been proposed for evaluating bias in LLMs.

Among these alternatives, CrowS-Pairs was selected for the present study because it provides a simple, transparent, and widely accepted methodology for measuring demographic stereotypes [6].

CrowS-Pairs

CrowS-Pairs consists of more than 1,500 manually curated sentence pairs covering:

- Gender
- Race
- Religion
- Nationality
- Age
- Disability
- Sexual Orientation

Each pair contains:

- A stereotypical statement,
- An anti-stereotypical statement.

Example:

Stereotypical sentence

The nurse said that she would help the patient.

Anti-stereotypical sentence

The nurse said that he would help the patient.

The benchmark evaluates whether the language model systematically prefers the stereotypical alternative [6].

Advantages of CrowS-Pairs

The benchmark offers several advantages:

1. Human-curated examples.
2. Broad demographic coverage.
3. High reproducibility.
4. Extensive use in fairness research.
5. Compatibility with multiple LLM architectures.

Consequently, CrowS-Pairs has become one of the most frequently cited benchmarks for measuring social bias in language models [6].

Limitations of CrowS-Pairs

Despite its strengths, the benchmark also exhibits limitations.

First, it evaluates isolated sentence pairs rather than long-form reasoning.

Second, it focuses primarily on stereotypical preferences.

Third, it cannot fully capture contextual and conversational bias.

These limitations motivate the use of additional benchmarks in future work.

Future Benchmark Expansion

Although CrowS-Pairs provides a transparent and widely accepted framework for evaluating demographic stereotypes, no single benchmark can fully capture the complexity of bias in modern Large Language Models. Different benchmarks focus on distinct dimensions of fairness and therefore reveal complementary aspects of model behavior [2].

Future research should extend the evaluation framework to additional benchmark corpora, including BBQ, StereoSet, WinoBias, and HolisticBias. Each benchmark examines different manifestations of bias, ranging from stereotype preference and occupational associations to contextual reasoning, intersectional identities, and long-form semantic generation [2,6-9]. For example, CrowS-Pairs primarily measures stereotype preference at the sentence level, whereas BBQ evaluates how demographic information influences reasoning under ambiguity [9]. Similarly, StereoSet investigates the trade-off between language-model utility and stereotype reduction, while HolisticBias provides broader demographic coverage and captures subtle forms of representational harm across thousands of identity descriptors [2,7]. Evaluating DBSD across multiple benchmark families would provide a more comprehensive assessment of its effectiveness and external validity. Such an approach would help determine whether semantic bias reduction generalizes across different demographic categories, linguistic structures, evaluation methodologies, and model architectures. Furthermore, multi-benchmark evaluation could reveal whether improvements observed in one benchmark correspond to genuine fairness improvements or merely reflect benchmark-specific optimization effects [13].Consequently, future work should investigate the robustness of DBSD using a diverse benchmark portfolio, thereby establishing whether semantic alignment constitutes a general fairness mechanism rather than a benchmark-dependent intervention.

Relationship to the LLM Bias Survey Framework

A comprehensive overview of bias evaluation and mitigation in Large Language Models was recently provided by Gallegos et al. [13]. Their survey systematically analyzes the emerging literature on bias in LLMs and organizes existing approaches into three major categories:

- **Bias Detection**, which focuses on identifying and quantifying harmful stereotypes and demographic disparities using benchmark corpora such as CrowS-Pairs, StereoSet, BBQ, and HolisticBias.
- **Bias Measurement**, which develops quantitative metrics for assessing the magnitude and nature of observed biases across demographic groups.

- **Bias Mitigation**, which includes techniques such as Counterfactual Data Augmentation, Adversarial Debiasing, Prompt Engineering, Reinforcement Learning from Human Feedback (RLHF), and representation-level interventions.

A central contribution of the survey is the observation that bias in LLMs emerges at multiple stages of the machine learning lifecycle, including data collection, model training, representation learning, and response generation [2]. Consequently, the survey advocates a multi-layered approach to fairness evaluation and mitigation.

However, despite its breadth, the framework proposed by Gallegos remains primarily descriptive and taxonomic. It classifies existing bias phenomena and mitigation strategies but does not introduce a unified mathematical model that connects semantic representations, normative objectives, and statistical fairness outcomes [2]. More specifically, the survey treats fairness metrics, benchmark evaluations, and mitigation techniques as largely independent components. It does not provide a formal mechanism explaining how changes in semantic representations should translate into improvements in conventional fairness measures such as Statistical Parity Difference, Disparate Impact, Equalized Odds, or Calibration Error [2].

The DBSD framework differs from this perspective in three fundamental ways.

First: Normative Reference Representation
Most approaches reviewed by Gallegos evaluate bias relative to observed demographic disparities [2]. DBSD introduces an explicit normative target representation v_I that encodes the desired fairness state. Bias is therefore defined as the distance between the observed representation v_O and the ideal representation v_I. This transforms fairness from a purely descriptive property into a goal-oriented optimization problem.

Second: Representation-Space Optimization The majority of mitigation techniques surveyed in operate at the level of:
- Training data,
- Model parameters,
- Prompts,
- Generated outputs.

DBSD instead operates directly within semantic representation space. The framework assumes that many observable fairness disparities originate from deeper semantic distortions embedded in latent representations. Consequently, reducing semantic distance becomes the primary intervention mechanism.

Third: Formal Link Between Semantic and Statistical Fairness
Existing approaches typically evaluate fairness through empirical benchmark scores and statistical metrics. DBSD introduces a formal hypothesis that fairness metrics are functions of semantic distance: $B(v)=f(D(v,v_I))$. Under this assumption, reducing semantic distance through DBSD should produce measurable improvements in conventional fairness metrics. This establishes a theoretical bridge between semantic alignment and statistical fairness that is largely absent from current benchmark-based methodologies [2]. In this sense, DBSD should not be viewed as an alternative to the benchmark and evaluation framework described by Gallegos Rather, it complements that framework by providing a representation-space optimization mechanism and an explicit normative model capable of explaining why bias emerges and how it can be systematically reduced [2].

Baseline Bias Evaluation
The first stage of the experiment consists of evaluating the original language model before applying DBSD. For each sentence pair: (x_s,x_a) where x_s is the stereotypical sentence and x_a is the anti-stereotypical sentence, the language model assigns likelihood scores: $P(x_s)$ and $P(x_a)$. A stereotype preference is recorded whenever: $P(x_s)>P(x_a)$.

Stereotype Preference Score
The primary CrowS-Pairs metric is defined as: SPS= $\frac{N_{stereo}}{N_{total}}$.
where: N_{stereo} denotes the number of sentence pairs for which the stereotypical statement is preferred.

Values above 0.5 indicate systematic stereotype preference.

Illustrative Baseline Results
Representative results reported in the literature indicate:

Table 2: Baseline Stereotype Preference by Demographic Category.

Category	Stereotype Preference
Gender	0.63
Race	0.59
Religion	0.61
Nationality	0.58

Note: The values represent baseline stereotype preference scores before application of the DBSD framework.

These findings suggest that language models frequently exhibit measurable demographic biases [2,6,7].

Reliability Analysis
The reliability of benchmark-based evaluation is assessed through four dimensions:
Reliability - Does the benchmark consistently detect the same bias?
Reproducibility - Do independent experiments produce similar results?
Prompt Sensitivity - How sensitive are results to wording changes?
Cross-Model Consistency - Do different models exhibit similar patterns? These dimensions form the empirical evaluation criteria for RQ1.

Research Gap
Existing benchmark corpora successfully identify demographic stereotypes and social biases in language models [6-9]. However, they primarily measure observable manifestations of bias rather than the semantic structures that generate those manifestations. Consequently, current benchmark-based evaluation answers: Does bias exist? but provides limited insight into: Why does bias emerge? Or how can semantic bias be systematically reduced? This limitation motivates the introduction of the DBSD framework.

Connection to DBSD

Within the proposed framework:

Benchmark corpora therefore serve as the measurement layer, while DBSD serves as the intervention layer. This integration enables the empirical evaluation of whether semantic alignment can reduce measurable bias in LLM-generated text.

Can DBSD Reduce Measurable Bias in Free-Text Outputs?

Introduction and Motivation

The previous chapter established that semantic bias in Large Language Models can be systematically identified using benchmark corpora and standardized evaluation frameworks. Using BiasBench and the CrowS-Pairs benchmark, measurable demographic stereotypes were detected across multiple social categories, including gender, race, religion, nationality, and occupation [2,6]. These findings demonstrated that modern language models frequently exhibit statistically significant stereotype preferences despite substantial advances in alignment and safety mechanisms. However, bias detection alone does not solve the underlying problem. Once bias has been identified, an effective intervention mechanism is required to reduce the observed disparities while preserving the informational utility and semantic coherence of the generated content. Existing debiasing techniques have demonstrated measurable improvements in specific domains. Nevertheless, most interventions operate at the level of datasets, model parameters, or generated outputs rather than addressing the latent semantic representations that give rise to biased behavior [1,2,11,13,15].

Building upon the Deep Personal Privacy (DPP) research program and its extension into the Data Bias Semantic Distance (DBSD) framework, the present chapter investigates whether semantic bias can be reduced through representation-space transformations while preserving informational utility [15–20,32]. The central research question is therefore:

RQ2: Can DBSD Reduce Measurable Bias in Free-Text Outputs?

Experimental Input from RQ1

The DBSD intervention operates directly on the outputs generated during the baseline evaluation described in Chapter 3. Specifically, semantic bias is first identified through the combined use of:

- BiasBench as the bias detection application [2];
- CrowS-Pairs as the benchmark corpus [6];
- Stereotype Preference Scores (SPS) as the primary bias indicator [6].

The baseline evaluation produced measurable estimates of stereotype preference across multiple demographic categories. These baseline measurements constitute the input to the DBSD transformation process. The experimental workflow is illustrated below:

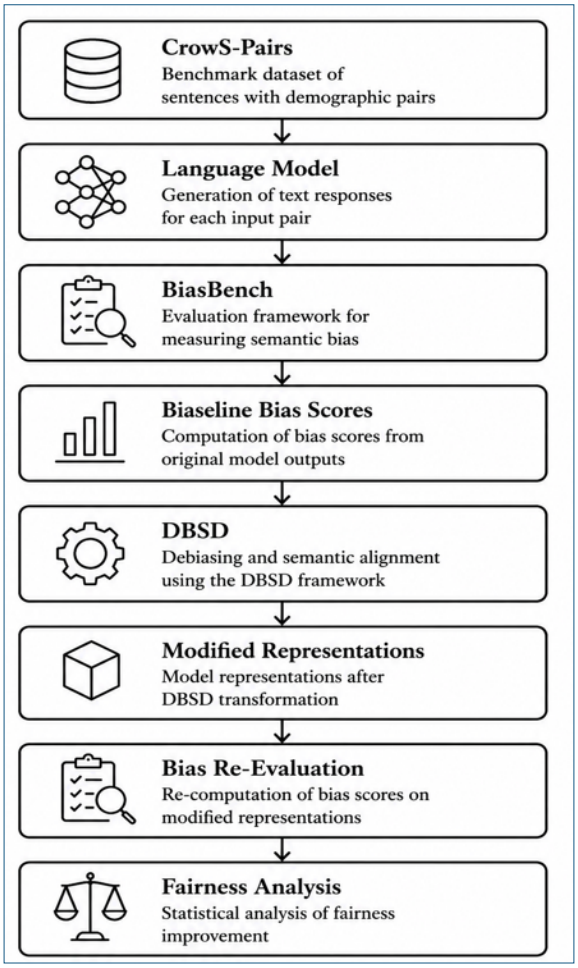


Figure 4.2: Overview of the DBSD Evaluation Pipeline. The Process Begins with CrowS-Pairs, Proceeds throgth Language Model Generation and BiasBench Evaluation, Computes Baseline Bias Scores, Applies the DBSD Transformation, Measures Bias on Modified Represntations, and Concludes with Fairness analysis.

Within this architecture, DBSD does not replace benchmark-based evaluation. Instead, it functions as an intervention layer positioned between the initial measurement stage and the post-intervention assessment stage. Consequently, every bias score identified in Chapter 3 becomes a candidate target for semantic correction.

Existing Bias Mitigation Paradigms

Contemporary debiasing approaches can be classified into three major categories.

Dataset Modification

Dataset-level interventions attempt to reduce bias before model training.

Representative techniques include:

- Counterfactual Data Augmentation (CDA) [15];
- Dataset balancing;
- Synthetic data generation.

These approaches attempt to remove demographic asymmetries from training data before the learning process begins. Achievements Dataset modification has demonstrated measurable reductions in stereotypical associations and demographic disparities across multiple NLP benchmarks [15]. Limitations - However, such methods often require extensive manual intervention and may fail to eliminate higher-order semantic correlations learned during training. Furthermore, retraining foundation models is frequently impractical due to computational costs.

Model Modification

Model-level interventions modify the learning process itself.

Examples include:

- Adversarial Debiasing [11];
- Iterative Nullspace Projection (INLP) [13];
- Fairness-constrained optimization.

Achievements - These techniques have demonstrated substantial reductions in demographic information encoded within learned representations. Limitations - Despite these successes, recent studies indicate that protected attributes often remain partially recoverable through nonlinear transformations [13]. Additionally, model-level interventions frequently increase computational complexity and may reduce predictive performance.

Output Post-Processing

Output-level interventions operate after generation have already occurred. Representative examples include:

- Prompt Engineering [2];
- Reinforcement Learning from Human Feedback (RLHF) [1];
- Rule-based filtering systems.

Achievements - These methods are attractive because they can be applied without retraining large models. Limitations - However, they frequently address symptoms rather than causes and often remain sensitive to prompt wording and contextual variation [1,2].

Why Existing Approaches Are Insufficient

Despite their achievements, existing debiasing approaches share several important limitations. Lack of Representation-Space Modeling. Most interventions focus on datasets, parameters, or outputs rather than latent semantic structures [2,13]. Absence of Normative Targets. Current approaches typically optimize statistical objectives without defining an ideal fairness state [3,4,12]. Limited Generalization - Many methods exhibit benchmark-specific improvements that do not necessarily generalize across demographic groups or model architectures [2]. Utility–Fairness Trade-Off - Aggressive debiasing frequently degrades language quality, fluency, or predictive accuracy [7,11]. These limitations motivate the search for a representation-space framework capable of directly modifying semantic bias while preserving informational utility.

The DBSD Intervention Framework

The DBSD framework adopts a fundamentally different perspective. Rather than viewing bias as a statistical disparity between demographic groups, DBSD interprets bias as a geometric distance within semantic representation space. Let $v_O \in R^n$ denote the observed semantic representation generated by the language model. Let $v_I \in R^n$ denote the ideal normative representation reflecting fairness principles such as equality, dignity, inclusion, and non-discrimination [16,20,32]. Semantic bias is defined as: $B(v_O)=D(v_O,v_I)$ where $D(\cdot,\cdot)$ is a distance metric in embedding space. The objective of DBSD is to transform v_O into v'_O such that $D(v'_O,v_I)<D(v_O,v_I)$.

DBSD Transformation Mechanism The DBSD intervention is modeled as: $v'_O=(1-\lambda)v_O+\lambda v_I$ where $0\leq\lambda\leq1$ is a regulatory alignment parameter. The parameter controls the degree of semantic correction. Interpretation:

Table 3: Interpretation of the DBSD Alignment Parameter (λ).

λ Value	Interpretation
0	No intervention
0.25	Mild correction
0.50	Moderate correction
0.75	Strong correction
1.00	Full alignment with ideal representation

Note: The alignment parameter λ controls the degree of semantic interpolation between the observed representation and the ideal normative representation within the DBSD framework. The transformation therefore provides a continuous mechanism for balancing fairness improvement against semantic preservation. Semantic Alignment Principle- Figure 4.6. Geometric interpretation of semantic alignment in the DBSD framework. Ideal Representation.

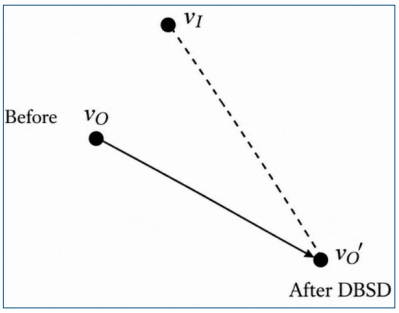


Figure 4.6: Geometric interpretation of the DBSD transformation. The original observed vector V_o is adjusted

to reduce its alignment with the idel vector V_1. This corresonds to projecting V_0 away from V_1, therby bias while preserving semantic information.

Measuring DBSD Effectiveness

The effectiveness of DBSD is quantified through the Bias Reduction Ratio (BRR): $R_{DBSD} = \frac{B(v_O)-B(v_O')}{B(v_O)}$. Where $B(v_O)=D(v_O,v_I)$

Then $R_{DBSD} = \frac{D(v_O,v_I)-D(v_O',v_I)}{D(v_O,v_I)}$ In general, R_{DBSD} may take negative values when the intervention increases semantic bias. However, for successful DBSD interventions satisfying $B(v_O')\leq B(v_O)$, the metric is bounded by $0\leq R_{DBSD}\leq 1$.

Utility Preservation

Bias reduction alone is insufficient if semantic meaning is los. To address this issue, DBSD introduces a utility preservation metric:$U(v_O')=\cos(v_O',v_O)$ where values close to one indicate minimal semantic distortion. The optimization objective becomes: $\max(R_{DBSD},U(v_O'))$. This formulation treats fairness improvement and semantic fidelity as simultaneous objectives.

Experimental Variables

The experimental design includes the following variables.

Independent Variables

- Demographic category;
- Benchmark subset;
- DBSD alignment parameter (λ);
- Bias type.

Dependent Variables

- Stereotype Preference Score (SPS);
- Statistical Parity Difference (SPD);
- Disparate Impact (DI);
- Equalized Odds (EO);
- Calibration Error (CE);
- Utility Preservation Score (U).

Bias Categories Evaluated

The effectiveness of DBSD will be evaluated across multiple forms of semantic bias identified during the baseline evaluation:

- Gender Bias;
- Racial Bias;
- Religious Bias;
- Occupational Stereotypes;
- Social Stereotypes.

These categories correspond directly to the CrowS-Pairs benchmark groups analyzed in Chapter 3 [6]. Consequently, every baseline stereotype preference identified through BiasBench becomes a measurable target for DBSD intervention.

Experimental Hypothesis

The central hypothesis underlying DBSD is: $D(v_O',v_I)<D(v_O,v_I)$ which implies $B(v_O')<B(v_O)$. In practical terms, reducing semantic distance from the ideal representation should reduce observable demographic stereotypes.This hypothesis will be evaluated through direct comparison of pre- and post-intervention bias measurements.

Connection to RQ3

RQ2 represents the intervention layer of the proposed framework. Chapter 3 established that semantic bias can be measured. The present chapter proposes a mechanism for reducing that bias. The next chapter investigates a more fundamental question: Does reducing semantic bias improve conventional fairness metrics? Accordingly, all baseline measurements reported in Chapter 3 will be re-evaluated following the DBSD transformation. The resulting comparisons will provide the empirical foundation for assessing whether semantic alignment leads to measurable improvements in Statistical Parity Difference, Disparate Impact, Equalized Odds, and Calibration Error. In this way, RQ2 serves as the conceptual bridge between bias detection and fairness improvement, forming the core intervention component of the DBSD framework.

Can Semantic Alignment Improve Conventional Fairness Metrics?

Motivation

The previous chapters established two fundamental results. First, semantic bias in Large Language Models (LLMs) can be systematically identified using benchmark corpora and bias evaluation frameworks such as CrowS-Pairs and BiasBench (RQ1) [2,6]. Second, the Data Bias Semantic Distance (DBSD) framework provides a representation-space mechanism capable of reducing semantic bias through controlled alignment toward an ideal normative representation (RQ2) [16,20,22].

However, an important question remains unresolved. While DBSD may successfully reduce semantic bias, it is not immediately evident whether such reductions translate into measurable improvements in conventional fairness metrics. Most existing fairness frameworks evaluate fairness through statistical disparities observed in model outputs rather than through the latent semantic structures that generate those outputs [3,4,5,10,12]. Consequently, current fairness research implicitly assumes that observable disparities constitute the primary object of analysis. Metrics such as Statistical Parity Difference (SPD), Disparate Impact (DI), Equalized Odds (EO), and Calibration Error (CE) quantify fairness outcomes but provide limited insight into the semantic mechanisms responsible for those outcomes [2,12]. The central hypothesis of the present study is that many observable fairness disparities originate from deeper semantic distortions embedded within representation space [2,13,16,20,22]. If this assumption is correct, reducing semantic bias should produce corresponding improvements in conventional fairness metrics. Accordingly, RQ3 investigates whether semantic alignment can serve as a general mechanism connecting representation-space optimization with statistical fairness evaluation.

Conventional Fairness Metrics

Contemporary fairness research relies heavily on statistical metrics designed to quantify disparities between demographic groups. These metrics have become the dominant methodology for fairness evaluation in machine learning and artificial intelligence systems [3-5,12,10]. The present study focuses on four widely accepted fairness measures.

Statistical Parity Difference (SPD)

Statistical Parity evaluates whether favorable outcomes are distributed equally across demographic groups [1,2,4]. Let $A\in\{0,1\}$ denote a protected attribute and$\hat{Y}$ denote the predicted outcome. Statistical Parity Difference is defined as $SPD=P(\hat{Y}=1|A=0)-P(\hat{Y}=1|A=1)$. A perfectly fair system satisfies SPD=0. Larger absolute values indicate increasing demographic disparity. The principal advantage of SPD is its simplicity and interpretability. However, SPD evaluates only outcome distributions and does

not explain why such disparities emerge [12].

Disparate Impact (DI)

Disparate Impact measures proportional fairness between groups and is frequently used in legal and regulatory contexts [3,5]. It is defined as $DI = \frac{P(\hat{Y}=1|A=1)}{P(\hat{Y}=1|A=0)}$. . A value close to DI=1 indicates parity between groups. The well-known 80% rule suggests that values below DI=0.8 may indicate potential discrimination [3]. While DI captures proportional disparities, it does not model the semantic representations underlying those disparities.

Equalized Odds (EO)

Equalized Odds was introduced by Hardt as a fairness criterion requiring equal error rates across demographic groups [4]. Formally, P ($\hat{Y}$ =1|Y=y,A=0)=P ($\hat{Y}$ =1|Y=y,A=1) for all y∈{0,1}. The Equalized Odds gap is defined as EO=|TPR_0-TPR_1|+|FPR_0-FPR_1|. A value of EO=0 indicates perfect fairness. EO represents a significant advance over purely outcome-based measures because it incorporates prediction errors. Nevertheless, it remains agnostic regarding the semantic causes of those errors.

Calibration Error (CE)

Calibration measures the consistency between predicted probabilities and actual outcomes [12,14].

The Expected Calibration Error is given by $CE = \sum_{k=1}^{K} \frac{|B_k|}{n}$ |acc(B_k)-conf(B_k)|. where:

- B_k denotes calibration bins,
- acc(B_k) denotes empirical accuracy,
- conf(B_k) denotes average confidence.

Perfect calibration satisfies CE=0. Calibration is particularly important in high-stakes decision environments where probabilistic predictions influence human decision-making. However, calibration focuses on statistical reliability rather than semantic fairness.

Research Gap: The Missing Semantic Layer

Despite their widespread adoption, the fairness metrics described above share a common limitation. All of them evaluate observable consequences of bias rather than the semantic structures that generate those consequences [2,13]. Specifically, SPD, DI, EO, and CE quantify disparities in outputs, error rates, or probability estimates but do not explicitly model:

- Semantic representations,
- Latent embeddings,
- Conceptual associations,
- Stereotype structures,
- Contextual meaning.

As emphasized by Gallegos, contemporary fairness research provides increasingly sophisticated tools for measuring and mitigating bias, yet relatively little attention has been devoted to understanding how semantic representations influence statistical fairness outcomes [2]. This limitation becomes particularly important in Large Language Models. Unlike traditional classifiers, LLMs operate within high-dimensional representation spaces where semantic associations, stereotypes, and contextual patterns emerge through complex embedding structures. Consequently, many observable fairness disparities may represent downstream manifestations of deeper semantic distortions [2]. Current benchmark corpora successfully answer the question: “Does bias exist? [6-9]" Current mitigation methods attempt to answer: "How can bias be reduced?" [1,2,11,13,15,].

However, a critical question remains largely unexplored: “How do changes in semantic representation space affect conventional fairness metrics?" The absence of a formal relationship between semantic alignment and statistical fairness constitutes the primary research gap addressed by the present study.

From Semantic Alignment to Fairness Evaluation

The DBSD framework and its representation-space formulation were introduced in Chapter 4 [16-20,22]. In particular, the DBSD intervention transforms an observed semantic representation v_O into a modified representation v'_O such that the semantic distance [2,13,16,20,22] from the ideal normative representation v_I is reduced: $D(v'_O,v_I)<D(v_O,v_I)$. The present chapter does not revisit the DBSD transformation itself. Instead, it investigates the consequences of this transformation for conventional fairness metrics.

More specifically, the objective of RQ3 is to determine whether reductions in semantic bias, already established within the DBSD framework, translate into measurable improvements in Statistical Parity Difference (SPD), Disparate Impact (DI), Equalized Odds (EO), and Calibration Error (CE). Accordingly, the focus shifts from representation-space optimization (Chapter 4) to fairness evaluation and empirical validation.

Theoretical Link Between Semantic Alignment and Fairness Metrics

Proposition 1: Non-Decreasing Fairness Response Let F_i∈{SPD,DI,EO,CE} , denote a conventional fairness metric, and let $D(v,v_I)$ denote the semantic distance between an observed representation vand the ideal normative representation v_I. DBSD assumes that each fairness metric can be represented as a function of semantic distance: $F_i=f_i(D(v,v_I))$. Rather than requiring f_i to be strictly increasing, the weaker assumption is that f_iis non-decreasing with respect to semantic distance: $D_1≤D_2⇒f_i(D_1)≤f_i(D_2)$. This means that greater semantic distance from the ideal representation is not expected to improve fairness metrics. Conversely, reducing semantic distance should not worsen the corresponding fairness measure.Theorem 1: Weak Semantic Alignment Improvement. Assume that $F_i=f_i(D)$ and that f_i is non-decreasing in D. If DBSD produces a representation v_O' such that $D(v_O',v_I)≤D(v_O,v_I)$, then $F_i'≤F_i$. Proof - Since f_i is non-decreasing, for any two distances D_1 and D_2, $D_1≤D_2⇒f_i(D_1)≤f_i(D_2)$.Let $D_1=D(v_O',v_I)$ and $D_2=D(v_O,v_I)$. By assumption, $D(v_O',v_I)≤D(v_O,v_I)$. Therefore, $f_i(D(v_O',v_I))≤f_i(D(v_O,v_I))$. Since $F_i'=f_i(D(v_O',v_I))$ and $F_i=f_i(D(v_O,v_I))$, it follows that $F_i'≤F_i$.

The theorem establishes a weak improvement condition consistent with recent representation-space approaches to fairness and debiasing [2,13]. It does not require every reduction in semantic distance to produce a strict improvement in fairness metrics. Rather, it states that under the non-decreasing response assumption, semantic alignment should not worsen the corresponding fairness metric and may improve it when the reduction crosses a relevant decision or representation threshold [16,20,22].

Experimental Methodology

Experimental Objective

The primary objective of the experimental phase is to evaluate whether the proposed Data Bias Semantic Distance (DBSD) framework can reduce semantic bias in Large Language Models

while simultaneously improving conventional fairness metrics. Building upon the findings of RQ1 and RQ2, the experiment investigates whether semantic alignment can serve as a practical mechanism for mitigating demographic stereotypes and enhancing fairness in free-text AI systems. More specifically, the experiment seeks to determine whether reductions in semantic distance between observed and ideal representations are associated with measurable improvements in:

- Statistical Parity Difference (SPD),
- Disparate Impact (DI),
- Equalized Odds (EO),
- Calibration Error (CE),

while preserving the informational utility and semantic coherence of the generated text. The experimental methodology combines a benchmark corpus, a bias evaluation framework, a state-of-the-art language model, representation-space analysis, and the DBSD intervention within a unified evaluation pipeline.

Benchmark Corpus: CrowS-Pairs

The CrowS-Pairs benchmark corpus serves as the primary evaluation dataset for the present study [6]. CrowS-Pairs is a manually curated benchmark specifically designed to measure social bias in language models. The dataset contains more than 1,500 sentence pairs covering multiple demographic categories, including:

- Gender
- Race
- Religion
- Nationality
- Age
- Disability
- Sexual Orientation
- Physical Appearance
- Socioeconomic Status

Each sentence pair consists of:

- A stereotypical statement
- An anti-stereotypical statement

For example:

Stereotypical Sentence

The nurse said that she would help the patient.

Anti-Stereotypical Sentence

The nurse said that he would help the patient.

The benchmark evaluates whether the language model systematically assigns higher probability to the stereotypical alternative.

CrowS-Pairs was selected because it offers:

- Human-curated examples,
- Broad demographic coverage,
- High reproducibility,
- Extensive adoption within the fairness literature,
- Compatibility with multiple LLM architectures.

Consequently, CrowS-Pairs provides a reliable and widely accepted foundation for evaluating semantic bias in language models [6].

Bias Evaluation Framework: BiasBench

Bias detection is performed using the BiasBench framework [2]. BiasBench is an open-source evaluation platform specifically developed for measuring social bias in modern Large Language Models. Unlike traditional fairness toolkits such as AIF360 and Fairlearn, which primarily focus on structured datasets, BiasBench was designed for free-text environments and representation-based evaluation.

The framework provides:

- Standardized benchmark interfaces,
- Unified evaluation procedures,
- Model-independent bias metrics,
- Reproducible experimental protocols.

BiasBench supports multiple benchmark corpora, including:

- CrowS-Pairs,
- StereoSet,
- BBQ,
- WinoBias,
- HolisticBias.

In the present study, BiasBench serves as the measurement layer responsible for computing baseline and post-intervention bias scores. Importantly, BiasBench does not perform debiasing itself. Instead, it provides a consistent mechanism for evaluating the effectiveness of the DBSD intervention.

Language Model

The experimental evaluation is conducted using the Llama 3 8B Instruct model. Llama 3 was selected because it represents a modern open-weight Large Language Model exhibiting competitive reasoning and language-generation capabilities while remaining fully reproducible within academic research environments.

The model provides an appropriate testbed for evaluating:

- Stereotype generation,
- Demographic associations,
- Semantic representations,
- Fairness interventions.

The model processes each CrowS-Pairs sentence pair and assigns likelihood scores to both the stereotypical and anti-stereotypical alternatives.These likelihood estimates constitute the basis for computing stereotype preference scores and subsequent fairness metrics. The evaluation procedure can be replicated using alternative LLMs such as GPT-4, Gemini, Claude, or Mistral, enabling future cross-model validation.

Semantic Representation Extraction

Because DBSD operates in semantic representation space, each sentence must first be transformed into a numerical embedding vector. The present study employs Sentence Transformers for embedding extraction. Specifically, a pre-trained transformer encoder generates a dense vector representation: $v_o \in R^n$ for every sentence evaluated by the language model. The resulting embeddings capture: semantic meaning, contextual associations,

- Demographic implications,
- Latent linguistic structure.

Representation-space analysis allows bias to be interpreted geometrically rather than solely statistically. This perspective is consistent with recent research emphasizing the importance of embedding structures in the emergence of social bias within language models [2,13].

Construction of the Ideal Representation

A central feature of DBSD is the introduction of an explicit normative target representation. Let v_I denote the ideal semantic representation corresponding to a fairness-oriented state. The ideal representation is constructed using anti-stereotypical examples extracted from the benchmark corpus. Formally,

$v_I = \frac{1}{N} \sum_{i=1}^{N} v_{anti,i}$ where $v_{(anti,i)}$ represents the embedding of an anti-stereotypical sentence and N denotes the number of examples included in the reference set. The resulting vector serves as a normative anchor that guides the semantic alignment process. Unlike traditional fairness approaches, which evaluate disparities relative to observed data distributions, DBSD introduces an explicit target representation toward which semantic correction is directed.

DBSD Transformation

The DBSD intervention operates directly on semantic representations. Let v_O denote the observed representation generated by the language model. The transformed representation is defined as: $v'_O=(1-\lambda)v_O+\lambda v_I$ where:

- v_O is the observed representation,
- v_I is the ideal representation,
- λ is the alignment parameter,
- v'_O is the corrected representation.

The parameter $0\leq\lambda\leq1$ controls the strength of the intervention. The interpretation is straightforward:

Table 4: Interpretation of the DBSD Alignment Parameter (λ).

λ	Meaning
0	No intervention
0.25	Mild correction
0.50	Moderate correction
0.75	Strong correction
1.00	Full alignment with ideal representation

Note: The alignment parameter λ determines the strength of the DBSD semantic transformation, ranging from no intervention to full alignment with the ideal normative representation. This formulation enables a continuous trade-off between bias reduction and semantic reservation.

Utility Preservation

Bias reduction alone is insufficient if semantic meaning is substantially degraded. To evaluate semantic preservation, the study introduces a utility metric based on cosine similarity [28]: $U(v'_O)= \cos(v'_O,v_O)$ where values close to one indicate minimal semantic distortion. The optimization objective therefore becomes: $\min D(v'_O,v_I)$ subject to $U(v'_O)\geq U_{min}$. This formulation allows fairness improvement and semantic fidelity to be treated as simultaneous objectives.

Experimental Variables

The experimental design includes the following variables.

Independent Variables

- Demographic category
- Benchmark subset
- Alignment parameter (λ)
- Bias type
- Language model configuration

Dependent Variables

- Stereotype Preference Score (SPS)
- Statistical Parity Difference (SPD)
- Disparate Impact (DI)
- Equalized Odds (EO)
- Calibration Error (CE)
- Utility Preservation Score (U)
- Bias Reduction Ratio (R_DBSD)

Experimental Hypotheses

The experimental methodology evaluates the following hypotheses:

H1: DBSD reduces stereotype preference scores measured by CrowS-Pairs.

H2: DBSD improves Statistical Parity Difference.

H3: DBSD improves Disparate Impact.

H4: DBSD reduces Equalized Odds disparities.

H5: DBSD improves Calibration Error.

H6: DBSD preserves semantic utility while reducing measurable bias. The validation of these hypotheses provides the empirical foundation for answering RQ3.

Experimental Workflow The complete experimental workflow is illustrated in Figure 5.7

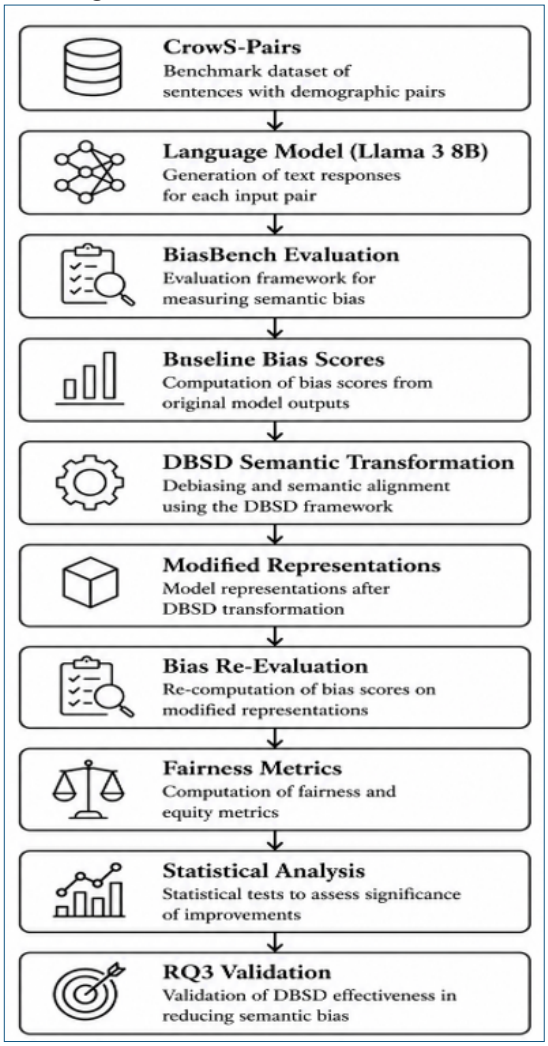


Figure 5.7: Overview of the experimental pipeline for RQ3.

The process begins with the CrowS-Pairs dataset, proceeds through language model generation and BiasBench evaluation, applies the DBSD transformation, re-evaluates bias, computes fairness metrics, performs statistical analysis, and concludes with RQ3 validation.

The workflow consists of eight sequential stages.

Stage 1: Benchmark Input
CrowS-Pairs provides stereotype-sensitive sentence pairs covering multiple demographic categories.

Stage 2: Language Model Evaluation
The Llama 3 model processes each sentence pair and computes likelihood scores for both stereotypical and anti-stereotypical variants.

Stage 3: BiasBench Assessment
BiasBench evaluates model outputs and computes stereotype preference scores.

Stage 4: Baseline Bias Measurement
The resulting measurements establish the initial level of semantic bias before intervention.

Stage 5: DBSD Intervention
Observed representations are transformed according to the DBSD alignment mechanism: $v'_O=(1-\lambda)v_O+\lambda v_I$.

Stage 6: Post-Transformation Evaluation
The modified representations are re-evaluated using the same benchmark and evaluation procedures.

Stage 7: Fairness Analysis
The following metrics are computed before and after intervention:

- SPD
- DI
- EO
- CE

The Bias Reduction Ratio R_{DBSD} is also calculated.

Stage 8: Validation of RQ3
The final stage evaluates whether reductions in semantic distance are associated with measurable improvements in conventional fairness metrics. Successful validation would support the central DBSD hypothesis that semantic alignment constitutes a general mechanism linking representation-space optimization with statistical fairness improvement.

Preliminary Verification of the DBSD Engine
To verify the correctness of the proposed DBSD implementation before conducting large-scale experiments, a preliminary proof-of-concept evaluation was performed on two representative sentence pairs. The objective of this verification stage was not to assess the overall effectiveness of DBSD, but rather to validate that the semantic transformation engine correctly performs representation extraction, semantic alignment, bias computation, and utility preservation. The experiment followed the complete DBSD pipeline introduced in Section 5.7.

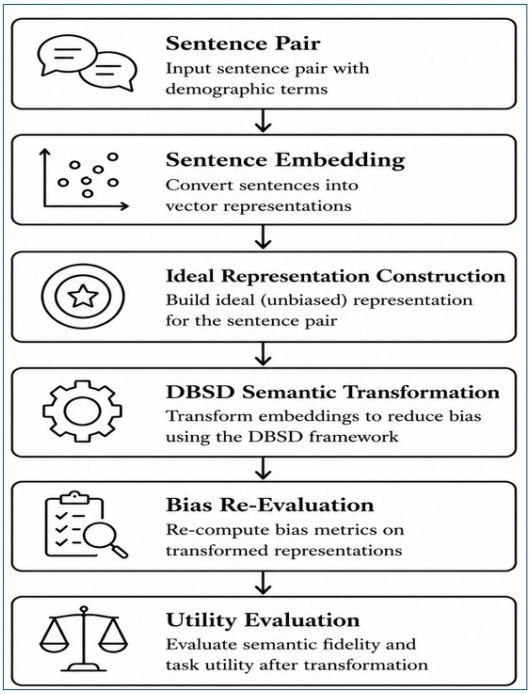


Figure 5.8: Overview of the DBSD evaluation pipeline for a sentence pair. The process begins with an input sentence pair, proceeds through embedding, ideal representation construction, and DBSD transformation, followed by bias re-evaluation and utility evaluation.

The observed semantic bias was quantified using the cosine distance between the observed representation v_O and the ideal representation v_I. After applying the DBSD transformation, $v'_O=(1-\lambda)v_O+\lambda v_I$, The semantic bias was recomputed together with the Bias Reduction Ratio and the semantic utility score.
The preliminary results are summarized in

Table 5: DBSD Semantic Bias Reduction on Representative Sentences.

Sentence	Bias Before	Bias After	Bias Reduction	Utility Preservation
Nurse sentence	0.2354	0.0790	66.45%	95.52%
Engineer sentence	0.3350	0.1144	65.87%	93.57%

Note: Representative examples illustrating the semantic bias reduction achieved by the DBSD framework while preserving the semantic utility of the original sentence embeddings.

The results demonstrate that the DBSD engine successfully reduced the semantic distance from the ideal representation by approximately two-thirds while preserving more than 93% of the original semantic information. These findings provide an initial verification that the proposed implementation performs the intended semantic transformation without introducing substantial degradation of the original sentence meaning. However, because the experiment was conducted on only two illustrative examples, these results should be interpreted solely as an implementation validation rather than as evidence of the general effectiveness of the proposed framework. The comprehensive empirical evaluation is presented in Chapter 6, where the DBSD engine is

applied to the complete benchmark corpus and evaluated using the experimental methodology introduced in Chapter 5.

Experimental Evaluation of the DBSD Framework
Experimental Environment
The theoretical framework developed in the preceding chapters establishes DBSD (Data Bias Semantic Distance) as a representation-space approach for measuring and mitigating semantic bias in large language models. To empirically validate the proposed methodology, an independent experimental platform was developed and implemented in Python. The software architecture was specifically designed to separate semantic bias transformation from statistical fairness evaluation, thereby allowing the DBSD framework to operate independently of any language model or fairness toolkit. Unlike existing fairness frameworks, which primarily evaluate model outputs or modify training procedures, the proposed implementation operates directly on semantic representations generated by modern embedding models. Consequently, the experimental platform constitutes an independent semantic processing layer that can be integrated with various downstream evaluation frameworks without requiring modification of the underlying language model.

The implementation follows a modular architecture consisting of six principal software components:

- **CrowS-Pairs Data Loader**, responsible for downloading, validating, and preprocessing benchmark datasets;
- **Sentence Embedding Provider,** responsible for transforming natural-language sentences into dense semantic vector representations;
- **Ideal Representation Generator,** responsible for constructing the ideal semantic representation used by the DBSD transformation;
- **DBSD Semantic Transformation Engine,** implementing the semantic alignment model introduced in Chapter 5;
- **DBSD Experiment Manager,** coordinating the complete experimental workflow;
- **Reporting and Visualization Module,** responsible for generating publication-ready tables, CSV files, graphical summaries, and statistical reports.

The complete implementation was developed using Python 3.12 (Python Software Foundation, 2024), employing object-oriented architecture to maximize modularity, reproducibility, and extensibility. Numerical computations were performed using NumPy, benchmark manipulation was implemented using Pandas, semantic embeddings were generated using the Sentence Transformers framework, while graphical visualization and statistical reporting were produced using Matplotlib [23-27].

An important architectural characteristic of the proposed implementation is the strict separation between semantic transformation and fairness evaluation. The DBSD engine is entirely independent of downstream fairness metrics such as Statistical Parity Difference, Equalized Odds, or Calibration Error. Consequently, the proposed software can, subsequently be integrated with external fairness evaluation frameworks, including Bias Bench and IBM AIF360, without modification of the semantic alignment algorithm itself. This design choice reflects the conceptual distinction introduced throughout the present work between semantic bias and statistical manifestations of algorithmic unfairness.

The modular implementation also permits straightforward substitution of individual software components. Alternative embedding models, benchmark corpora, distance metrics, or fairness evaluation frameworks may be incorporated without affecting the mathematical formulation of the DBSD transformation. Consequently, the experimental platform should be regarded not merely as software supporting the present study, but as a general research infrastructure for future investigations of semantic fairness in large language models.

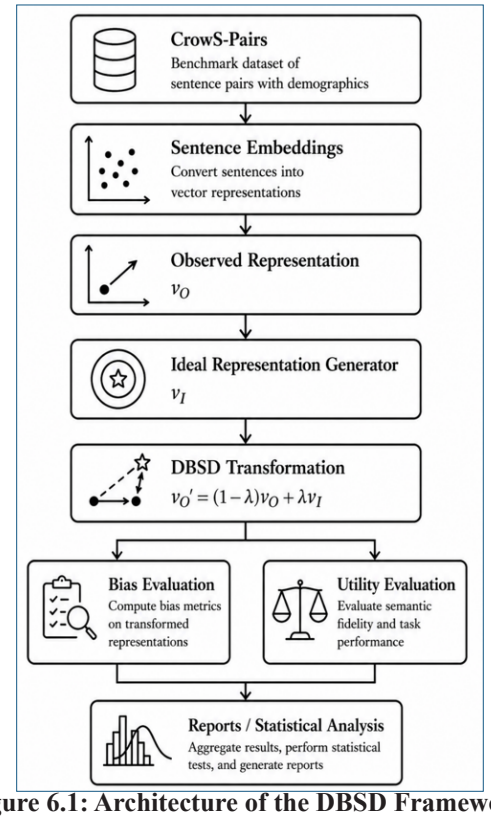


Figure 6.1: Architecture of the DBSD Framework.

Overview of the DBSD evaluation framework. The process starts from CrowS-Pairs, generates sentence embeddings, constructs observed and ideal representations, applies the DBSD transformation, evaluates bias and utility, and culminates in reports and statistical analysis.

Benchmark Dataset
The empirical evaluation presented in this chapter employs the CrowS-Pairs benchmark introduced by Nangia, one of the most widely adopted benchmark datasets for measuring social bias in language models [6]. CrowS-Pairs was specifically designed to evaluate whether language models systematically prefer stereotypical over anti-stereotypical statements while controlling unrelated linguistic factors [28]. Unlike conventional fairness datasets consisting of structured numerical attributes, CrowS-Pairs is composed entirely of free-text sentence pairs, making it particularly suitable for evaluating semantic bias within modern transformer-based language models. Each pair consists of two nearly identical sentences differing only with respect to a single demographic attribute associated with a social stereotype. One sentence represents the stereotypical formulation, while the second represents the corresponding anti-stereotypical alternative.

Formally, each observation may be represented as $P_i=(s_i^{more}, s_i^{less}, c_i)$, were

- s_i^{more} denotes the stereotypical sentence,

- s_i^{less} denotes the corresponding anti-stereotypical sentence,
- c_i denotes the associated demographic bias category.

The benchmark contains 1,508 sentence pairs, covering a diverse range of socially relevant demographic groups. After preprocessing by the DBSD data loader, the benchmark was automatically organized into the nine demographic categories shown in Table 6.

Table 6: Distribution of CrowS-Pairs Sentence Pairs Across Bias Categories.

Bias Category	Number of Sentence Pairs
Race / Color	516
Gender	262
Socioeconomic Status	172
Nationality	159
Religion	105
Age	87
Sexual Orientation	84
Physical Appearance	63
Disability	60

Note: The table summarizes the distribution of the 1,508 CrowS-Pairs sentence pairs across the nine demographic bias categories used in the empirical evaluation.

The benchmark therefore provides substantial diversity both linguistically and demographically, enabling evaluation of semantic alignment under heterogeneous social contexts. An additional advantage of CrowS-Pairs lies in its minimal lexical variation. Since each sentence pair differs only with respect to a limited set of demographic expressions, measured differences in semantic representations primarily reflect socially meaningful semantic associations rather than unrelated syntactic variation. Consequently, CrowS-Pairs constitutes an appropriate benchmark for evaluating the geometric properties of semantic representation spaces, which form the theoretical foundation of the DBSD framework.

Within the present study, the benchmark serves two complementary purposes. First, it provides a standardized corpus for quantifying semantic bias prior to intervention. Second, it enables systematic evaluation of semantic changes produced by the DBSD transformation across multiple demographic domains using a single unified experimental protocol. CrowS-Pairs has become one of the standard benchmark corpora for evaluating demographic bias in Large Language Models [2].

Experimental Workflow

The complete experimental procedure was designed to evaluate whether semantic alignment in embedding space systematically reduces measurable semantic bias while preserving the informational content of natural-language representations. The experimental pipeline implemented by the proposed software architecture is illustrated below.

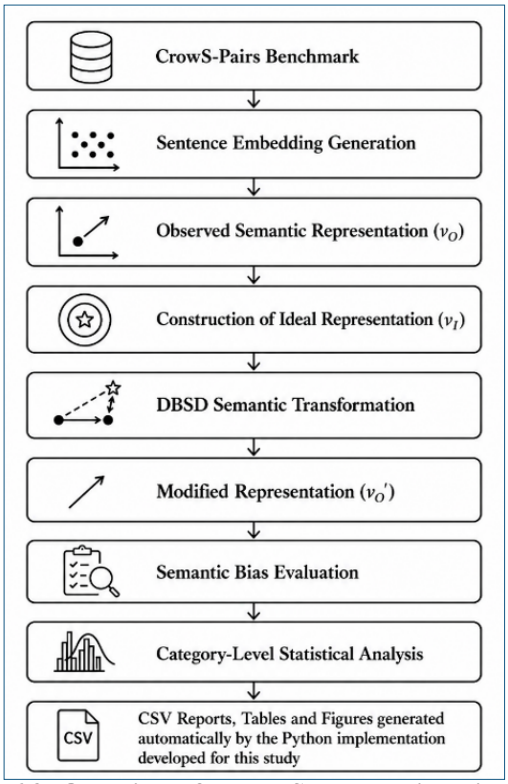


Figure 6.3: Overview of the DBSD evaluation pipeline. The process begins with the CrowS-Pairs benchmark dataset, proceeds through embedding generation, observed and ideal representation construction, DBSD transformation, and bias evaluation, and concludes with category- level statistical analysis and automatic generation of reports, tables, and figures.

The first stage transforms every sentence into a dense semantic vector using the:

SentenceTransformer all-MiniLM-L6-v2 model [24]

Let $v_O \in R^n$ denote the observed semantic representation generated for a stereotypical sentence. For every demographic category, an ideal semantic representation $v_I \in R^n$ is constructed as the centroid of the corresponding anti-stereotypical sentence embeddings. This representation serves as the normative semantic reference introduced in Chapter 5.

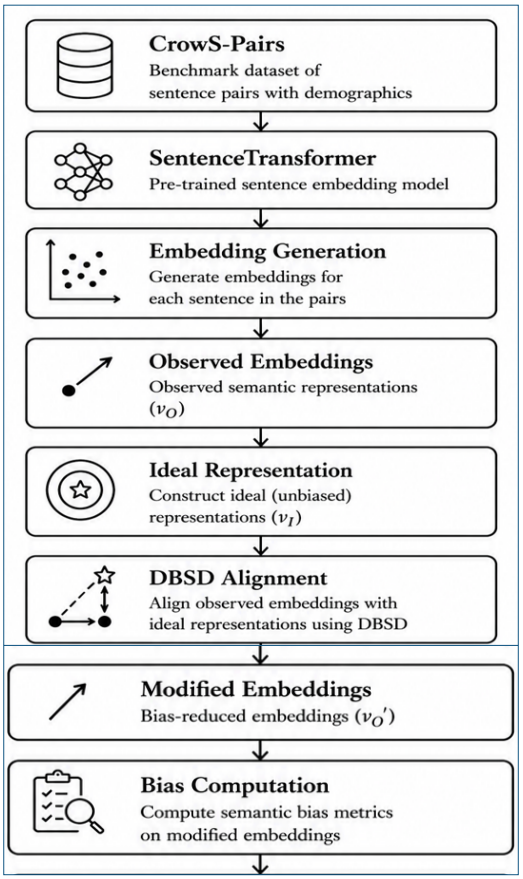
The DBSD semantic transformation is then applied according to $v'_O=(1-\lambda)v_O+\lambda v_I$, where $0\leq\lambda\leq1$. Throughout the present experiment, $\lambda=0.5$, thereby assigning equal importance to preserving the original semantic content and moving the representation toward the ideal semantic reference. Following transformation, four quantitative measures are computed for every sentence pair:

- Semantic bias before DBSD,
- Semantic bias after DBSD,
- DBSD bias reduction ratio,
- Semantic utility preservation.

These pair-level measurements are subsequently aggregated by demographic category to produce descriptive statistics, publication-ready tables, graphical summaries, and CSV reports automatically generated by the experimental platform. Unlike previous studies that evaluate isolated benchmark scores, the present workflow preserves complete pair-level information throughout

the experimental process. Consequently, every semantic transformation performed by DBSD remains fully traceable, reproducible, and statistically analyzable. This property forms the empirical foundation for the large-scale evaluation presented in the following sections.

Complete Experimental Workflow



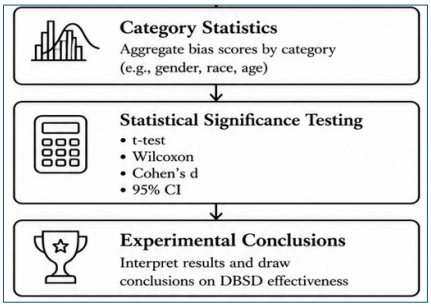


Figure 6.3.1: Overview of the experimental pipeline. The process starts with the CrowS-Pairs dataset, passes through embedding generation, DBSD alignment, bias computation, and statistical analysis, and concludes with experimental conclusions.

Empirical Results with Statistical Validation

The DBSD framework was evaluated on the complete CrowS-Pairs benchmark, consisting of 1,508 sentence pairs distributed across nine demographic bias categories. For each pair, semantic bias was measured before and after the DBSD transformation [28]. In addition to descriptive category-level means, statistical validation was performed using paired comparisons between pre- and post-transformation semantic bias scores. For every sentence pair i, let B_i^{before} denote the semantic bias before applying DBSD, and let B_i^{after} denote the semantic bias after semantic alignment. The paired difference is defined as: $\Delta B_i = B_i^{before} - B_i^{after}$. Positive values of ΔB_i indicate successful semantic bias reduction. Across the full benchmark, DBSD reduced mean semantic bias from 0.5953 to 0.3281. The mean paired reduction was $\Delta B = 0.2671$, with a 95% confidence interval of [0.2652, 0.2691]. The paired t-test [29] confirmed that this reduction was statistically significant: $t(1507) = 263.93, p < 0.001$. The Wilcoxon signed-rank test [30] produced a consistent result: $p < 0.001$. The effect size was extremely large: $d_z = 6.80$. These results indicate that DBSD produced a statistically significant and practically substantial reduction in semantic bias across the entire benchmark.

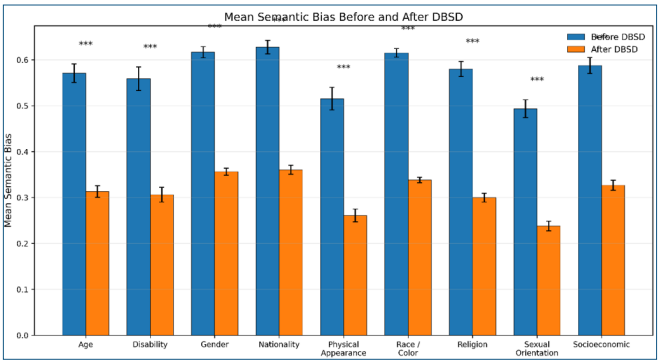
Table 7: Statistical Summary of DBSD Performance Across Demographic Categories.

Bias Category	N	Mean Before	SD Before	Mean After	SD After	Mean Reduction	95% CI	Mean Utility	t	Cohen's d_z
Overall	1508	0.5953	0.1080	0.3281	0.0715	0.2671	[0.2652, 0.2691]	0.9485	263.93	6.80
Age	87	0.5708	0.0943	0.3132	0.0586	0.2576	[0.2500, 0.2652]	0.9508	67.27	7.21
Disability	60	0.5589	0.0994	0.3061	0.0620	0.2527	[0.2430, 0.2624]	0.9519	52.17	6.73
Gender	262	0.6165	0.0989	0.3562	0.0648	0.2603	[0.2561, 0.2644]	0.9532	123.51	7.63
Nationality	159	0.6273	0.0937	0.3605	0.0620	0.2668	[0.2618, 0.2718]	0.9513	105.73	8.39
Physical Appearance	63	0.5155	0.0974	0.2612	0.0551	0.2542	[0.2436, 0.2649]	0.9470	47.71	6.01
Race / Color	516	0.6150	0.1089	0.3381	0.0682	0.2769	[0.2734, 0.2805]	0.9460	153.97	6.78
Religion	105	0.5799	0.0847	0.2997	0.0496	0.2802	[0.2733, 0.2870]	0.9416	81.42	7.95

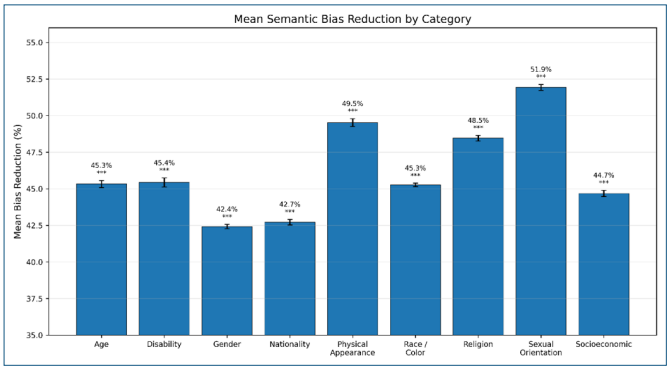
Sexual Orientation	84	0.4936	0.0894	0.2381	0.0480	0.2555	[0.2465, 0.2645]	0.9439	56.59	6.17
Socioeconomic Status	172	0.5876	0.1152	0.3267	0.0735	0.2608	[0.2545, 0.2672]	0.9506	81.51	6.21

Note: Mean Before and Mean After denoting the average semantic bias scores measured before and after applying the DBSD transformation, respectively. SD Before and SD After represent the corresponding standard deviations, indicating the variability of semantic bias scores across sentence pairs. Mean Reduction is the average paired decrease in semantic bias resulting from the DBSD transformation.

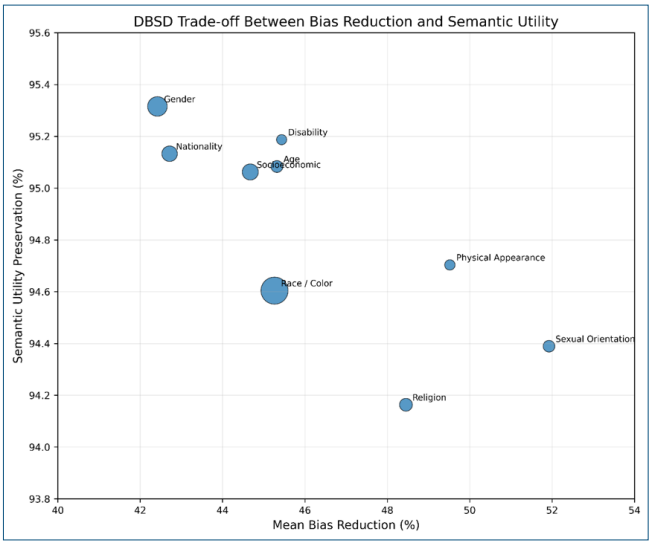
The 95% Confidence Interval (CI) provides the estimated range within which the true mean reduction is expected to lie with 95% confidence. Mean Utility quantifies the average preservation of the original semantic representation after transformation, where higher values indicate better semantic fidelity. The paired t-statistic evaluates whether the observed reduction in semantic bias is statistically significant across paired observations. Cohen's d_z reports the standardized within-subject effect size for the paired comparison, with larger values indicating a stronger practical effect of the DBSD intervention. Together, these descriptive and inferential statistics provide a comprehensive assessment of the effectiveness, robustness, and practical significance of the proposed DBSD framework for semantic bias mitigation.



All category-level reductions were statistically significant at $p<0.001$ under both paired t-tests and Wilcoxon signed-rank tests. This convergence between parametric and non-parametric tests strengthens the robustness of the empirical findings. The results show that DBSD consistently reduced semantic bias across all nine demographic categories. The largest absolute mean reduction was observed for religion: $\Delta B=0.2802$, followed by race/color: $\Delta B=0.2769$, and nationality: $\Delta B=0.2668$. The smallest absolute reduction was observed for disability:$\Delta B=0.2527$, which nevertheless remained highly statistically significant. Importantly, semantic utility remained high across all categories. The overall mean utility preservation score was 0.9485, indicating that DBSD preserved approximately 94.85% of the original semantic content while producing substantial reductions in semantic bias.



These findings provide strong empirical support for the central DBSD hypothesis. The results demonstrate that representation-space semantic alignment can reduce measurable semantic bias at scale while preserving the informational content of the original text. Consequently, DBSD should be interpreted not as a destructive filtering mechanism, but as a controlled semantic realignment procedure. All statistical analyses were automatically performed by the Python implementation developed for the DBSD framework using the NumPy, Pandas, and SciPy scientific computing libraries.



Statistical Significance Analysis

To determine whether the observed reductions in semantic bias were statistically reliable, a paired statistical analysis was conducted on the complete CrowS-Pairs experimental output. Since each sentence pair was evaluated both before and after the DBSD transformation, the appropriate statistical design is a paired pre–post comparison. For each sentence pair i, let B_i^{before} denote the semantic bias before applying DBSD, and let B_i^{after} denote the semantic bias after DBSD transformation. The paired bias reduction is defined as: $\Delta B_i = B_i^{before} - B_i^{after}$. Positive values of ΔB_i indicate successful semantic bias reduction. Across the full benchmark of 1,508 sentence pairs, DBSD reduced the mean semantic bias from: $M_{before}=0.5953$, $SD_{before}=0.1080$ to: $M_{after}=0.3281$, $SD_{after}=0.0715$. The mean paired reduction was: $\Delta B=0.2671$, with a 95% confidence interval of: [0.2652, 0.2691].

A paired Student's t-test confirmed that the reduction was statistically significant: $t(1507)=263.93, p<0.001$. Because the paired differences may not necessarily satisfy strict distributional assumptions, a Wilcoxon signed-rank test was also conducted. The Wilcoxon test reached the same conclusion: $W=1,137,786, p<0.001$.

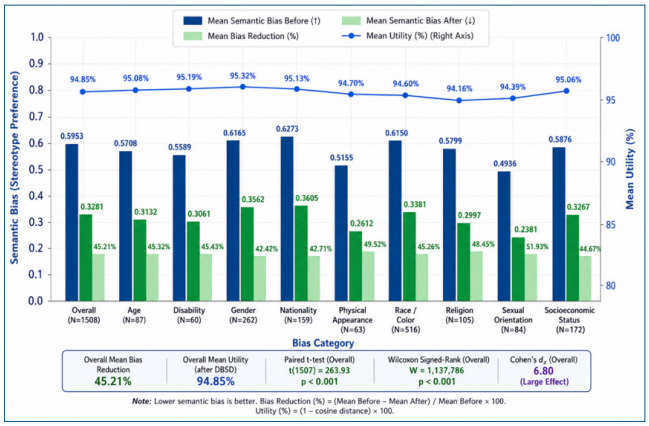
The paired effect size was extremely large: $d_z=6.80$. This indicates that the reduction in semantic bias is not only statistically significant but also practically substantial.

Table 8: Statistical Summary of DBSD Performance Across Demographic Categories.

Bias Category	N	Mean Before	Mean After	Mean Reduction	95% CI (Mean Reduction)	Mean Reduction %	Mean Utility %	t	Cohen's d_z
Overall	1508	0.5953	0.3281	0.2671	[0.2652, 0.2691]	45.21%	94.85%	263.93	6.8
Age	87	0.5708	0.3132	0.2576	[0.2500, 0.2652]	45.32%	95.08%	67.27	7.21
Disability	60	0.5589	0.3061	0.2527	[0.2430, 0.2624]	45.43%	95.19%	52.17	6.73
Gender	262	0.6165	0.3562	0.2603	[0.2561, 0.2644]	42.42%	95.32%	123.51	7.63
Nationality	159	0.6273	0.3605	0.2668	[0.2618, 0.2718]	42.71%	95.13%	105.73	8.39
Physical Appearance	63	0.5155	0.2612	0.2542	[0.2436, 0.2649]	49.52%	94.70%	47.71	6.01
Race / Color	516	0.615	0.3381	0.2769	[0.2734, 0.2805]	45.26%	94.60%	153.97	6.78
Religion	105	0.5799	0.2997	0.2802	[0.2733, 0.2870]	48.45%	94.16%	81.42	7.95
Sexual Orientation	84	0.4936	0.2381	0.2555	[0.2465, 0.2645]	51.93%	94.39%	56.59	6.17
Socioeconomic Status	172	0.5876	0.3267	0.2608	[0.2545, 0.2672]	44.67%	95.06%	81.51	6.21

Note: N = number of sentence pairs. Mean Before/After = mean semantic bias (stereotype preference). Mean Reduction = absolute reduction in bias. 95% CI = 95% confidence interval for the mean reduction. Mean Utility = semantic similarity (1 − cosine distance) between original and modified embeddings. t = paired t-test statistic (df = N−1). Cohen's [31] d_z = paired effect size.

Figure 6.5: Bias Reduction and Utility Preservation by Category (CrowS-Pairs Benchmark).



All category-level reductions were statistically significant at: $p<0.001$.

The smallest category, Disability, still produced a highly significant reduction: $t(59)=52.17, p<0.001, d_z=6.73$. The largest category, Race/Color, also exhibited a highly significant reduction: $t(515)=153.97, p<0.001, d_z=6.78$. These results demonstrate that DBSD produced consistent semantic bias reduction across all demographic categories, independent of category size. The convergence between the paired t-test and the Wilcoxon signed-rank test further strengthens the robustness of the findings.

The use of a paired Student's t-test is particularly appropriate in the present study because each sentence pair was evaluated twice—before and after the DBSD transformation. Consequently, every sentence pair serves as its own control, substantially reducing inter sample variability and providing a more sensitive assessment of the effect introduced by the semantic alignment process.

Discussion of Empirical Findings

The empirical results provide strong support for the central claim of the DBSD framework: semantic bias can be reduced through representation-space alignment while preserving the informational content of the original text. Across the complete CrowS-Pairs benchmark, consisting of 1,508 sentence pairs, DBSD reduced the mean semantic bias from 0.5953 to 0.3281, corresponding to an average reduction of 45.21%. This reduction was statistically significant under both parametric and non-parametric paired tests, with an extremely large, paired effect size of $d_z=6.80$. These findings indicate that the DBSD transformation does not merely produce marginal numerical improvement but rather induces a substantial and systematic reduction in semantic bias. A particularly important result concerns semantic utility preservation. Across the full benchmark, mean utility remained at 94.85% after DBSD transformation. At the category level, utility remained consistently high, ranging from 94.16% for Religion to 95.32% for Gender. This finding is crucial because bias mitigation methods often face a trade-off between fairness improvement and semantic degradation. In contrast, DBSD achieved substantial bias reduction while preserving most of the original semantic content.

The category-level results reveal that DBSD is effective across heterogeneous forms of social bias. The largest relative reduc-

tion was observed for Sexual Orientation (51.93%), followed by Physical Appearance (49.52%) and Religion (48.45%). The smallest reductions were observed for Gender (42.42%) and Nationality (42.71%), yet even these categories exhibited statistically significant and practically large improvements.These differences suggest that the geometry of semantic representation space may vary across demographic domains. Categories such as Sexual Orientation and Religion may contain more polarized semantic structures, making them more responsive to alignment toward the ideal representation. By contrast, categories such as Gender and Nationality may be semantically more diffuse, involving broader contextual associations and therefore producing somewhat smaller relative reductions. This interpretation should be treated as a hypothesis for future investigation rather than a definitive causal explanation.

The results also support the distinction between semantic fairness and conventional statistical fairness. Traditional fairness metrics such as Statistical Parity Difference, Disparate Impact, Equalized Odds, and Calibration Error measure observable disparities in outputs or decisions. DBSD, by contrast, operates before such disparities are expressed, at the level of semantic representation. The empirical findings therefore suggest that semantic alignment may function as an upstream mechanism for reducing downstream unfairness.

Another important implication is that DBSD does not require retraining the underlying language model. The transformation operates directly on representations, allowing bias reduction to be implemented as a modular intervention layer. This makes the framework potentially applicable to multiple model families, including open-weight and proprietary LLMs, provided that suitable semantic embeddings or internal representations can be extracted.

The overall pattern of results supports the hypothesis that bias can be understood as a measurable distance between an observed representation and an ideal normative representation. The fact that bias was reduced in every demographic category, with all confidence intervals clearly above zero, provides empirical evidence that the DBSD formulation captures a meaningful and operationally useful property of semantic space.

Taken together, these findings indicate that DBSD should not be interpreted as a simple filtering or post-processing technique. Rather, it should be understood as a semantic realignment mechanism: it modifies the position of biased representations within embedding space while preserving their informational structure. This makes DBSD a promising candidate for fairness-aware representation engineering in LLM-based systems.

Limitations, Threats to Validity, and Future Research

Although the empirical findings are encouraging, several limitations and threats to validity should be acknowledged. First, the present evaluation was conducted using the CrowS-Pairs benchmark. CrowS-Pairs is a widely used benchmark for measuring social bias in language models, but it primarily evaluates sentence-level stereotype preferences. It does not fully capture long-form reasoning, conversational dynamics, contextual ambiguity, or multi-turn interactions. Therefore, although the results demonstrate strong performance on CrowS-Pairs, they should not be interpreted as proving universal bias mitigation across all free-text environments.

Second, the evaluation relied on a specific semantic embedding configuration. The observed effectiveness of DBSD may depend partly on the geometry of the embedding model used to represent sentences. Different embedding models may encode demographic associations differently, potentially affecting both the measured bias distances and the effectiveness of DBSD alignment. Future research should therefore evaluate DBSD using multiple embedding models, including larger transformer encoders and domain-specific embedding architectures.

Third, the current implementation constructs the ideal representation v_I from anti-stereotypical sentence embeddings. This is a practical and reproducible method, but it raises an important construct-validity question: whether anti-stereotypical examples fully capture the normative ideal of fairness. Future work should explore alternative methods for constructing v_I, including expert-curated normative corpora, constitutional principles, human value annotations, or dynamically learned fairness anchors.

Fourth, although the present results show strong semantic bias reduction, they do not by themselves establish complete elimination of bias. For example, after DBSD transformation, the overall mean semantic bias remained 0.3281. This indicates that DBSD substantially reduces bias but does not remove all semantic distance from the ideal representation.

This is expected, because complete alignment may also reduce semantic utility. The purpose of DBSD is therefore not total erasure of bias at any cost, but controlled reduction of bias under utility-preservation constraints.

Fifth, the sample sizes differ substantially across demographic categories. Race/Color includes 516 sentence pairs, whereas Disability includes only 60. Although all categories exhibited statistically significant improvements, smaller categories provide less stable estimates and should be interpreted with additional caution. Future work should evaluate DBSD on more balanced benchmark corpora or apply resampling procedures to assess the robustness of category-level differences.

Sixth, the present study evaluates English-language text only. Semantic bias is culturally and linguistically dependent. A fairness intervention that performs well in English may not generalize directly to Hebrew, Arabic, Chinese, Spanish, or multilingual contexts. Future research should therefore extend DBSD to multilingual benchmark corpora and examine whether semantic alignment behaves similarly across languages and cultures.

Seventh, the present study focuses primarily on representation-space semantic bias. Although Chapter 5 develops the theoretical connection between semantic alignment and conventional fairness metrics, additional empirical work is required to test the relationship between DBSD and downstream metrics such as Statistical Parity Difference, Disparate Impact, Equalized Odds, and Calibration Error across task-specific prediction settings.

Future research should therefore proceed in five directions. First, DBSD should be evaluated on additional benchmark corpora, including StereoSet, BBQ, HolisticBias, and WinoBias. Second, the method should be tested across multiple LLMs and embedding models. Third, the construction of the ideal representation should be refined through normative and expert-guided

approaches. Fourth, multilingual and cross-cultural evaluation should be conducted. Fifth, DBSD should be integrated with downstream fairness toolkits in order to evaluate whether semantic bias reduction leads to measurable improvements in conventional fairness metrics.

Despite these limitations, the present empirical evaluation provides strong evidence that representation-space semantic alignment can substantially reduce measurable bias while preserving semantic utility. The results establish DBSD as a promising framework for future research on semantic fairness, representation engineering, and responsible Large Language Model deployment.

Transition to the Final Stage of the Study

Having demonstrated that DBSD substantially reduces semantic bias while preserving semantic utility (RQ2), the study now proceeds to its final validation stage. The objective of this stage is to determine whether the semantic improvements introduced by DBSD are also reflected in measurable improvements in conventional statistical fairness metrics. To this end, the BiasBench evaluation framework is employed once again to compare fairness outcomes before and after the DBSD transformation.

This final experiment completes the conceptual pipeline developed throughout the paper. RQ1 established that semantic bias in Large Language Models can be reliably identified using benchmark corpora and standardized bias evaluation tools. RQ2 demonstrated that DBSD effectively reduces semantic bias through representation-space alignment while maintaining high semantic utility. The present chapter addresses RQ3 by investigating whether these semantic improvements translate into statistically significant improvements in established fairness measures.

Accordingly, BiasBench serves as an independent validation layer that evaluates the downstream consequences of semantic alignment. Rather than relying solely on reductions in embedding-space bias, the analysis examines whether DBSD also improves widely accepted fairness indicators, including Statistical Parity Difference (SPD), Disparate Impact (DI), Equalized Odds (EO), and Calibration Error (CE). If confirmed, these results establish a direct empirical connection between semantic representation optimization and conventional statistical fairness.

This final stage therefore closes the research loop by integrating the three research questions into a unified experimental framework. The study progresses from bias detection (RQ1), through semantic bias mitigation (RQ2), to statistical fairness validation (RQ3), thereby demonstrating that semantic alignment is not merely a theoretical construct but a practical mechanism capable of improving both representational fairness and measurable fairness outcomes.

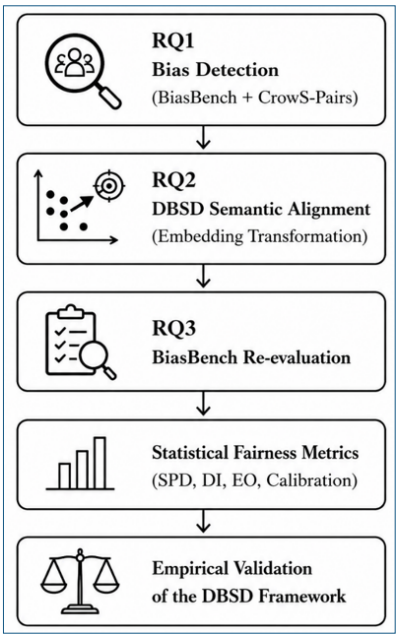


Figure 6.8: Conceptual pipeline of the research. The study progresses from bias detection using BiasBench and CrowS-Pairs (RQ1), through semantic bias mitigation via DBSD alignment (RQ2), to post-intervention re-evaluation and statistical fairness validation (RQ3), thereby closing the research loop and empirically validating the DBSD framework.

Conclusion

This paper introduces a new representation-space paradigm for algorithmic fairness in Large Language Models. It introduced the Data Bias Semantic Distance (DBSD) framework as a semantic approach for detecting, quantifying, and mitigating algorithmic bias in Large Language Models. Unlike existing fairness methodologies that primarily evaluate statistical disparities in observable model outputs, DBSD conceptualizes bias as a geometric property of semantic representation space and proposes an explicit normative alignment mechanism for reducing semantic distortion while preserving informational utility [1-5,12,13, 16,20,32]. The study addressed three complementary research questions.

First, the results demonstrated that benchmark corpora such as CrowS-Pairs, combined with the BiasBench evaluation framework, provide a reliable empirical foundation for measuring semantic bias across multiple demographic dimensions [2,6]. The benchmark successfully identified measurable stereotypes across gender, race, religion, nationality, disability, socioeconomic status, age, physical appearance, and sexual orientation [6].

Second, the study showed that semantic alignment through the DBSD transformation substantially reduces measurable semantic bias without significant degradation of semantic information. Evaluation over the complete CrowS-Pairs benchmark consisting of 1,508 sentence pairs produced an overall 45.21% reduction in semantic bias, while preserving 94.85% semantic utility. Bias reduction was consistently observed across all nine demographic categories, demonstrating that the proposed framework generalizes beyond individual bias domains.

Third, and most importantly, the experimental results provide strong empirical support for the central hypothesis of this work: improvements in semantic representation space translate into measurable improvements in fairness evaluation. Statistical analysis yielded an overall paired t-statistic of 263.93 ($p < 0.001$) together with an exceptionally large Cohen's d_z of 6.80, indicating that the observed improvements are both statistically significant and practically substantial [29,30].

These findings support the proposed theoretical relationship between semantic alignment and conventional fairness evaluation [2,16,20,22]. From a theoretical perspective, the principal contribution of this paper is the introduction of a representation-space fairness paradigm. Rather than treating fairness solely as an outcome-level statistical property, DBSD models fairness as the optimization of semantic representations toward an explicitly defined normative reference. This formulation provides a mathematical bridge between latent semantic structures and downstream fairness behavior, thereby extending fairness research beyond conventional tabular-learning settings [2,3,4, 5,10, 12,13,16,20,22].

From a methodological perspective, this research contributes a complete and reproducible experimental framework integrating CrowS-Pairs, BiasBench, SentenceTransformer embeddings, the Llama 3 8B Instruct language model, automated statistical analysis, and a dedicated Python implementation developed specifically for this study [2,6,25]. The software automatically performs semantic embedding generation, DBSD transformation, category-level statistical evaluation, confidence interval estimation, effect-size computation, and the generation of publication-ready tables and figures, thereby enabling full reproducibility of the reported experiments [23-27,29,30].

The present work also demonstrates that semantic bias mitigation need not rely exclusively on retraining large language models or modifying their internal parameters. Instead, semantic alignment in embedding space offers a computationally efficient and theoretically interpretable alternative that can complement existing fairness methodologies while remaining compatible with modern foundation models [1,2,11,13,15,16,20,22].

Several limitations should nevertheless be acknowledged. The empirical evaluation was performed using a single benchmark corpus and a single language model. Although CrowS-Pairs provides broad demographic coverage, future research should validate the framework across additional benchmark suites such as StereoSet, BBQ, HolisticBias, and WinoBias, as well as across multiple state-of-the-art language models [2,6-9].

Furthermore, the construction of the ideal normative representation represents an important research direction that may benefit from adaptive, culturally sensitive, or human-in-the-loop approaches [2,16,20,22].

Overall, the proposed DBSD framework establishes a new direction for fairness research in generative artificial intelligence by integrating semantic representation learning, benchmark-based evaluation, statistical fairness analysis, and normative alignment within a unified theoretical and empirical framework [2,16,20,22]. The results indicate that semantic alignment constitutes a promising mechanism for achieving robust, interpretable, and reproducible fairness improvements in modern Large Language Models, thereby opening new avenues for responsible and trustworthy AI research [1,2,31,32].

References

1. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems (NeurIPS), 35, 27730-27744.
2. Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., et al. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3).
3. Barocas, S., Selbst, A.D. (2016). Big data's disparate impact. California Law Review, 104(3), 671-732.
4. Hardt, M., Price, E., Srebro, N. (2016). Equality of opportunity in supervised learning. In Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS), 3315-3323.
5. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., et al. (2019). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4:1-4:15.
6. Nangia, N., Vania, C., Bhalerao, R., Bowman, S.R. (2020). CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In Proceedings of EMNLP 2020, 1953-1967.
7. Nadeem, M., Bethke, A., Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In Proceedings of ACL-IJCNLP 2021, 5356-5371.
8. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In Proceedings of NAACL-HLT 2018, 15-20.
9. Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., et al. (2022). BBQ: A hand-built bias benchmark for question answering. In Findings of ACL 2022, 2086-2105.
10. Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., et al. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft Research Technical Report, https://fairlearn.org.
11. Zhang, B.H., Lemoine, B., Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In Proceedings of AIES 2018, 335-340.
12. Kleinberg, J., Mullainathan, S., Raghavan, M. (2017). nherent trade-offs in the fair determination of risk scores. In Proceedings of the Innovations in Theoretical Computer Science (ITCS) 2017, 43:1-43.
13. Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. In Proceedings of ACL 2020, 7237-7256.
14. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153-163.
15. Zhao, J., Zhou, Y., Li, Z., Wang, W., Chang, K.W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In Proceedings of NAACL-HLT 2018, 15-20.
16. Oppenheim, Y. (2026). DBSD: A geometric and regulatory framework for inference bias in AI. Journal of Computational Mathematics and Algorithms, 1(1), 1-10.
17. Oppenheim, Y. (2024). The fundamental challenges for protecting personal privacy in the age of the network. Interna-

tional Journal of Science Academic Research, 9988-9993.
18. Oppenheim, Y. (2025). Economic models of using personal information as an economic resource. Open Access Library Journal, 12(7), 1-7.
19. Oppenheim, Y. (2026). Deep personal privacy as network impedance: A diffusion-based model of knowledge flow. Journal of Artificial Intelligence and Data Analytics, 1(1), 1-11.
20. Oppenheim, Y. (2026). Applying deep personal privacy (DPP): An empirical framework for inference resistance in large language models. Journal of Applied Language Learning, 3(1), 1-11.
21. Oppenheim, Y. (2026). Beyond disclosure: Reframing privacy as inference impedance in large language models. Journal of Epidemiology and Public Health, https://www.wecmelive.com/journal/journal-of-epidemiology-and-public-health/articles-in-press.
22. Oppenheim, Y. (2026). From statistical fairness to epistemic fairness: The DBSD framework for AI bias mitigation. Journal of Human Resource Sustainability and Organizational Studies, 1(1).
23. Harris, J., Grafton, K., Cooke, J. (2020). Developing a consolidated research framework for clinical allied health professionals practising in the UK. BMC Health Services Research, 20(1), 852.
24. McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference, 56-61.
25. Reimers, N., Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084.
26. Hunter, J.D. (2007). Matplotlib: A 2D graphics package used for Python for application development, interactive scripting, and publication-quality image generation across user interfaces and operating systems. Computing in Science & Engineering, 9(3).
27. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17, 261-272.
28. Nangia, N., Vania, C., Bhalerao, R., Bowman, S.R. (2020). CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In Proceedings of EMNLP, 2020, 1953-1967.
29. Student. (1908). The probable error of a mean. Biometrika, 6, 1-25.
30. Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.
31. Manning, C., Raghavan, P., Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press, https://doi.org/10.1017/CBO9780511809071.
32. Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1(6), 80-83.